Universidad de San Carlos de Guatemala
Facultad de Ingeniería
Escuela de Estudios de Postgrado
Maestría en Artes en Estadística Aplicada

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

Ing. Pedro Pablo Morales Ortiz
Asesorado por el Mtro. Ing. José Rolando Chávez Salazar

Guatemala, enero de 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

TRABAJO DE GRADUACIÓN

PRESENTADO A LA JUNTA DIRECTIVA DE LA
FACULTAD DE INGENIERÍA
POR

ING. PEDRO PABLO MORALES ORTIZ
ASESORADO POR EL MTRO. ING. JOSÉ ROLANDO CHÁVEZ SALAZAR

AL CONFERÍRSELE EL TÍTULO DE

MAESTRO EN ARTES EN ESTADÍSTICA APLICADA

GUATEMALA, ENERO DE 2024

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA

FACULTAD DE INGENIERÍA

NÓMINA DE JUNTA DIRECTIVA

DECANO a.i.	Ing. José Francisco Gómez Rivera
VOCAL II	Ing. Mario Renato Escobedo Martínez
VOCAL III	Ing. José Milton de León Bran
VOCAL IV	Ing. Kevin Vladimir Cruz Lorente
VOCAL V	Ing. Fernando José Paz González
SECRETARIO	Ing. Hugo Humberto Rivera Pérez

TRIBUNAL QUE PRACTICÓ EL EXAMEN DE DEFENSA DE TESIS

DECANO a.i.	Ing. José Francisco Gómez Rivera
EXAMINADORA	Mtra. Inga. Aurelia Anabela Córdova Estrada
EXAMINADOR	Mtro. Ing. Edwin Adalberto Bracamonte Orozco
EXAMINADOR	Mtro. Ing. William Eduardo Fagiani Cruz
SECRETARIO	Mtro. Ing. Hugo Humberto Rivera Pérez

HONORABLE TRIBUNAL EXAMINADOR

En cumplimiento con los preceptos que establece la ley de la Universidad de San Carlos de Guatemala, presento a su consideración mi trabajo de graduación titulado:

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA

Tema que me fuera asignado por la dirección de la Escuela de Estudios de Postgrado, con fecha 10 de noviembre de 2022.

Ing. Pedro Pablo Morales Ortiz

Decanato
Facultad de Ingeniería
24189101- 24189102
secretariadecanato@ingenieria.usac.edu.gt

LNG.DECANATO.OI.039.2024

El Decano de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer la aprobación por parte del Director de la Escuela de Estudios de Posgrado, al Trabajo de Graduación titulado: **CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA,** presentado por: **Ing. Pedro Pablo Morales Ortiz,** que pertenece al programa de Maestría en artes en Estadística aplicada después de haber culminado las revisiones previas bajo la responsabilidad de las instancias correspondientes, autoriza la impresión del mismo.

IMPRÍMASE:

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DECANO a.i.
Facultad de Ingeniería

Ing. José Francisco Gómez Rivera

Decano a.i.

Guatemala, enero de 2024

JFGR/gaoc

Guatemala, enero de 2024

LNG.EEP.OI.039.2024

En mi calidad de Directora de la Escuela de Estudios de Postgrado de la Facultad de Ingeniería de la Universidad de San Carlos de Guatemala, luego de conocer el dictamen del asesor, verificar la aprobación del Coordinador de Maestría y la aprobación del Área de Lingüística al trabajo de graduación titulado:

"CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA"

presentado por **Ing. Pedro Pablo Morales Ortiz** correspondiente al programa de **Maestría en artes en Estadística aplicada** ; apruebo y autorizo el mismo.

Atentamente,

"Id y Enseñad a Todos"

UNIVERSIDAD DE SAN CARLOS DE GUATEMALA
DIRECTORA
ESCUELA DE ESTUDIOS DE POSTGRADO
FACULTAD DE INGENIERÍA

Mtra. Inga. Aurelia Anabela Cordova Estrada
Directora
Escuela de Estudios de Postgrado
Facultad de Ingeniería

Guatemala, 21 de octubre de 2023

M.A. Inga. Aurelia Anabela Cordova Estrada
Directora
Escuela de Estudios de Postgrado
Presente

Estimada M.A. Inga. Cordova Estrada

Por este medio informo a usted, que he revisado y aprobado el **INFORME FINAL y ARTÍCULO CIENTÍFICO** titulado: **CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA** del estudiante **Pedro Pablo Morales Ortiz** quien se identifica con número de carné **201403531** del programa de Maestria En Estadistica Aplicada.

Con base en la evaluación realizada hago constar que he evaluado la calidad, validez, pertinencia y coherencia de los resultados obtenidos en el trabajo presentado y según lo establecido en el ***Normativo de Tesis y Trabajos de Graduación aprobado por Junta Directiva de la Facultad de Ingeniería Punto Sexto inciso 6.10 del Acta 04-2014 de sesión celebrada el 04 de febrero de 2014.*** Por lo cual el trabajo evaluado cuenta con mi aprobación.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

UNIVERSIDAD DE SAN CARLOS
ESCUELA DE POSTGRADO
FACULTAD DE INGENIERIA
DE GUATEMALA

Msc. Ing. Edwin Adalberto Bracamonte Orozco
Coordinador
Maestria En Estadistica Aplicada
Escuela de Estudios de Postgrado

Guatemala, 21 de octubre de 2023

M.A. Inga. Aurelia Anabela Cordova Estrada
Directora
Escuela de Estudios de Postgrados
Presente

Estimada M.A. Inga. Cordova Estrada

Por este medio informo a usted, que he revisado y aprobado el Trabajo de Graduación y el Artículo Científico: **" CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA"** de el/la estudiante **Pedro Pablo Morales Ortiz** del programa de **Maestria En Estadistica Aplicada** identificado(a) con número de carné 201403531.

Agradeciendo su atención y deseándole éxitos en sus actividades profesionales me suscribo.

Ing. José Rolando Chávez Salazar
Ingeniero Industrial
Colegiado No. 4,317

Mtro. Ing. José Rolando Chávez Salazar
Colegiado No. 4317
Asesor de Tesis

ACTO QUE DEDICO A:

Mis padres	Por guiarme siempre por el buen camino, buscando mi superación personal
Mis abuelos	Por su constante apoyo en mi proceso de aprendizaje y motivación para continuar en los momentos más difíciles.
Helena Reyes	Por ser mi apoyo incondicional y mi fuente de motivación.
Mis amigos	Por brindarme siempre su ayuda y conocimientos durante nuestra etapa de estudios.

AGRADECIMIENTOS A:

Universidad de San Carlos de Guatemala	Por brindarme una educación de calidad y un segundo hogar.
Mi familia	Por brindar un entorno de confianza para mi desarrollo académico.
Ing. Rolando Chávez	Por su apertura a brindar asesoría en el proceso y los buenos consejos.
Ing. Carlos Beltetón	Por su apoyo en el desarrollo de este proyecto en su unidad de negocio.
Ing. William Fagiani	Por su incontable apoyo para la elección y calibración de los modelos más adecuados.
Mis catedráticos	Por brindarme su incontable conocimiento y consejos durante mi etapa estudiantil.

ÍNDICE GENERAL

ÍNDICE DE ILUSTRACIONES

FIGURAS

TABLAS

LISTA DE SÍMBOLOS

Símbolo	Significado
F	Frecuencia de compra
H1	Hipótesis alterna
H0	Hipótesis nula
M	Monto promedio
%	Porcentaje
R	Reciencia

GLOSARIO

Centroide Es la ubicación real o imaginaria que representa el centro de un grupo de datos. También puede considerarse como el centro geométrico en un espacio de k variables o dimensiones.

Clientes profesionales Son los clientes que, por su modelo de operación comercial, proveen servicios a otros clientes finales. Se encuentran identificados de esta forma en la base de datos de la empresa.

Clúster Son agrupaciones de datos o registros que comparten características o están relacionadas entre sí.

ERP Se refiere a un *software, enterprise resource planning,* que se traduce como sistema de planificación de recursos empresariales. Es un sistema que ayuda a administrar los procesos contables y comerciales de una empresa, registrando sus procesos de ventas, cadena de suministro, operaciones, recursos humanos, entre otros.

EPV Se refiere al término *events per value,* que, en el análisis de variables categóricas, se refiere a la cantidad de observaciones u eventos que hay en cada nivel o factor de la variable.

Machine learning	Es una rama de la inteligencia artificial que, a través de diferentes algoritmos, da la capacidad a sistemas de información para identificar patrones en datos masivos y elaborar predicciones.
Minería de datos	Es la exploración y análisis de datos, automático y semiautomático, que analiza grandes cantidades de información para descubrir patrones o reglas que sean significativas.
Multicolinealidad	Es la relación de dependencia lineal entre dos o más variables independientes en un proceso de regresión.
Quintiles	Son cuatro valores que dividen un conjunto de datos ordenados en cinco segmentos del mismo tamaño.
Residuo estandarizado	Es la transformación que se obtiene al dividir un valor residual de un modelo dentro de su desviación estándar estimada.
Reciencia	Es la cantidad de días transcurridos desde la última transacción del cliente.
RFM	Método de análisis del comportamiento de clientes basado en las variables reciencia, frecuencia y monto promedio de compras.

Variable de intervalo	Es una variable de tipo cuantitativo, donde los intervalos entre sus clases son iguales; sin embargo, el cero no implica el valor nulo de un atributo.
Variable de razón	Variable de tipo cuantitativo donde el cero sí indica la ausencia total de un atributo, por lo que sí es posible hacer razones en la medición.
Variable dicotómica	Es un tipo de variable cualitativa que solo puede tomar dos valores que son mutuamente excluyentes, denotando la ausencia o presencia de una característica.
Variables ordinales	Son variables cualitativas donde cada clase posee una misma relación posicional con la siguiente, por lo que muestra situaciones escalonadas.

RESUMEN

El presente informe tuvo como objetivo principal la construcción de un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales en una empresa minorista de materiales de construcción en Guatemala.

Para lograr este objetivo, se planteó una investigación de enfoque cuantitativo, con alcance descriptivo correlacional y diseño no experimental que extrajo datos transaccionales del ERP de la empresa de forma transversal.

En el desarrollo de la investigación se utilizaron herramientas estadísticas para segmentar a los clientes profesionales en grupos similares, identificar las variables que influyen en la retención de clientes y construir un modelo de regresión logística. Se utilizaron técnicas como el método de simulación por K medias y pruebas de independencia y correlación entre otras.

Los resultados obtenidos fueron estructurados en función de los objetivos establecidos. Se logró una segmentación efectiva de los clientes profesionales utilizando variables como reciencia, frecuencia y monto promedio de compra. Además, se identificaron las variables que tienen mayor influencia en la retención de clientes. El modelo de regresión logística desarrollado permitió cuantificar las probabilidades de pérdida y compra de cada cliente con una exactitud máxima del 69.39 %.

La conclusión principal destaca la utilidad del modelo de regresión logística como una herramienta para la toma de decisiones estratégicas; usando

como base de predicción, las variables de comportamiento y segmentación de los clientes en la empresa minorista de materiales de construcción. Se recomienda a la empresa utilizar el modelo y los hallazgos de esta investigación para mejorar sus estrategias de retención y fidelización de clientes.

PLANTEAMIENTO DEL PROBLEMA

- Contexto general

El mantenimiento de una cartera de clientes es una de las labores de mayor impacto dentro de una organización. Sin embargo, para que esto sea efectivo, el seguimiento constante de estos clientes debe hacerse de forma estratégica, buscando predecir su comportamiento para aprovechar los momentos donde estos son más propensos a comprar, y determinar cuando están en riesgo de perderse.

El modelo de negocio a estudiar es el de una empresa minorista de productos de construcción, con base de operaciones en la ciudad de Guatemala. Por el tipo de productos que la empresa comercializa, la clientela se suele segmentar en clientes profesionales y minoristas. Los primeros son clientes que, por la naturaleza de su negocio, suelen tener una relación comercial de largo plazo con la empresa. Por lo tanto, son el segmento en el que la empresa tiene más oportunidades de construir lealtad.

Se ha identificado que hay una brecha de información sobre el comportamiento de los clientes y su frecuencia de compras, que impide hacer un análisis profundo de las estrategias de generación de lealtad. Dicha situación causa que el área estratégica de la empresa no pueda predecir los momentos críticos de los ciclos de compra de sus clientes. Por ello, todos los clientes son tratados de la misma forma sin discriminar variables útiles, como la probabilidad de compra o pérdida del cliente, su región o su segmento de mercado.

- Descripción del problema

Dado que no se cuenta con estudios válidos sobre el comportamiento de los clientes profesionales, en la empresa no es posible desarrollar herramientas estadísticas que permitan impulsar técnicas predictivas sobre el comportamiento de los clientes actuales. Por este motivo, se necesita conocer cuál es la probabilidad de compra y pérdida de cada cliente profesional, tomando en consideración las variables cuantitativas y cualitativas de cada consumidor.

Por otro lado, no se ha estudiado con profundidad qué variables de segmentación de los clientes tienen una correlación significativa con la compra o pérdida de los clientes, y cuáles de estas están relacionadas entre sí. Este vacío de información existe para las variables de clasificación de los clientes, que son categóricas, como las variables cuantitativas de reciencia, frecuencia y monto de compras de cada cliente.

- Formulación del problema

A partir de la descripción del problema anterior, es posible identificar las preguntas generadoras que permiten la formulación de la investigación y su esquema de solución.

  - Pregunta central

¿Cómo se comporta la probabilidad de retención de los clientes profesionales, en función de sus variables cuantitativas y cualitativas?

- Preguntas auxiliares

    - ¿Cuáles son las agrupaciones óptimas de los clientes, en función de sus variables reciencia, frecuencia y monto promedio (RFM)?

    - ¿Qué variables de los clientes profesionales provocan una variación significativa en la retención o pérdida de los clientes?

    - ¿Cuál es el modelo óptimo para describir la probabilidad de retención o pérdida de los clientes profesionales?

- Delimitación del problema

El problema abarca al segmento de clientes que están categorizados como clientes profesionales en Guatemala, que cuenten con un mínimo de dos transacciones en los últimos dos años. Todos aquellos clientes profesionales que no cumplan con este parámetro serán considerados como cartera inactiva.

OBJETIVOS

General

Construir un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales, en una empresa minorista de materiales de construcción en Guatemala.

Específicos

1. Agrupar a los clientes profesionales en segmentos similares, basado en las variables de reciencia, frecuencia y monto de compras, aplicando métodos de simulación por K Medias.

2. Identificar las variables de clientes que interfieren en la pérdida o retención de clientes, usando pruebas de independencia y pruebas de correlación.

3. Relacionar las variables cuantitativas y cualitativas de los clientes profesionales, para construir un modelo de regresión logística que permita cuantificar las probabilidades de pérdida y compra de cada cliente profesional.

RESUMEN DEL MARCO METODOLÓGICO

A continuación, se muestran las características metodológicas de la presente investigación.

- Características del estudio

El enfoque del estudio fue cuantitativo, debido a que se tomaron mediciones de variables continuas por medio del método RFM, y estas se combinaron con otros factores categóricos, como el método de pago, sucursal de origen, canal y otras variables de clasificación de los clientes para estimar la probabilidad de retención de estos.

El alcance del proyecto fue descriptivo, pues el presente trabajo de investigación requirió la construcción de un modelo estadístico funcional, para predecir la retención o pérdida de clientes profesionales. Además, fue de tipo correlacional porque se analizó la relación entre las variables regresoras con la variable de respuesta.

El diseño adoptado fue de tipo no experimental (observacional), ya que los datos para el estudio fueron obtenidos por medio de minería de bases de datos. La información del comportamiento de los clientes se analizó en su estado original sin ninguna manipulación; además, fue transversal, pues se hizo un análisis de los datos históricos transaccionales de los clientes en un punto en específico, sin evaluar la evolución de estos.

- Unidades de análisis

La población en estudio se constituyó por los clientes profesionales de la empresa minorista de materiales de construcción, de la cual se extrajo la totalidad de la población de la base de datos transaccional para su estudio y estimación.

- Variables

A continuación, se muestra el detalle de la operativización de las variables, con su correspondiente definición a nivel operativo.

Tabla 1.

Operativización de variables

Variable	Definición teórica	Definición operativa
Estado del cliente	Es la variable de respuesta del modelo. Define si el cliente se perdió o se retuvo en un periodo determinado. De tipo cuantitativa dicotómica.	Si el cliente tuvo al menos una compra desde la fecha de referencia se consideró como cliente activo. Por el contrario, se designó como perdido.
Reciencia (R)	Es la cantidad de días transcurridos desde la última transacción del cliente. Variable de tipo cuantitativa de razón.	Diferencia entre la fecha de cálculo y la última fecha de transacción, medido en días.
Frecuencia (F)	Es la cantidad de días promedio que hay entre cada transacción. Variable de tipo cuantitativa de razón.	Se definió como los días promedio entre cada factura, medido en días
Monto promedio (M)	Es la cantidad promedio que gastó un cliente en cada visita. Variable de tipo cuantitativa de razón.	Se calculó como el promedio del total de las facturas emitidas al cliente, será medido en la divisa local.

Continuación de la tabla 1.

Variable	Definición teórica	Definición operativa
Canal	Es el medio del que se captó al cliente en su primera compra, con 3 opciones: *retail,* proyectos y digital. Variable categórica nominal.	Se obtuvo la variable mediante la agrupación de sucursales por canal. Puede considerarse como una simplificación de la variable sucursal.
Condición de pago	Es la forma en la que el cliente paga por los productos obtenidos en cada transacción. Variable de tipo categórica con escala nominal.	Se extrajo de la base de datos el tipo de pago asociado a cada cliente.
Cliente multicategoría	Mide si el cliente está asociado a la empresa transaccionando solo un tipo de productos o múltiples categorías de estos. Esta variable está definida como dicotómica.	Se hizo un conteo de las categorías compradas por el cliente en la ventana de tiempo. Si este era mayor que uno, se consideró como cliente multicategórico.
Forma de entrega	Es la forma en la que el cliente recibe su producto, pudiendo tener 3 diferentes valores: ruta, recolección en tienda y múltiples. Variables de tipo cualitativo nominal.	Se extrajo la forma de entrega predeterminada de cada cliente por medio de la minería de datos.
Antigüedad de cotización	Mide cuántos días han pasado desde la última cotización realizada al cliente. Variable de tipo cualitativa de razón.	Se calculó como la diferencia, en días, entre la última cotización y la fecha de cálculo.

Nota. Listado de variables a utilizar, con su correspondiente definición teórica y definición operativa. Elaboración propia, realizado con Word.

- Fases del estudio

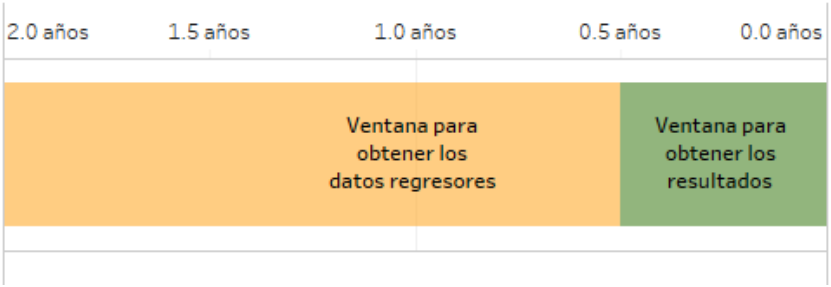
El presente estudio se dividió en varias etapas que se describen a continuación.

En la primera fase, se llevó a cabo una exhaustiva revisión de literatura. Se recopiló bibliografía relevante para fundamentar la temática de la investigación, centrándose en conocimientos estadísticos como el análisis de correlación y la construcción de modelos de regresión logística. También se incluyó una sección dedicada a la temática específica de la empresa y al análisis del comportamiento de los clientes.

La segunda fase se centró en la minería y limpieza de datos. La empresa proporcionó acceso a sus bases de datos transaccionales, lo que permitió realizar tareas de minería de datos utilizando Microsoft SQL Server. Se extrajeron las variables de segmentación y comportamiento de los clientes a partir de los documentos transaccionales del período comprendido entre 2021 y 2022.

Figura 1.

Ventana de extracción de información histórica



Nota. Detalle de los periodos de tiempo utilizados para la extracción de la información transaccional de las bases de datos. Elaboración propia, realizado con Tableau.

Esta fase se basó principalmente en la observación directa, recopilando información de manera no participante a través del análisis de los datos transaccionales de la empresa. Se creó un modelo de entrenamiento utilizando las transacciones de los últimos seis meses para medir el comportamiento de la variable objetivo (retención o pérdida del cliente) y las transacciones de los 18 meses anteriores para obtener las variables regresoras. La figura anterior describe las ventanas temporales usadas para crear el modelo de entrenamiento.

La tercera fase consistió en el análisis de correlación entre las variables regresoras. Se utilizaron los datos extraídos mediante la minería de datos para analizar la interacción entre estas variables. Se calcularon coeficientes de correlación entre las variables cuantitativas obtenidas mediante el método RFM para evaluar su grado de interacción. Además, se realizaron pruebas de independencia entre las variables regresoras y la variable dependiente para determinar su asociación y seleccionar las variables adecuadas para la construcción del modelo.

La cuarta fase se enfocó en la construcción del modelo de regresión. Se seleccionaron las variables significativas que demostraron tener un efecto en la variable de respuesta durante el análisis de correlación. Se construyeron múltiples modelos de regresión logística binomial utilizando *software* R. También se realizó un análisis de multicolinealidad para identificar aquellas variables que aportaban poca información al modelo de datos.

La quinta fase consistió en el análisis de los resultados obtenidos en cada modelo de regresión. Se compararon los resultados utilizando el criterio de información de Akaike (AIC) y la exactitud de predicción medida a través de matrices de confusión. Se obtuvieron valores predichos a partir de los modelos construidos en la fase anterior para determinar la probabilidad de retención de

los clientes. Si la probabilidad calculada por el modelo era mayor al 50 %, se consideraba que el modelo predecía la retención del cliente.

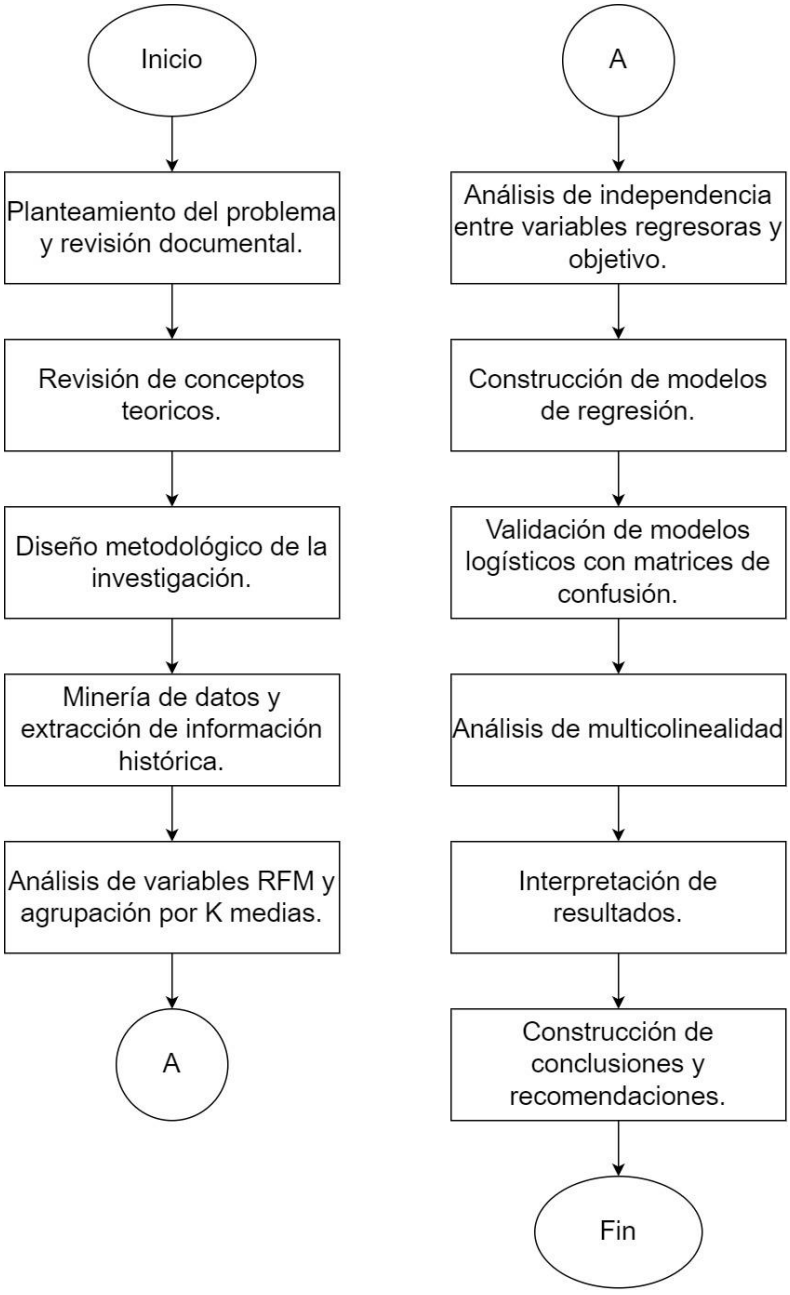
En la sexta y última fase, se redactó el informe final y se presentaron los resultados de la investigación. Se incluyeron conclusiones, recomendaciones y el informe final basado en los hallazgos del estudio.

- Flujograma del proceso de investigación

A continuación, se detalla el flujograma que describe las fases del proceso de desarrollo e investigación.

Figura 2.

Flujograma del proceso de investigación



Nota. Se detallan las etapas del proceso de investigación y redacción de informe final. Elaboración propia, realizado con diagrams.net.

INTRODUCCIÓN

El presente estudio consistió en la sistematización del proceso de análisis y segmentación de clientes profesionales mediante el cálculo de la probabilidad de retención, para su posterior tratamiento en la empresa minorista de materiales de construcción en Guatemala.

En un análisis preliminar, se determinó que la empresa no tenía la capacidad de desarrollar técnicas predictivas sobre el comportamiento de los clientes, y tampoco se había estudiado con profundidad qué variables de segmentación de estos tienen una correlación significativa con la compra o pérdida de los clientes.

La importancia del estudio radica en la necesidad de la empresa de generar acciones comerciales para su crecimiento sostenido, donde contar con este tipo de información permite diseñar estrategias de personalización masiva que son de gran apoyo a la fidelización de cartera de clientes y a la atención de las necesidades de servicio en aquellos que lo requieran, incrementando la satisfacción del cliente y la rentabilidad de la empresa.

La metodología de la investigación fue de enfoque cuantitativo, con diseño no experimental u observacional y un alcance descriptivo correlacional.

El resultado del presente estudio permitió a la empresa disponer de una estimación puntual automatizada, mediante un modelo de regresión logística, de la probabilidad de retención de cada uno de sus clientes profesionales. Para su segmentación e identificación dentro de los sistemas de información.

El desarrollo del estudio estuvo compuesto por seis fases. En la primera fase se hizo una recopilación bibliográfica que fue de utilidad para fundamentar los temas a analizar en esta investigación. En la segunda se hicieron tareas de minería de datos para extraer las variables de segmentación y de comportamiento de los clientes, utilizando documentos transaccionales del 2021 al 2022. La tercera fase del estudio consistió en un análisis de correlación para inferir si cada variable tiene un efecto significativo en la retención o pérdida de los clientes profesionales. La cuarta fase tomó las variables que durante el análisis de correlación probaron ser significativas en su efecto sobre la variable de respuesta, para construir un modelo de regresión logística binomial. Por último, se evaluó la multicolinealidad y exactitud de los modelos construidos, para iterar nuevamente consiguiendo múltiples modelos que se ajusten a los datos presentados. La quinta fase del estudio consistió en el análisis de los resultados obtenidos para elegir el modelo de regresión con mejor ajuste. La última fase del estudio permitió presentar los resultados obtenidos de la investigación, así como la redacción de conclusiones, recomendaciones e informe final.

La factibilidad del estudio fue adecuada, ya que se contó con todos los recursos tecnológicos, financieros y de información necesarios para la realización de las fases descritas anteriormente.

El informe final está constituido por los siguientes capítulos:

En el primer capítulo hizo un análisis del marco referencial, que incluyó otros estudios referentes a la estimación de probabilidades en la cartera de los clientes, análisis del comportamiento de estos y su segmentación.

En el segundo capítulo se detalla el marco teórico compuesto por dos partes. La primera se compone de todos los fundamentos, teoremas y métodos

estadísticos que sustentaron la investigación. La segunda consiste en las definiciones y conceptos que complementaron la aplicación en el ámbito profesional.

En el tercer capítulo se presentan los resultados estructurados por objetivos. Inicialmente se analizan las variables reciencia, frecuencia y monto promedio de compra para su segmentación adecuada. En la segunda parte de la presentación de resultados se hace un análisis de correlación utilizando pruebas independencia de datos. Por último, se concluyó la presentación de resultados con el análisis de los múltiples modelos de regresión logística para profundizar en su exactitud y ajuste, se eligió el modelo que mejor describe a los datos.

En el cuarto capítulo se discuten los resultados de la investigación. En principio, se hizo un análisis interno para resaltar la validez de los resultados y permitió la generalización de algunos conceptos. Posteriormente, se contrastaron los resultados con otras investigaciones de la misma rama de aplicación, para validar los criterios de validez interna y externa.

Finalmente, se presentan las conclusiones y recomendaciones obtenidas a partir de los resultados. Se resumen los hallazgos clave, se discuten las implicaciones y se destacan las contribuciones de la investigación al campo de estudio.

1. MARCO REFERENCIAL

Todos los pasos y movimientos que dan las empresas dejan un rastro digital dentro de los sistemas de planificación de recursos (ERP). Si estos datos son ordenados y transformados para mostrar información relevante, es posible facilitar la toma de decisión para la organización. Esto mismo aplica con las decisiones sobre los clientes, donde las técnicas de segmentación y predicción son cada vez más acertadas.

Un estudio realizado por Aleksandrova (2018) sobre una empresa de fabricación de concreto, utilizó los datos transaccionales del ERP de la empresa para extraer variables de reciencia, frecuencia y monto de compra de sus diferentes clientes, y así calcular la probabilidad de pérdida de los clientes. Para determinar que un cliente se perdió, analizaron sus ciclos de compra, determinando que para ese mercado la inactividad de seis meses o más indica que pertenece a dicha categoría (clientes perdidos).

Basados en esta ventana temporal, dividieron los periodos de tiempo en dos. Para cada cliente calcularon la reciencia, frecuencia y monto de compra que tuvieron previo a los últimos seis meses, y determinaron por último si los clientes habían estado activos o inactivos en los últimos seis meses. De esta forma, estimar la probabilidad de pérdida de los clientes seis meses atrás, y validar este modelo con el periodo de tiempo que le procedió. Dado que la pérdida o retención de un cliente es una opción binaria, se podría considerar que la probabilidad de pérdida de un cliente es complementaria a su probabilidad de compra, lo que podría facilitar su estimación. Aun así, la investigadora hace la recomendación

de buscar otras variables que permitan la segmentación de los clientes para hacer estimaciones más precisas.

La investigación mencionada anteriormente fue utilizada en la investigación como base para construir el modelo de entrenamiento, en el proceso de minería de datos, se utilizó una ventana temporal de respuesta de 6 meses para determinar si un cliente fue retenido o perdido.

Por otro lado, en el estudio de Cuadros, Gonzales y Jiménez (2017) se buscó hacer una segmentación de cartera de clientes, basada en parámetros de reciencia, frecuencia y monto de compra por medio de la aplicación de técnicas de análisis estadístico multivariado. Adicional a las variables clásicas del modelo RFM, se extrajeron los datos de utilidades, margen neto y días de crédito vencidos. En el proceso de minería de datos se procedió a normalizar los datos para comparar las variables en la misma escala de medición, para medir la covarianza, colinealidad y correlación.

Aunque algunas variables cualitativas relevantes no fueron consideradas en el análisis, el estudio proporcionó una base sólida para la segmentación de la clientela actual de la empresa. Además, este estudio sirvió como punto de partida para explorar otras variables cualitativas que pudieran contribuir a la comprensión del comportamiento de los clientes. En este sentido, la inclusión de la variable días de antigüedad de la última cotización como variable regresora resultó en una mejora significativa en los modelos desarrollados.

En un estudio reciente sobre la segmentación avanzada de clientes, de Yoseph y AlMalaily (2019), se hizo el análisis de segmentación de clientela en una empresa *retail* de Malasia. El autor afirma que basarse únicamente en las variables tradicionales de segmentación, que provienen de aspectos

demográficos, suele ser una fuente de errores en mercadeo, ya que no describe en su totalidad a los clientes. Por ello, también se requiere llevar a cabo un análisis de los aspectos de comportamiento de los clientes. Para el estudio de segmentación, los autores tomaron los datos de frecuencia y monto de compra para clasificar a los clientes en cuatro segmentos: frecuentes, de alto gasto, mejores clientes e inciertos. Sin embargo, por el modelo de negocio que se describe en el estudio, este estudio no tomó en consideración la variabilidad que tiene la reciencia de compra. Para tomar en cuenta una tercera variable, este tipo de segmentación resulta en una agrupación de al menos ocho estratos, que es muy numerosa para desarrollar estrategias de mercadeo.

El mencionado estudio brindó una perspectiva profesional a la presente investigación al destacar una de sus premisas fundamentales: las variables de segmentación clásicas proporcionan información limitada sobre el comportamiento del cliente, aunque no deben descartarse por completo. Por lo tanto, se procedió a realizar una minería de datos exhaustiva para extraer todas las variables de comportamiento posibles.

Otro enfoque utilizado por Dogan, Aycin & Bulut (2018) para solucionar la problemática de una óptima segmentación de clientes, basada en su comportamiento, es la aplicación de un algoritmo de clasificación de K medias. Este algoritmo hace simulaciones de clasificación para obtener las agrupaciones que disminuyan la variación total. Los autores utilizan este método con un *set* de datos de clientes y sus valores de RFM. Finalmente, determinaron cuál disposición redujo la variación total en los datos e identificaron qué registros pertenecen a cada estrato. Los resultados del estudio, que fue aplicado a un comercio minorista, reflejaron que con tres estratos se redujo el error total al agruparlos. La metodología descrita puede ser útil para la construcción de modelos predictivos, pues convierte tres variables cuantitativas en una sola

variable categórica. Sin embargo, se debe validar si este tipo de agrupación tiene un mejor ajuste a la realidad. comparado con la regresión de las tres variables por separado.

El estudio previo realizado por Dogan, Aycin & Bulut (2018) se utilizó como referencia para la segmentación basada en variables de comportamiento RFM desde una perspectiva de análisis de datos multivariados. Gracias a este enfoque en el ordenamiento de los datos, fue posible obtener segmentos de clientes más homogéneos en términos de comportamiento, los cuales también resultaron ser variables predictivas efectivas para el modelo en cuestión.

Jain, Khunteta & Srivastava (2020) efectuaron un estudio experimental en una empresa de telecomunicaciones de Estados Unidos, donde compararon diferentes modelos, dos modelos predictivos de la probabilidad de pérdida de sus clientes. Para ello, extrajeron información directamente de la plataforma de gestión de relaciones de los clientes (CRM), y la dividieron aleatoriamente en dos muestras. Una para trabajar por medio de *machine learning* y la otra, por medio de una regresión logística binaria, donde 0 es para los clientes que siguieron activos, y 1 para los que se perdieron.

El resultado de esta comparación de métodos fue similar para ambos casos; con un 85.23 % de exactitud para la regresión logística y 85.17 % de exactitud para los algoritmos de aprendizaje cerrados. Esta diferencia podría considerarse despreciable, pero ambos modelos podrían ser válidos para la predicción. Sin embargo, por la complejidad y costos de las plataformas de inteligencia artificial, en ocasiones se recomienda obtener la información de los métodos más sencillos, como lo es la regresión logística binaria. También se debe mencionar que el modelo de negocio sobre el que se realizó este estudio es manejado por contratos fijos con la clientela, con lo cual se tiene información

inmediata del momento de pérdida de un cliente; en el caso de la empresa de materiales de construcción, esto es un valor fijo de referencia (Jain, Khunteta & Srivastava, 2020).

A pesar de que el estudio mencionado se llevó a cabo en un ámbito de negocio totalmente diferente al de la presente investigación, proporcionó una comparativa real entre métodos de clasificación y predicción, lo cual permitió seleccionar la regresión logística binaria cómo el modelo más adecuado para explicar la retención de clientes.

De la misma forma, en el estudio de Hargreaves (2019) se toma la pérdida o retención del cliente como una variable binaria para la construcción de un modelo binario de regresión logística. Este modelo fue elegido por ser considerado un modelo de clasificación sencillo con buenos resultados, pues en el diagnóstico el modelo obtuvo un 76.7 % de precisión. En dicho estudio se aplicó una combinación de 20 variables cualitativas y cuantitativas. Sin embargo, destacan que la principal suposición es que todas las variables independientes no tienen una multicolinealidad significativa, por lo que hace la recomendación de llevar a cabo un análisis de regresión para las variables cuantitativas, y de independencia con las variables a un nivel de significancia adecuado.

Tomando como referencia la recomendación de Hargreaves (2019), en el presente estudio se llevó a cabo un análisis preliminar de las variables predictoras de comportamiento, examinando su interdependencia y correlación. Asimismo, se analizó la multicolinealidad de los modelos ajustados para evitar la influencia de la covarianza de otras variables predictoras y minimizar la variabilidad causada por las relaciones entre las variables regresoras.

En esta línea, Senaviratna y Cooray (2019) hicieron un estudio donde buscaron diagnosticar y optimizar la multicolinealidad de un modelo de regresión logística. Los autores utilizaron el coeficiente de Pearson (r) y el factor de inflación de la varianza (VIF) para identificar las variables que están relacionadas entre sí. Destacan que, en ese momento, no había posibilidad de aumentar el tamaño de la muestra para corregir la multicolinealidad, por lo que se procedió a eliminar variables relacionadas. Para ello eliminaron la variable con mayor VIF y analizaron nuevamente la multicolinealidad; posteriormente, repitieron estos pasos hasta tener una agrupación de variables independiente.

El enfoque propuesto por Senaviratna y Cooray (2019) se utilizó en la fase de análisis de la multicolinealidad de los mejores modelos. Se realizaron cambios iterativos en las variables predictoras con el objetivo de obtener una combinación adecuada que proporcionara precisión, minimizará la inflación de la varianza y presentara un coeficiente de Pearson apropiado.

Boateng & Abaye (2019) realizan una revisión de las recomendaciones para aplicar una regresión logística, discutiendo las mejores prácticas para la elección de variables independientes. En este, indican que el hecho de incluir una variable adicional a un modelo de regresión logística puede incrementar su exactitud al explicar cierta parte de la variación. Sin embargo, indican también que se debe ser muy cuidadoso para no incluir todas las variables que existan, pues hay posibilidades de que algunas de estas variables no tengan una correlación significativa con la variable de respuesta; esto tiende a inflar la validez aparente del modelo construido. Se recomienda ampliamente la revisión literaria de la variable a estudiar, como uno de los métodos para facilitar la elección de las variables independientes.

Un antecedente importante en la elección de las variables es el estudio de Van *et. al.* (2018), indican que el tamaño de la muestra para regresiones logísticas se debe analizar en función del número de eventos u observaciones por valor (EPV), y que hay una tendencia a asumir que si la variable tiene EPV>10 es útil para la construcción de modelos. En el análisis, los autores obtuvieron errores de calibración de sus modelos cuando las variables contaban con menos de 10 eventos por variable. También observaron una mejora de la exactitud de los modelos al aumentar el número de EPV, pero no obtuvieron una mejora de sus resultados más allá de 20 EPV.

Estos estudios destacan la importancia de tener un conjunto de variables predictoras (tanto cualitativas como cuantitativas) que sean independientes entre sí, pero que al mismo tiempo estén adecuadamente correlacionadas con la variable de respuesta. Durante el proceso de análisis de correlación, fue fundamental contar con los criterios proporcionados para seleccionar adecuadamente las variables.

En un estudio sobre la pérdida de clientes en el sector de telecomunicaciones de Pakistán de Ullah *et. al.* (2019), se construyeron varios modelos de *machine learning* para predecir la pérdida de clientes. Para determinar cuál de los modelos construidos presentaba un mejor ajuste con los clientes, se realizó una matriz de desempeño, donde compararon la exactitud, precisión, tasa de positivos y tasa de falsos positivos de cada predicción. Estas mediciones fueron obtenidas a través una matriz de confusión, donde se consideró que, si la predicción de pérdida del cliente es de 50 % o más, el modelo los clasificaba como clientes a perderse, de lo contrario, serían considerados clientes a retenerse. Al comparar esta clasificación con los datos reales, midieron cuántas predicciones fueron correctas, la cantidad de falsos negativos y la cantidad de falsos positivos.

El enfoque descrito por los autores en este estudio permitió crear una matriz de confusión para cada modelo ajustado, lo que proporcionó información sobre su exactitud y el riesgo de predicción (expresado como el porcentaje de falsos positivos determinado por el modelo). Esto permitió determinar el modelo de mejor ajuste a los resultados reales de los clientes.

El estudio descrito aplicó conceptos de informática fuera del alcance, pero se puede considerar válida la técnica de comparación de modelos. Para ello, fue necesario hacer una segmentación en la técnica de minería de datos como la descrita en el estudio de Aleksandrova (2018).

La revisión de estudios realizados, principalmente en otros países, evidenció el alto desarrollo que existe en el área analíticas avanzadas de clientes, para desarrollar modelos predictivos en otros modelos de negocio. Para el presente estudio, fueron utilizadas para la construcción de un esquema metodológico y para la construcción de un esquema de solución estadística.

2. MARCO TEÓRICO

2.1. Estadística

Del Castillo y Salazar (2018) definen la estadística como la ciencia que se encarga de recolectar, ordenar, representar, analizar e interpretar datos generados en una investigación sobre hechos, individuos o agrupaciones de estos, para realizar conclusiones o estimaciones precisas.

2.1.1. Estadística descriptiva

Los mismos autores indican que la estadística descriptiva, también conocida como matemática, es aquella que permite analizar un conjunto de datos, extrayendo conclusiones que son válidas únicamente para este conjunto de observaciones. Por ello, con los estudios realizados con este tipo de análisis, únicamente es posible describir los resultados que se obtuvieron.

2.1.2. Estadística inferencial

Por otro lado, se define a la estadística inferencial como la rama que pretende obtener estimaciones generales de una población, mediante el estudio de una muestra proveniente de la población. Para este proceso, es necesario conocer los conceptos de valores estadísticos y parámetros; inicialmente, se calculan los valores descriptivos de la muestra, obteniendo estadísticos referentes al conjunto de datos, con lo que se busca estimar los parámetros que describen a la población general.

2.1.3. Análisis de normalidad

La normalidad estadística es una premisa fundamental en el análisis de datos, ya que es esencial para aplicar muchas pruebas estadísticas. En un contexto de distribución normal, los datos se distribuyen simétricamente alrededor de su media, lo que permite realizar inferencias precisas y confiables. Las pruebas de normalidad evalúan si un conjunto de datos sigue una distribución normal, ayudando a determinar si se pueden aplicar métodos paramétricos o si se requieren enfoques no paramétricos.

La normalidad de los datos se puede corroborar de dos maneras fundamentales: la primera, usando métodos gráficos como el gráfico cuantil-cuantil; y la segunda, aplicando pruebas estadísticas, como la prueba de normalidad de Kolmogórov-Smirnov y la prueba de normalidad de Shapiro-Wilk.

2.1.3.1. Gráfico cuantil-cuantil (Q-Q)

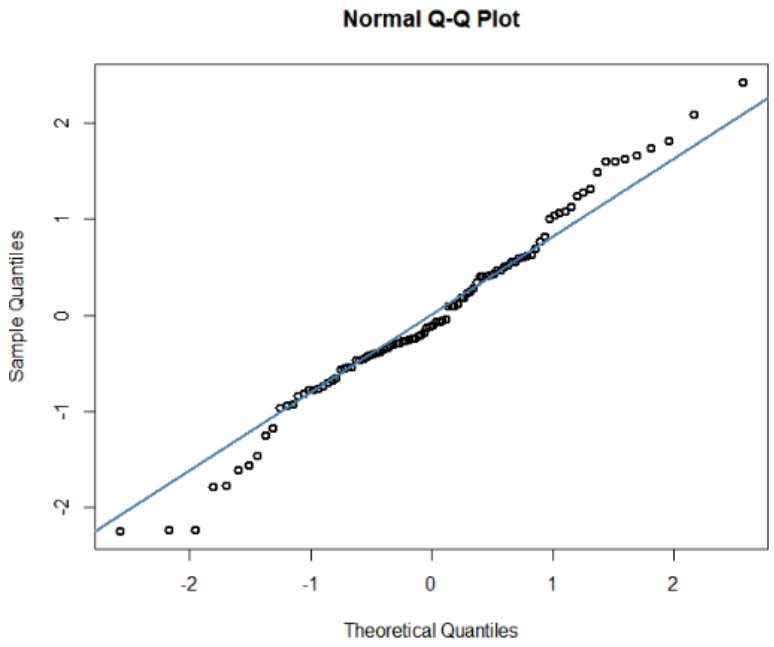
Para Castillo y Damian (2007), el gráfico cuantil-cuantil (Q-Q) es una herramienta gráfica utilizada en estadísticas para evaluar la similitud entre la distribución de los datos observados y una distribución teórica, a menudo la distribución normal. Su objetivo es determinar si los datos siguen una distribución específica o si presentan desviaciones significativas.

En un gráfico Q-Q, los cuantiles de los datos observados se representan en el eje vertical, mientras que los cuantiles esperados de la distribución teórica se trazan en el eje horizontal. Si los puntos en el gráfico siguen aproximadamente una línea recta, sugiere que los datos se ajustan a la distribución teórica. Las desviaciones de esta línea indican diferencias en la forma de la distribución.

Este gráfico es particularmente útil para detectar desviaciones en los valores extremos, lo que puede tener un impacto significativo en los análisis estadísticos. Además de evaluar la distribución normal, el gráfico Q-Q también puede revelar la adecuación de los datos a otras distribuciones, como la exponencial o log normal.

Figura 3.

Gráfico cuantil – cuantil (Q-Q)



Nota. Gráfico cuantil – cuantil de una serie de 100 observaciones muestrales. Elaboración propia, realizado con Rstudio.

La figura 3 muestra un ejemplo de gráfico cuantil – cuantil que compara datos muestrales con una distribución normal. En el eje horizontal se muestran los cuantiles teóricos, y en el eje vertical se colocan los cuantiles muestrales.

Asimismo, la línea diagonal representa la distribución teórica, mientras que los puntos representan los datos muestrales.

2.1.3.2. Prueba de normalidad Kolmogórov-Smirnov

Para Flores y Flores (2021), la prueba de Kolmogórov-Smirnov es una prueba de bondad de ajuste, que tiene un amplio uso en la comprobación de normalidad en datos muestrales. Se basa en el planteamiento de una hipótesis nula que afirma que la distribución empírica de los datos es igual a una distribución teórica (a menudo, la distribución normal). Dichas hipótesis se plantean de la siguiente forma:

$$H_0: F_n(x) = F(x)$$
$$H_a: F_n(x) \neq F(x) \quad \text{(Ec. 1)}$$

Donde $F_n(x)$ es la distribución muestral o distribución empírica, mientras que $F(x)$ es la distribución teórica. Por lo tanto, no rechazar la hipótesis nula implica que ambas distribuciones son estadísticamente iguales.

El análisis de la prueba descrita se realiza comparando la función de distribución muestral acumulada con la distribución esperada de la función teórica Si la diferencia, que se expresa en términos del valor absoluto observada es suficientemente grande, se rechaza la hipótesis que implica la normalidad de la población. El estadístico de prueba D_n se calcula utilizando la ecuación 2.

$$|D_n| = max|F_n(x) - F(x)| \quad \text{(Ec. 2)}$$

En dicha ecuación, el operador max indica que es necesario obtener la máxima diferencia entre las funciones de distribución acumuladas teóricas y

muestrales. El estadístico de prueba puede compararse con los valores críticos dados en el anexo 1; si el estadístico de prueba es mayor que el valor crítico al nivel de confianza α fijado, se procede a rechazar la hipótesis nula.

2.1.3.3. Prueba de normalidad Shapiro-Wilk

Dietrichson (2019) afirma que la prueba de hipótesis de Shapiro-Wilk, es una prueba de bondad de ajuste para medir el ajuste de una serie de datos muestrales con la distribución normal, utilizada en muestras menores o iguales de 50 observaciones. Esta prueba puede aplicarse a muestras de gran tamaño mediante modificaciones propuestas por Royston en 1982 utilizando algoritmos en *software* especializado en estadística; produciendo resultados similares a los obtenidos con la prueba Kolmogórov-Smirnov.

Para ello, se plantea una hipótesis nula que afirma la aproximación de los datos muestrales a la distribución normal, como la descrita en la ecuación 3.

$$H_0: F_n(x) = N(\mu, \sigma)$$
$$H_a: F_n(x) \neq N(\mu, \sigma) \qquad \text{(Ec. 3)}$$

El estadístico de prueba, W, que permite analizar el rechazo o aceptación de las hipótesis planteadas anteriormente, se calcula utilizando la siguiente ecuación:

$$W = \frac{[\sum_{i=1}^{m} a_i * (x_{n+1-i} - x_i)]^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \qquad \text{(Ec. 4)}$$

Donde $m = n/2$, x_i y x_{n+1-i} se obtienen ordenando los datos muestrales de forma ascendente y a_i es un factor obtenido de la tabla de Shapiro-Wilk mostrada en el anexo 2.

El resultado obtenido del estadístico de prueba W se compara con la tabla de valores de P presentada en el anexo 3; interpolando a los valores más cercanos según sea el caso. De esta forma, es posible contrastar el estadístico de prueba con un valor de significancia dado.

2.1.4. Correlación entre variables

La construcción de modelos de regresión implica elegir adecuadamente las variables independientes que intervienen en el fenómeno estudiado. Para ello, se requiere conocer qué tan relacionadas están las variables independientes con la variable de estudio. La técnica de correlación se encarga de determinar cuál es el grado de asociación que existe entre dos variables, ya sea positiva o negativa, como indican Badii *et. al.* (2014). Este análisis puede realizarse entre variables cuantitativas y cualitativas mediante diferentes métodos que se describen a continuación.

Cuando se determina la correlación entre variables, se debe prestar especial atención al tipo de variables que se examinan, pues según Hernández *et. al.* (2018), la validez de las conclusiones que se extraen en una inferencia estadística depende del cumplimiento de los supuestos bajo los cuales se han construido los modelos.

2.1.4.1. Coeficiente de Pearson

También conocido como R de Pearson, es el método clásico para medir la correlación entre variables cuantitativas que están relacionadas de forma lineal. Con él, se obtiene la magnitud, sentido y significación de la asociación entre dos variables intervalo (una dependiente y otra independiente). Las muestras de estas variables deben ser obtenidas con muestreo aleatorio y los datos se deben presentar de forma separada (para cada valor de la variable independiente, hay un valor de la variable que a priori se define como dependiente).

Para que los resultados del análisis de correlación por medio del coeficiente de Pearson sean válidos, los datos deben cumplir con varios supuestos que se conocen por medio de un análisis previo.

2.1.4.1.1. Normalidad

La evaluación de la normalidad es un requisito vital al aplicar el coeficiente de correlación de Pearson en el análisis de datos. Para verificar la normalidad, se recurre a pruebas estadísticas como el test de Kolmogórov-Smirnov o el test de Shapiro-Wilk descritos anteriormente, además de gráficos de cuantiles normales. Si los datos no siguen una distribución normal, puede ser necesario considerar transformaciones o emplear correlaciones no paramétricas.

Aunque el coeficiente de Pearson puede ser resistente a leves desviaciones de la normalidad con tamaños de muestra grandes, es fundamental evaluar su presencia y cómo podría impactar la interpretación de los resultados, asegurando así la validez de las conclusiones obtenidas de la correlación.

2.1.4.1.2. Ausencia de datos atípicos

Los valores atípicos son observaciones inusuales que difieren significativamente del patrón general de los datos y pueden tener un impacto desproporcionado en el cálculo del coeficiente de correlación.

La correlación de Pearson se basa en la magnitud de las diferencias entre los valores de las variables. Cuando hay valores atípicos presentes en los datos, pueden ejercer una influencia desproporcionada en el cálculo del coeficiente de correlación, tirando de la línea de mejor ajuste y distorsionando la relación real entre las variables. Esto puede llevar a interpretaciones erróneas sobre la fuerza y la dirección de la correlación.

Prykhodko et al. (2017) afirman que el análisis de datos atípicos desde una perspectiva de datos multivariados se puede simplificar por medio de la normalización o estandarización de los datos a valores o puntajes de Z.

$$Z_i = \frac{x_i - \mu}{\sigma} \qquad \text{(Ec. 5)}$$

Donde Z_i es el valor Z para cada observación x_i es cada valor observado, μ es la media de los datos y σ es la desviación estándar. Al realizar este procedimiento con ambas variables, se obtiene una comparación de ambas variables en términos de su desviación estándar, considerando el rango donde los valores se consideran típicos es $-3 \leq Z_i \leq 3$.

2.1.4.1.3. Independencia entre cada par de observaciones

La independencia de datos es un supuesto fundamental en la estadística y en muchos métodos de análisis. Esta es crucial para asegurar que los resultados de la correlación de Pearson sean confiables, interpretables y útiles para tomar decisiones fundamentadas.

Sin embargo, la falta de independencia entre las observaciones de una variable, que también se conoce como autocorrelación, es una cualidad difícil de observar en los datos sin conocer el contexto de cómo fueron recolectados; pues a menudo, los datos no independientes provienen de eventos recolectados muy cerca en el tiempo.

Por los motivos descritos anteriormente, el análisis de independencia entre pares de observaciones debe realizarse desde la perspectiva de series de tiempo, utilizando el orden en el que se obtuvieron las observaciones como variable temporal. La prueba de Durbin-Watson es una de los más populares para detectar la autocorrelación. Para esta prueba, se plantea una hipótesis nula afirmando la inexistencia de autocorrelación de la siguiente forma.

$$H_0: \rho = 0$$
$$H_a: \rho \neq 0 \qquad \text{(Ec. 6)}$$

El estadístico de prueba que permite comprobar dichas hipótesis se calcula en términos del error y el orden de cada variable, y está determinado por la siguiente ecuación:

$$d = \frac{\sum_{t=2}^{n}(e_t - e_{t-1})^2}{\sum_{t=1}^{n} e_t^2} \qquad \text{(Ec. 7)}$$

Donde e denota a los residuos del modelo lineal subyacente generado por la relación entre variables, y t denota el orden en que fueron tomadas las observaciones. De forma que el estadístico de prueba d se calcula en términos de la variación observable entre cada muestra de datos pareados.

Los resultados del estadístico de prueba d son comparados con la tabla del anexo 4, con los valores críticos d_L y d_U; que son comparados con tres criterios: si $d < d_L$, se rechaza la hipótesis nula; si $d > d_U$, no se rechaza la hipótesis nula. Pero si $d_L \leq d \leq d_U$, se determina que la prueba es no concluyente.

2.1.4.1.4. Definición

Una vez se ha verificado el cumplimiento de los supuestos anteriormente descritos, es posible el cálculo del coeficiente de correlación de Pearson. Sea X la variable independiente e Y la variable independiente y los datos están presentados en la forma de *n* pares ordenados; según Bewick, Cheek & Ball (2003), el coeficiente de correlación de Pearson está dado por la siguiente ecuación:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \qquad \text{(Ec. 8)}$$

El valor obtenido de r estará siempre entre valores -1 y +1. El símbolo de este coeficiente indica el sentido de la correlación; siendo negativo para correlaciones inversas y, positivo, para correlaciones directas. Por otro lado, el

valor del coeficiente indica qué tan fuerte es esta asociación; los valores más extremos indican relaciones más fuertes.

Es evidente que un valor de r cercano a 0 indica que no hay relación entre las variables; sin embargo, vale la pena repasar el supuesto de linealidad de la asociación de las variables.

Dado que el modelo de Pearson para determinar la correlación está basado en un modelo lineal, un valor 0 de r indica que no existe una relación lineal, pero se debe profundizar el análisis para determinar si la relación entre las variables obedece a otro tipo de función.

Para realizar este análisis, es factible utilizar el coeficiente de Pearson con transformaciones de variables, siempre y cuando estas cumplan con los supuestos mencionados anteriormente.

2.1.4.1.5. Análisis

El valor obtenido del coeficiente de Pearson es útil para la determinación de datos muestrales; sin embargo, para hacer estimaciones de la proporción se suele utilizar una prueba de hipótesis para determinar si la asociación es significativa o no. Para ello, en su análisis de métodos de correlación, Badii *et. al.* (2014) indica que es posible plantear las hipótesis de la siguiente forma:

$$H_0: r = 0$$
$$H_a: r \neq 0 \qquad \text{(Ec. 9)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional.

Para probar dicha hipótesis, Sánchez & Borges (2003) proponen el procedimiento de transformación $\mathbb{Z}$ de Fisher, que busca centralizar los valores del coeficiente de correlación de Pearson. Dicha transformación está descrita por la siguiente ecuación:

$$\mathbb{Z} = \operatorname{atanh}(r) = 0.5 * \ln\left(\frac{1+r}{1-r}\right) \qquad \text{(Ec. 10)}$$

La transformación propuesta por Fisher asume una distribución normal, con una desviación estándar equivalente a $\frac{1}{\sqrt{n-3}}$, por ende, el estadístico de prueba se calcula de la siguiente forma:

$$Z = \frac{\mathbb{Z} - Z_{\alpha}}{\frac{1}{\sqrt{n-3}}} \qquad \text{(Ec. 11)}$$

Dado que la prueba de hipótesis se plantea a dos colas, Z_{α} toma un valor igual a cero. El resultado del estadístico de pruebas puede analizarse comparando la tabla de valores de la distribución normal del anexo 5, para decidir sobre la aceptación o rechazo de la hipótesis planteada.

2.1.4.2. Coeficiente de Spearman

Cuando las dos variables bajo estudio no cumplen con los supuestos de normalidad y linealidad que supone el coeficiente de Pearson, una de las alternativas no paramétricas para medir la correlación es el método de Spearman, que es una medida de la correlación entre las posiciones relativas de cada variable.

2.1.4.2.1. Supuestos

Para la aplicación de este método de correlación, los datos deben cumplir con los siguientes supuestos:

- Las variables deben ser de intervalo, de razón u ordinales, de forma que puedan ordenarse.

- No es necesario que se asuma una relación lineal entre las variables.

- No es requerido que las variables sigan una distribución normal.

2.1.4.2.2. Definición

El coeficiente de correlación de Spearman, denotado como $r_{spm,}$ no está basado en el valor que toma cada variable en su *i-ésima* observación, sino en la posición relativa que esta observación toma respecto al conjunto de datos. Para ello, definidas las variables x e y, se calcula el rango o posición media de cada observación; a esta variable se le denotará como R_x y R_y.

El coeficiente de Spearman estará dado por la siguiente ecuación:

$$r_{spm} = 1 - \frac{\sum_{i=1}^{n} D_i^2}{n(n^2 - 1)} \qquad \text{(Ec. 12)}$$

Donde *n* es el tamaño de la muestra y, *D*, es la diferencia entre los rangos de *X* e *Y* de cada observación. Lo que algebraicamente puede definirse como:

$$D_i = R_{xi} - R_{yi} \qquad \text{(Ec. 13)}$$

Al igual que el análisis del coeficiente de Pearson, el valor de r_{spm}varía desde -1 hasta +1 de forma adimensional. Siendo los valores más extremos los que describen las correlaciones entre variables más fuertes, y los cercanos a cero implican falta de correlación.

2.1.4.2.3. Análisis

La correlación de Spearman también puede ser analizada por medio de pruebas de hipótesis para hacer afirmaciones correspondientes a la población. Para ello, se plantean las hipótesis de la siguiente forma:

$$H0: r_{spm} = 0$$
$$H1: r_{spm} \neq 0 \qquad \text{(Ec. 14)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional. La comparativa del estadístico de prueba con los valores críticos puede definirse desde la tabla de Spearman, que contiene valores estándar para un tamaño de muestra y nivel de significancia dados.

Según May & Looney (2022), al igual que el coeficiente de Pearson, el coeficiente de Spearman puede transformarse utilizando la transformación de Fisher, que para este método es conocida como $\mathbb{Z}_s$, y está definida por la ecuación 15.

$$\mathbb{Z}_s = \operatorname{atanh}(r_s) = 0.5 * \ln\left(\frac{1 + r_s}{1 - r_s}\right) \qquad \text{(Ec. 15)}$$

Para esta transformación, que también asume una distribución normal estándar, la desviación estándar equivale a $\frac{1.03}{\sqrt{n-3}}$. Por lo tanto, el estadístico de prueba para la inferencia se determina por:

$$Z = \frac{\mathbb{Z}_s - Z_\alpha}{\frac{1.03}{\sqrt{n-3}}} \qquad \text{(Ec. 16)}$$

El valor del estadístico de prueba puede compararse con la tabla de distribución normal estándar mostrada en el anexo 5, permitiendo aceptar o rechazar las hipótesis propuestas.

2.1.4.3. Coeficiente de Kendall

Para Serna (2019), el coeficiente de Kendall, denotado τ, es una alternativa no paramétrica para el cálculo de la correlación entre variables que no presentan un comportamiento paramétrico. Se basa en los intervalos jerarquizados de las observaciones, lo que permite que la distribución de τ sea independiente de los valores que presentan las variables x e y.

2.1.4.3.1. Supuestos

Para la aplicación del modelo de Kendall, los datos deben cumplir con los mismos supuestos utilizados para el coeficiente de Spearman, los cuales son:

- Las variables deben ser de intervalo, de razón u ordinales, de forma que puedan ordenarse.

- No es necesario que se asuma una relación lineal entre las variables.

- No es requerido que las variables sigan una distribución normal.

2.1.4.3.2. Definición

Dado que la estimación de τ de Kendall está basado en la concordancia y discordancia de sus pares ordenados. Esto lo define Serna (2019) como "Sea $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ una muestra aleatoria de *n* observaciones de variables continuas o al menos ordinales, un par de observaciones (x_i, y_i) y (x_j, y_j) son concordantes si $x_i < x_j$ y $y_i < y_j$, o si $x_i > x_j$ y $y_i > y_j$". Por ende, cada par de datos será concordante o discordante, con lo que podrá calcularse el τ de Kendall como:

$$\tau = \frac{Nc - Nd}{Nc + Nd} \qquad \text{(Ec. 17)}$$

Donde N_c denota el número de pares concordantes, y N_d denota el número de pares discordantes.

El resultado del coeficiente de Kendall será un valor entre -1 y 1, que define el grado de asociación que existe entre dos variables. Si se trata de variables independientes, el valor esperado para el coeficiente de Kendall es cero.

2.1.4.3.3. Análisis

El valor calculado de τ también puede ser usado como un estadístico de prueba para hacer pruebas de hipótesis que ayuden a hacer aseveraciones sobre la población. Para ello, la hipótesis a plantear sigue la siguiente lógica:

$$H_0: \tau = 0$$
$$H_a: \tau \neq 0 \qquad \text{(Ec. 18)}$$

Si la hipótesis nula es aceptada, se afirma con un nivel de significancia α que no existe correlación significativa entre las variables a nivel poblacional. El coeficiente de correlación de Kendall también puede centralizarse utilizando la transformación de Fisher según May & Looney (2022), utilizando la ecuación 19.

$$\mathbb{Z}_\tau = \operatorname{atanh}(\tau) = 0.5 * \ln\left(\frac{1+\tau}{1-\tau}\right) \qquad \text{(Ec. 19)}$$

Para esta distribución, el valor de la desviación estándar equivale a $\frac{0.661}{\sqrt{n-3}}$, por lo tanto, el valor del estadístico de pruebas para la comprobación de las hipótesis planteadas se define por:

$$Z = \frac{\mathbb{Z}_\tau - Z_\alpha}{\frac{0.661}{\sqrt{n-3}}} \qquad \text{(Ec.20)}$$

Dicho estadístico de prueba se puede comparar con los valores descritos en la tabla de distribución normal del anexo 5.

2.1.4.4. Pruebas de Independencia Chi (X^2) cuadrado

El análisis de correlación para variables categóricas difiere de los métodos utilizados en variables cuantitativas, ya que en estos no es posible utilizar sus valores o jerarquías para definir coeficientes de interés. El análisis de este tipo de relaciones se realiza con una tabulación ordenada de las frecuencias de los datos ordenados mediante tablas de contingencia.

Lopez & Fachelli (2015) definen una tabla de contingencia como una tabla de frecuencias que resulta de la distribución conjunta al relacionar o cruzar dos o más variables cualitativas. Estas pueden representar la tabulación cruzada de *n* variables; sin embargo, para el propósito de medir la independencia de los datos, se hará énfasis en las tablas de contingencia bidimensionales.

Tabla 2.

Tabla de contingencia bidimensional

Estructura de Tablas de Contingencia

	Segmento			
Método de envío	Cliente	Empresa	Pequeña empresa	**Total general**
Estándar	3,165	1,864	1,144	6,173
Mismo día	272	150	93	515
Rápido	1,097	618	382	2,097
Urgente	759	417	293	1,469
Total general	5,293	3,049	1,912	10,254

Nota. Tabla de contingencia denotando la cantidad de observaciones por cada categoría de dos variables. Elaboración propia, realizado con Tableau.

Las tablas de contingencia se completan con los totales de filas y columnas, que serán utilizados para hacer diferentes estimaciones de la distribución bivariada, observando la homogeneidad e independencia de los datos. Para ellas, es posible la implementación de pruebas de independencia como una alternativa para determinar la correlación entre variables categóricas.

2.1.4.4.1. Supuestos

Para aplicar una prueba de independencia es necesario que los datos cumplan con los siguientes criterios:

- Se presentan tabulados en una tabla de contingencia de dos dimensiones.
- Las variables son cualitativas (medidas a nivel nominal u ordinal, o son tratadas en esa escala de medición).

2.1.4.4.2. Definición

Las pruebas de independencia, a diferencia de los métodos de correlación para variables cuantitativas, no tiene como resultado un coeficiente que indique el grado de asociación, más bien se trata de una prueba de hipótesis basada en la distribución X^2, que permite establecer o no una relación significativa entre las variables. Para esto, se plantean la hipótesis nula e hipótesis alterna de la siguiente forma:

$$\begin{aligned} &\textit{H0: Las variables son independientes} \\ &\textit{Ha: Las variables no son independientes.} \end{aligned} \qquad \text{(Ec.21)}$$

Estas hipótesis serán probadas por medio de un estadístico de prueba basado en la distribución X^2, que algebraicamente está definido como:

$$X^2 = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{\left(f_{ij} - fe_{ij}\right)^2}{fe_{ij}} \qquad \text{(Ec.22)}$$

Donde f_{ij} es la frecuencia de la *i-ésima* fila en la *j-ésima* columna, fe_{ij} es la frecuencia de la *i-ésima* fila en la *j-ésima* columna. Esta última, puede calcularse utilizando la siguiente expresión:

$$fe_{ij} = \frac{total\ de\ fila\ i * total\ de\ columna\ j}{total\ de\ observaciones} \quad \text{(Ec.23)}$$

Los grados de libertad que toma esta distribución bivariada están definidos por la cantidad de niveles que presentan ambas variables. Por tanto, los grados de libertad v están dados por:

$$v = (i-1) * (j-1) \quad \text{(Ec.24)}$$

Con estos cálculos realizados, es posible concluir mediante el cálculo del p-valor, aceptando o rechazando el supuesto de independencia de datos. Por la naturaleza de este análisis, el resultado no incluye un grado de asociación con una dirección definida; únicamente es posible afirmar la dependencia o independencia con un nivel de significancia definido.

2.1.5. Regresión logística

El análisis de correlación descrito en la sección anterior es únicamente útil para conocer la relación que hay entre dos variables. Posterior a ello, es común hacer inferencias estadísticas por medio de técnicas de regresión. En este contexto, una de las tareas de mayor importancia es conocer la naturaleza de los datos que se desean explicar o predecir.

La regresión logística es un método de análisis estadístico que permite predecir, por medio de probabilidades, el resultado de una variable dependiente

de tipo categórica. Aunque estos modelos son capaces de relacionar una variable dependiente politómica (que admite múltiples valores), esta es en especial potente cuando solamente existen dos alternativas (dicotómicas).

2.1.5.1. Regresión logística binomial

En su análisis, Sagaró & Zamora (2019) indican que una regresión logística expresa la probabilidad de que ocurra determinado evento en función de sus variables regresoras; y aunque es posible hacerlo con variables politómicas, es la versión dicotómica o binomial la más potente y utilizada en el ámbito de la investigación científica.

El problema de observación propuesto implica el análisis de la retención de clientes profesionales, lo cual es un escenario con dos posibles resultados: la pérdida o retención de estos. Por esta razón, la revisión documental sobre los métodos de regresión logística estará enfocada únicamente en la regresión logística binomial.

2.1.5.1.1. Elección de las variables

Una de las fases más importantes dentro del análisis de regresión es la elección de las variables regresoras que definirán el modelo. Se considera necesario realizar análisis de correlación o independencia de datos según sean los casos presentados, y utilizar dentro del análisis únicamente aquellas con una asociación significativa.

Aun así, se recomienda incluir también en el proceso aquellas variables que demostraron una correlación débil contra la variable dependiente, pues,

aunque en solitario tengan una débil asociación, es posible que, al ser analizadas y probadas con el resto de covariables, estas sean más importantes.

Se recomienda excluir de este análisis las variables que, por causalidad, no pueden estar relacionadas al problema de estudio, las que sean redundantes o que estén estrechamente relacionadas, para evitar la multicolinealidad o que sean el desenlace de la variable objetivo que está por probar.

2.1.5.1.2. Tratamiento de los datos

El modelo de cálculo de una regresión logística requiere acomodar la información extraída adecuadamente, según Sagaró & Zamora (2019). Para ello, los autores recomiendan las siguientes mecánicas para el tratamiento de datos previo a la construcción del modelo de regresión logística:

- El modelo está basado en el uso de variables dicotómicas que toman el valor 1 para la presencia de la característica y 0 para su ausencia.

- Si las variables son nominales pero politómicas, se ve en la necesidad de convertir cada nivel de esta en una variable *dummy*. Por ende, si se cuenta con una variable politómica con n posibles valores, esta se convertirá en n variables dicotómicas, como se puede observar en la tabla 3.

- Las variables ordinales, por otro lado, también se crean como variables *dummy* en cada nivel, pero su orden si implica un orden jerárquico.

- Además, para las variables cuantitativas, se asume que cada cambio de una unidad sobre la variable regresora tiene la misma magnitud.

Tabla 3.

Estructura de las variables dummy

Estructura de Variables *Dummy*

ID	Color	Color Azul	Color Blanco	Color Negro
1	Blanco	0	1	0
2	Azul	1	0	0
3	Negro	0	0	1
4	Azul	1	0	0
5	Negro	0	0	1
6	Blanco	0	1	0
7	Negro	0	0	1
8	Blanco	0	1	0
9	Negro	0	0	1
10	Blanco	0	1	0
11	Negro	0	0	1
12	Blanco	0	1	0

Nota. Ejemplo de conversión de variables categóricas a variables *dummy.* Elaboración propia, realizado con Tableau.

2.1.5.1.3. Definición

Los mismos autores indican que el modelo de regresión logística binaria, que expresa la probabilidad p de que ocurra un evento en función de ciertas variables viene dado por:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k)}} \qquad \text{(Ec.25)}$$

Donde p representa la probabilidad de ocurrencia del evento dicotómico representado por la variable dicotómica, $\beta_1, \beta_2, \ldots, \beta_k$ son los coeficientes de regresión asociados a cada variable $x_1, x_2, \ldots, x_k$. La estimación de los

coeficientes que se asocian a cada variable es usual realizarla mediante un método iterativo de Newton-Rhapson, utilizando un proceso de máxima verosimilitud. Para esto, se hace una transformación logarítmica dividiendo la probabilidad por su complementario de la siguiente forma:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k \qquad \text{(Ec.26)}$$

Dada la cantidad de variables que se suelen utilizar para la construcción de los modelos de regresión logística, hay necesidad de iterar la construcción de múltiples modelos iterativos para elegir el que mejor se ajuste a los datos.

2.1.5.2. Validación de los modelos de regresión logística

Una vez definido el modelo de regresión logística, es necesario desarrollar la fase de diagnóstico, donde se corrobora el cumplimiento de los supuestos, validando si otra función u otra combinación de variables describe mejor el problema planteado. Este análisis debe enfocarse en el cumplimiento de dos aspectos: primero, el modelo debe ser congruente en cuanto a las variables elegidas y su interacción, y segundo, se debe cumplir el principio de parsimonia, que busca la menor cantidad de variables para explicar el modelo.

El diagnóstico de los supuestos del modelo de regresión logística debe enfocarse desde dos perspectivas. La primera es el análisis de los residuos del modelo, que pueden ser de tres tipos: estandarizados, estudentizados y absolutos; el modelo con el menor error es el que mejor describe el problema de investigación. El segundo método corresponde al análisis de las medidas de influencia que cuantifican el efecto de cada observación sobre el vector de

predicciones, como las medidas de apalancamiento del método Leverage y las medidas de distancia del método Cook.

2.1.5.3. Exactitud de modelo de regresión logística

Oshaki *et. al.* (2017) indican que la exactitud en modelos de clasificación, como es el caso de un modelo de regresión logística puede obtenerse utilizando una matriz de confusión, donde se denoten los valores predichos del modelo comparados con los valores reales de la variable de respuesta.

Tabla 4.

Matriz de confusión para resultados de predicción

	Respuesta positiva	**Respuesta negativa**
Predicho positivo	Verdaderos positivos (VP)	Falso positivo (FP)
Predicho negativo	Falso negativo (FN)	Verdadero negativo (VN)

Nota. Estructura de matriz de confusión para la evaluación de exactitud de predicción. Elaboración propia, realizado con Word.

La matriz de confusión se obtiene haciendo un conteo de los posibles resultados al comparar los resultados predichos del modelo y los resultados reales. Estos pueden ser:

- Verdaderos positivos (VP): resultados predichos como exitosos por el modelo que en la práctica también fueron exitosos.

- Falsos positivos (FP): resultado que se predijo positivo por el modelo, pero no tuvieron un resultado exitoso.

- Falso negativo (FN): resultado que se predijo negativo en el modelo, pero fue positivo en la práctica.

- Verdadero negativo (VN): resultado que se predijo negativo por el modelo y fue negativo en la realidad.

Con los resultados de la matriz anteriormente descrita, es posible calcular la exactitud del modelo, que está determinada por:

$$Exactitud = \frac{VP + VN}{N} \qquad \text{(Ec.27)}$$

Donde *VP* representa la cantidad de verdaderos positivos, *VN* la cantidad de verdaderos negativos y *N* la cantidad total de observaciones.

2.1.6. Algoritmo de K medias

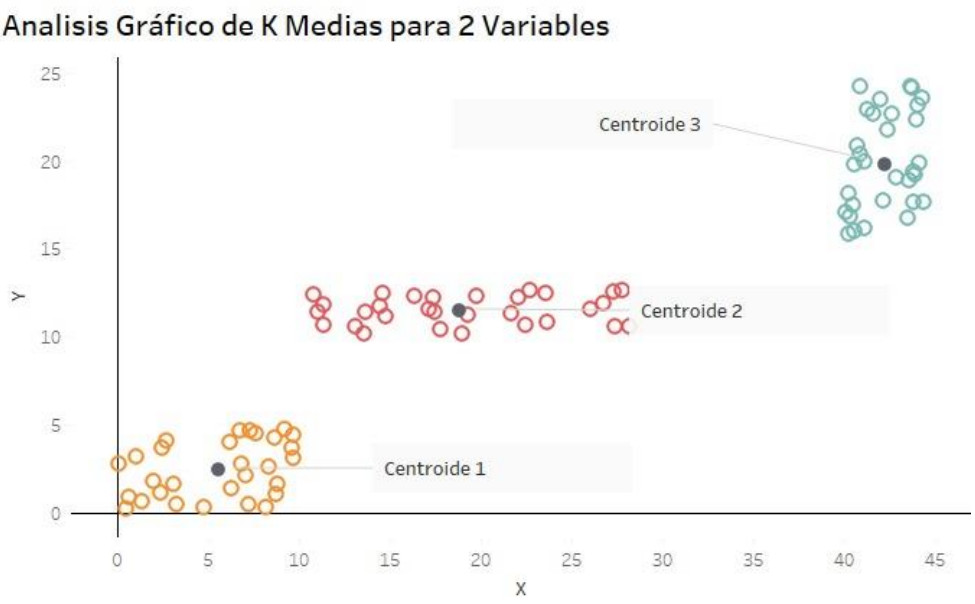
El método de segmentación por K medias es un algoritmo no supervisado, valioso en el análisis de grandes conjuntos de datos, según indican Anitha & Patil (2022).

El proceso implica la identificación de K centroides en los datos, siendo K un valor definido a priori. Posteriormente, cada registro del conjunto de datos se asigna al centroide más cercano en función de la medida de distancia, generalmente la distancia euclidiana. Este proceso se repite iterativamente hasta que la asignación de los registros a los centroides converge, resultando en grupos distintos y significativos en los datos.

Si se hace un análisis de esto en dos dimensiones, se podría observar que las agrupaciones se asemejan a la figura 4, donde cada color representa un segmento asociado a su centroide.

Figura 4.

Análisis gráfico de K medias para dos variables



Nota. Análisis visual de agrupación utilizando el algoritmo de K medias a un conjunto de datos de 2 variables con 3 centroides definidos. Elaboración propia, realizado con Tableau.

Para realizar la clasificación de los clientes, el algoritmo presenta variables o dimensiones, y busca construir k grupos donde se minimiza la suma de distancias de los objetos a su centroide. Esta distancia se puede calcular mediante la distancia euclidiana, que se define matemáticamente de la siguiente forma:

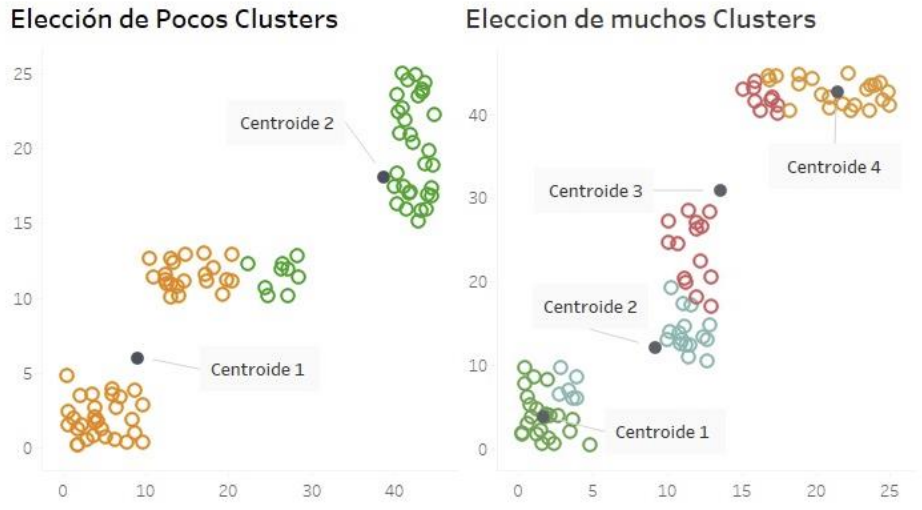
$$s = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \qquad \text{(Ec.28)}$$

Donde q_i es el valor de la variable y p_i es el valor del centroide. Esto es calculado para todas las variables de segmentación.

La cantidad de segmentos K que se utiliza para segmentar debe ser una definición del investigador; sin embargo, cabe mencionar que a medida que aumenta la cantidad de segmentos, naturalmente disminuye el error total de segmentación artificialmente. Según Hernández *et. al.* (2018), esta es una de las principales desventajas de esta técnica de clasificación.

Figura 5.

Elección incorrecta de clústeres



Nota. Análisis visual de agrupación mediante el algoritmo de K medias, utilizando tanto una cantidad K inferior como superior a la considerada adecuada. Elaboración propia, realizado con Tableau.

Elegir la cantidad adecuada de segmentos sobre los cuales optimizar, es una de las tareas críticas en el contexto del algoritmo de k medias. Esto es particularmente sencillo si se trata de un *set* de datos con dos o tres dimensiones, pues el análisis se puede realizar de forma visual. Sin embargo, si incrementa la cantidad de variables, ya no es posible representar este modelo en 3 dimensiones, por lo que se debe recurrir a otros métodos de amplia aceptación: Método Silhouette y Método Elbow.

2.1.6.1. Método Silhouette

Este método utiliza un coeficiente conocido como Silhouette, que está definido como la diferencia entre la distancia promedio de los elementos a su centroide más cercano y la distancia intra-*clúster* dividido por el máximo de los dos. Se itera este algoritmo con diferentes valores de K; cuando el coeficiente de Silhouette se maximiza, se obtiene la cantidad óptima de centroides sobre los cuales maximizar

$$CS = \frac{1}{N}\sum_{i=1}^{N}\frac{b(x) - a(x)}{\max\{a(x), b(x)\}} \qquad \text{(Ec.29)}$$

Donde *CS* denota el valor del coeficiente de Silhouette, $b(x)$ es la distancia media de los centroides y $a(x)$ es la distancia intra-clúster y *N* denota cada grupo sobre los cuales se está haciendo el cálculo del coeficiente.

2.1.6.2. Método Elbow

Este método calcula la suma de cuadrados del error como la métrica principal de ajuste para esta agrupación. Naturalmente, al incrementar la cantidad de segmentos o clústeres la distancia entre los puntos y sus

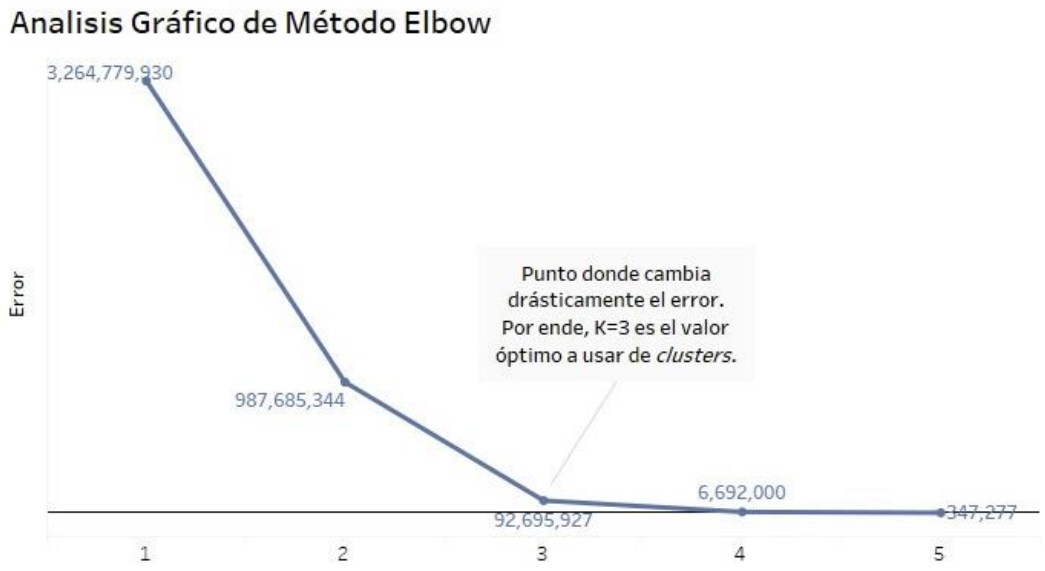
correspondientes centroides disminuye, por lo que la suma de los cuadrados del error también disminuye.

A medida que se itera para los diferentes valores de K, el error total del modelo se reduce drásticamente para los valores más pequeños, pero a medida que el valor de K incrementa, la reducción del error se hace más pequeña.

En el análisis gráfico de este comportamiento, se obtiene un punto de inflexión en el gráfico, que indica el valor óptimo de centroides a utilizar.

Figura 6.

Análisis gráfico con el método Elbow



Nota. Gráfico de error total obtenido con algoritmo de agrupación por K media con diferentes valores de K (eje x) y su correspondiente punto de inflexión. Elaboración propia, realizado con Tableau.

2.2. Empresa

El desarrollo del presente trabajo de investigación se realizará en una empresa comercializadora de materiales de construcción en Guatemala.

2.2.1. Características de la empresa

El segmento con el que opera la empresa implica que hay una gran cantidad de clientes eventuales, que compran por una construcción o remodelación puntual, así como otros clientes que se dedican a prestar este tipo de servicios.

La naturaleza del primer tipo de clientes no permite la gestión de la lealtad; sin embargo, para la empresa sí es un objetivo estratégico mejorar la lealtad de los clientes que están considerados como profesionales.

Para el desarrollo de lealtad, la empresa tiene planeada la generación de estrategias mercadológicas enfocadas en las necesidades de cada usuario, para lo cual requiere efectuar un análisis profundo del comportamiento de los clientes y qué tan probable es que estos se pierdan.

2.2.2. Análisis del comportamiento de compras

Según Rahim *et. al.* (2021), el análisis del comportamiento de los clientes en negocios minoristas es una de las actividades de mayor beneficio en el ámbito del mercadeo estratégico. Indica que analizar el comportamiento de los clientes implica conocer a profundidad la forma en la que se distribuyen anuncios, espacios y la forma en que interactúan con estos elementos. Esto último, hace

que el análisis requiera de herramientas tecnológicas complejas, que normalmente no están disponibles en los comercios.

Para simplificar este análisis, es posible medir el comportamiento de compra de los clientes por medio del análisis de su comportamiento transaccional, definiendo tres variables que describen cómo un cliente se relaciona con un comercio: su frecuencia de compra, la antigüedad desde su última compra y el nivel de gasto que presenta en cada visita. A este método se le conoce como RFM.

2.2.2.1. Reciencia (R)

Este indicador se refiere al tiempo, medido en días, que ha sucedido desde la última compra del cliente. Es de especial importancia porque permite determinar qué clientes se pueden considerar como perdidos. Algebraicamente, la reciencia está definida como:

$$R = Fecha\ Act - Ult.fecha\ Compra \qquad \text{(Ec.30)}$$

El resultado de este cálculo se representa como una variable continua para facilitar su análisis estadístico.

2.2.2.2. Frecuencia (F)

La frecuencia de compra, en el análisis del comportamiento de clientes, se refiere al tiempo medio que ocurre entre cada compra.

Para su cálculo, se debe definir una ventana temporal para la revisión de datos; posteriormente, se toma la diferencia entre la última y primera transacción

en esa ventana temporal y se divide por la cantidad de transacciones. Esta expresión, puede definirse como:

$$F = \frac{\text{Última } fecha\ de\ Compra - Primera\ Fecha\ de\ Compra}{Cantidad\ de\ Transacciones} \qquad \text{(Ec.31)}$$

Para facilitar el análisis y la comparación, la frecuencia de compra debe mantener las mismas dimensionales que la variable de reciencia.

2.2.2.3. Monto promedio (M)

El monto promedio de compras es la media aritmética de los montos facturados a cada cliente. Al igual que la variable de frecuencia, se define la misma ventana temporal sobre la cual se hace la extracción de datos. Esta está definida de la siguiente forma:

$$M = \frac{Monto\ Total\ Facturado}{Cantidad\ de\ Transacciones} \qquad \text{(Ec.32)}$$

Las unidades dimensionales para esta variable deben estar en formato de moneda, para realizar estimaciones adecuadas.

2.2.2.4. Clasificación de clientes por percentiles de RFM

Las variables RFM obtenidas en la sección anterior proporcionan una perspectiva numérica sobre el comportamiento de los clientes, sin embargo, para fines mercadológicos es necesario agrupar los clientes en grupos similares de cada variable para su tratamiento masivo.

Según Murad (2021), se basa en la suposición de que los percentiles (usualmente se utilizan deciles o quintiles) de reciencia, frecuencia y monto promedio tienen una diferente tasa de respuesta a los estímulos mercadológicos que se envían.

Para la frecuencia, luego del cálculo de los percentiles, se asigna la clasificación 1 para los clientes que presentan un menor tiempo medio entre compras, mientras que se asigna la clasificación más baja al percentil que tiene más tiempo promedio entre transacciones. De esta forma, la clasificación de la variable frecuencia clasifica mejor a los clientes más leales.

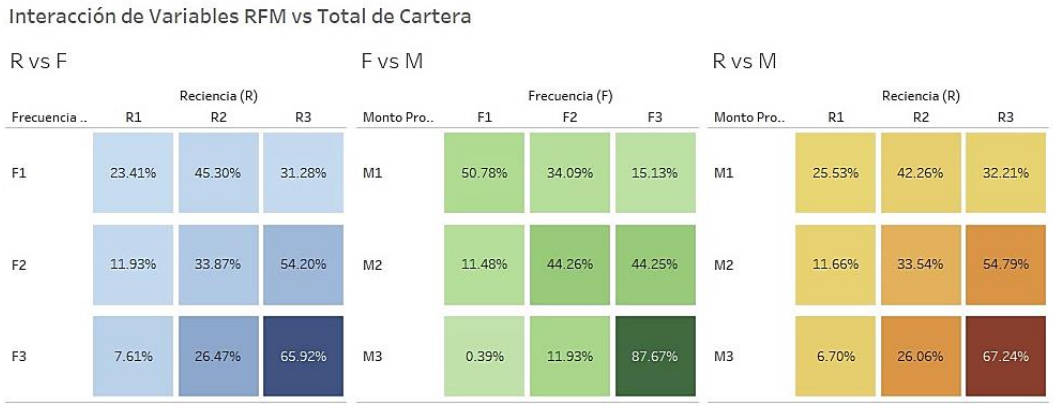
Esta misma lógica debe aplicarse para la medición de la variable de reciencia, donde se asigna la primera clasificación al percentil con menor antigüedad desde su última compra.

Por último, la clasificación de los clientes desde su nivel de gasto se debe hacer al inverso del resto de variables; pues se debe asignar la mejor clasificación de la variable al percentil de clientes con mayor gasto promedio.

Con esta clasificación, todos los clientes activos de la base de datos obtienen una clasificación de cada variable, y las combinaciones permiten hacer agrupaciones adecuadas de los clientes. Por ejemplo, los clientes clasificados como R1, F1 y M1 son aquellos de mayor interés para la empresa, y los clasificados como R3, F3 y M3 son los clientes de menor valor.

Figura 7.

Análisis de interacción de variables RFM



Nota. Análisis de la interacción entre variables RFM, mostrando la combinación reciencia – frecuencia, frecuencia – monto promedio y reciencia – monto promedio. Elaboración propia, realizado con Tableau.

2.2.3. Segmentación de la clientela

Para Gajanova, Nadanyiova & Moravcikova (2019), identificar qué tan leales a la marca son los clientes es una tarea de especial interés para el éxito mercadológico y el tratamiento adecuado de estos. Por ende, la tarea de segmentación de los clientes basado en la actividad de estos es de gran valor. Este tipo de segmentación se logra mediante el análisis de variables de comportamiento descritos en el método RFM.

Sin embargo, los autores indican que es un error común tratar de analizar el comportamiento de los clientes únicamente por la forma en la que estos interactúan transaccionalmente. En realidad, se requiere mucho más contexto referente a otras variables, que son útiles para clasificar a los clientes. Uno de los

métodos que los autores toman para resolver este problema, es analizando a los clientes desde cuatro tipos de segmentación que se describen a continuación.

2.2.3.1. Segmentación geográfica

La segmentación de los clientes por su ubicación geográfica se refiere, principalmente, a agrupar a los clientes por las regiones en las que estos se encuentran. La lógica detrás de este tipo de segmentación es que hay barreras físicas que pueden diferenciar el comportamiento de los clientes. Por ejemplo, el tamaño del mercado o los incentivos fiscales pueden ser determinantes en el comportamiento de los clientes.

Entre las variables más comunes de segmentación geográfica se encuentran: país, ciudad, idioma, clima, densidad y población, entre otras. De estar disponibles todas estas variables de segmentación, son útiles para el análisis. Sin embargo, debe considerarse un análisis de independencia de datos entre la variable de segmentación y la variable objetivo, para establecer si estas representan un factor determinante en el comportamiento de los clientes.

2.2.3.2. Segmentación demográfica

Los criterios demográficos de segmentación suelen ser los más utilizados para segmentar a los clientes. En el contexto de comercios minoristas, se busca encontrar el género, edad, estilo de vida o nivel socioeconómico de los clientes. Este enfoque funciona si se estudia a clientes individuales, pero el enfoque de clientes profesionales del presente estudio implica que la mayoría de los clientes por analizar son empresas, por lo que estos factores no son útiles para la segmentación.

Esto obliga a tomar en consideración otros factores de segmentación, como los descritos por Möllering (2018) se sugiere que para la segmentación se tome en cuenta principalmente las características de la organización, como el tamaño de la organización, segmento de mercado en el que opera o incluso la tecnología con la que operan.

En segundo plano, se debe tomar en consideración las características demográficas de los compradores con quienes la empresa tiene contacto; para ello se requiere tener un sistema de manejo de relaciones con clientes (CRM) con esta información.

Por último, también se recomienda evaluar las condiciones financieras del cliente, tales como el estatus de cobro, límites de crédito y términos de pago.

2.2.3.3. Segmentación psicográfica

Este tipo de segmentación tiene como objetivo explicar las diferencias en las que actúan los clientes, basados en sus predisposiciones sociales y psicológicas, y trata de agrupar a clientes con características demográficas y geográficas similares que se comportan de diferente manera. Entre este tipo de variables, se puede explorar todo lo que conlleve el comportamiento de los clientes en torno a sus transacciones.

Mucha de esta información es difícil de conseguir para toda la población de clientes; sin embargo, hay otras variables de segmentación que pueden obtenerse desde el análisis de los resultados transaccionales. En el análisis por estas variables puede considerarse la modalidad de entrega de los productos, la amplitud de categorías que compran o la forma de pago, entre otras.

2.2.3.4. Segmentación por satisfacción y lealtad

Gajanova, Nadanyiova & Moravcikova (2019) sugiere que uno de los principales predictores de la lealtad de los consumidores es la satisfacción de estos. Por ello, propone segmentar a los clientes en cuatro diferentes categorías, según su satisfacción y percepción hacia la marca:

- Clientes leales y altamente satisfechos: se trata de los clientes que son leales a la marca y se encuentran muy satisfechos con los servicios que se prestan. Tienden a ser defensores de la marca por lo que tienen baja probabilidad de perderse.

- Clientes leales por factores externos: son clientes que no están satisfechos con los servicios que presta la marca, pero son leales por algún factor externo. Este tipo de clientes suele perderse ante una mejor oferta de un competidor de la marca o la aparición de una nueva marca rival.

- Clientes satisfechos, pero no leales: se clasifican de esta forma a los clientes que están satisfechos, pero no son leales a la marca, por lo que están propensos a perderse. Este tipo de clientes suelen originarse cuando la mayor propuesta de valor es el precio.

- Clientes insatisfechos y no leales: son clientes que no son leales ni están satisfechos con los servicios.

Contar con este tipo de segmentación es útil para predecir los movimientos futuros de los clientes. Sin embargo, se requiere un profundo entendimiento de los clientes, ya que el nivel de satisfacción de los clientes se puede conocer únicamente a través de programas de encuestas con los clientes.

3. PRESENTACIÓN DE RESULTADOS

A través de diversas etapas que incluyeron revisión de literatura, minería y limpieza de datos, análisis de correlación, construcción de modelos de regresión y análisis de resultados, se obtuvieron los resultados que se muestran a continuación.

Los datos obtenidos mediante la fase de minería de datos fueron los detalles de transacciones de clientes profesionales de los últimos dos años. De estos, se utilizaron los primeros dieciocho meses de la ventana temporal para obtener las variables regresoras; y los últimos seis meses para obtener la variable de respuesta (recompra o pérdida del cliente).

Adicionalmente, la información obtenida fue colocada en una matriz de individuos por variables, que permitió tener una vista completa de cada cliente profesional con sus correspondientes variables regresoras para un análisis desde la perspectiva multivariada.

Al hacer esta segmentación y arreglo de datos fue posible determinar cuáles clientes eran activos en la ventana inicial. Si un cliente no tuvo transacciones en los dieciocho meses de este periodo, se consideraron inactivos y se excluyeron del modelo, manteniendo un total de 2,538 clientes profesionales en el modelo.

La variable de respuesta permitió identificar cuáles de estos clientes se mantuvieron activos. Como se puede observar en la tabla 5, solamente 1,162

clientes se mantuvieron activos, dando una tasa de retención del 45.8 % globalmente.

Tabla 5.

Resumen de los resultados obtenidos

Tipo de cliente	Cantidad de clientes	Clientes retenidos	Tasa de retención
Cliente profesional	2 538	1 162	45.8 %

Nota. Resumen del total de clientes profesionales, cantidad de clientes retenidos y su correspondiente tasa de retención. Elaboración propia, realizado con Tableau.

Con estos resultados preliminares, fue posible el análisis de las variables predictoras y la construcción del modelo de acuerdo con los objetivos propuestos para la investigación.

3.1. Objetivo 1: agrupar a los clientes profesionales en segmentos similares, basado en las variables de reciencia, frecuencia y monto de compras, aplicando métodos de simulación por K medias

Al obtener la información del modelo de entrenamiento, se procedió al análisis de las principales variables predictoras: reciencia, frecuencia y monto promedio de compras para segmentar a los clientes en agrupaciones similares.

Dado que el objetivo requirió agrupar a los clientes en segmentos suficientemente heterogéneos, fue necesario analizar si las variables del método RFM estaban correlacionadas entre sí. Para realizar estas pruebas de correlación, también fue necesario analizar la distribución de estas variables,

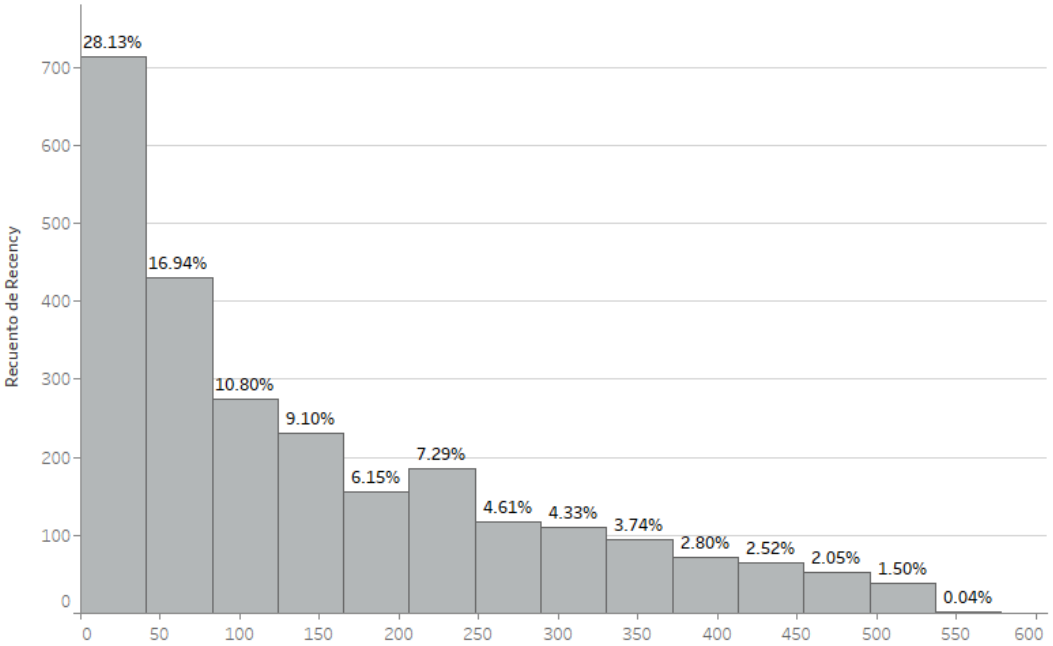
para determinar si esto se haría desde la perspectiva paramétrica o no paramétrica.

3.1.1. Análisis de la variable reciencia (R)

El análisis de la variable reciencia se inició gráficamente haciendo un histograma. La cantidad de clases incluidas en este histograma y en los posteriores fue determinada por el método de Sturges. En la figura 8 se observa que hay una alta concentración de clientes con una reciencia baja, sin embargo, en los niveles más altos de la variable no se ve una disminución sustancial.

Figura 8.

Histograma de reciencia (R)



Nota. Histograma de la variable reciencia (R), con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

Tal como se muestra en la tabla 6, la variable reciencia tiene una media de 145.4 días y una desviación estándar de 135.7 días con lo que su coeficiente de variación es del 93.4 %. Esto puede considerarse como una variación fuerte considerando la cantidad de observaciones disponibles; sin embargo, al analizar la distribución de esta variable, se obtienen muchos datos extremos que provocan el incremento en la variabilidad. Por el motivo anterior, se planteó la transformación de los datos para su tratamiento en la investigación.

Tabla 6.

Caracterización de la variable reciencia (R)

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	145.4	135.7	93.4 %

Nota. Media, desviación estándar y coeficiente de variación para variable reciencia (R). Elaboración propia, realizado con Word.

3.1.1.1. Transformación logarítmica de reciencia (R)

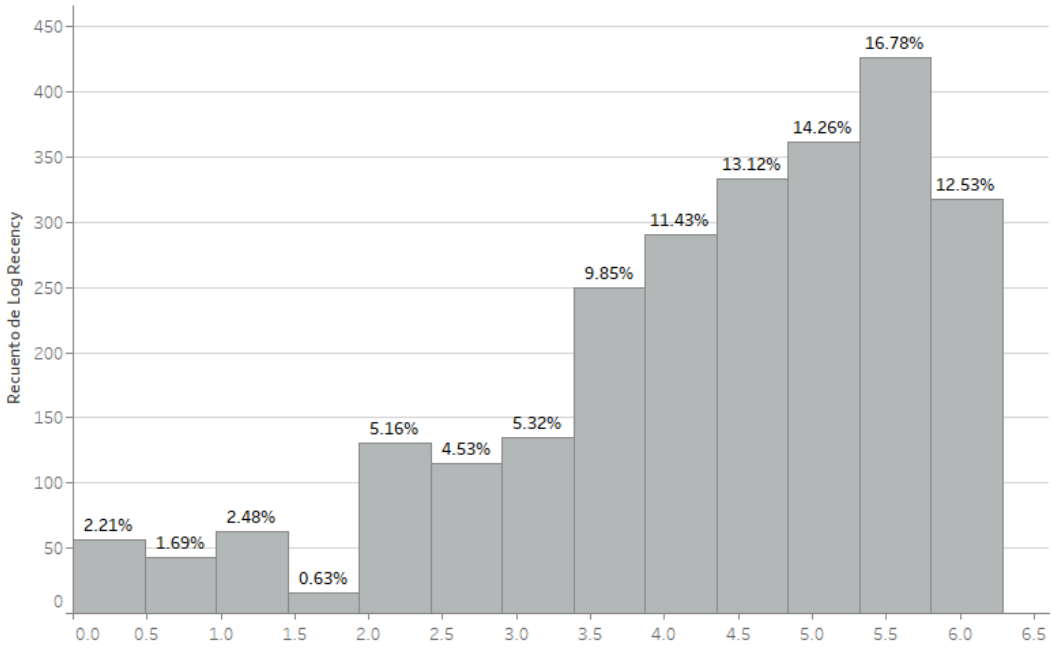
Se eligió hacer una transformación logarítmica de la variable para reducir su variabilidad, y también buscar que esta tuviera un comportamiento cercano a la distribución normal. Tal como se observa en el histograma que se muestra en la figura 9.

A diferencia de la distribución sin transformación, se observa que los datos están distribuidos alrededor del punto de máxima densidad. El comportamiento descrito es un indicador de que la variable transformada no cambia su

distribución, estando aún lejos de una distribución normal. Sin embargo, esto es solo una apreciación hecha a partir de un análisis gráfico.

Figura 9.

Histograma de reciencia (R) con transformación logarítmica



Nota. Histograma de la variable reciencia (R) con transformación logarítmica, con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

Sin embargo, como se muestra en la tabla a 7, la transformación de la variable sí redujo su variabilidad, pues el coeficiente de variación para la variable transformada fue de 33.9 %. El cambio descrito anteriormente, permitió tener las observaciones de manera más condensada, punto que fue clave en las etapas posteriores de análisis desde la perspectiva no paramétrica.

Por lo tanto, a pesar de no contar con esta variable en un comportamiento normal, fue posible realizar un análisis profundo de la misma, y utilizarla como una variable regresora de los diferentes modelos construidos.

Tabla 7.

Caracterización de la variable reciencia (R) con transformación logarítmica

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	4.297	1.456	33.9 %

Nota. Media, desviación estándar y coeficiente de variación para variable reciencia (R) con transformación logarítmica. Elaboración propia, realizado con Word.

3.1.1.2. Análisis de normalidad para variable reciencia (R)

A pesar de tener una reducción la variabilidad con la transformación, aun se tuvo la necesidad de probar la normalidad de ambas versiones de la variable.

Tabla 8.

P-valores de prueba de normalidad para reciencia (R)

Variable	P valor de prueba de normalidad	
	Variable original	Variable transformada
Reciencia	0.000	0.000

Nota. P-valores de las pruebas Kolmogórov-Smirnov para probar la normalidad de reciencia (R) en su versión original y transformada. Elaboración propia, realizado con Word.

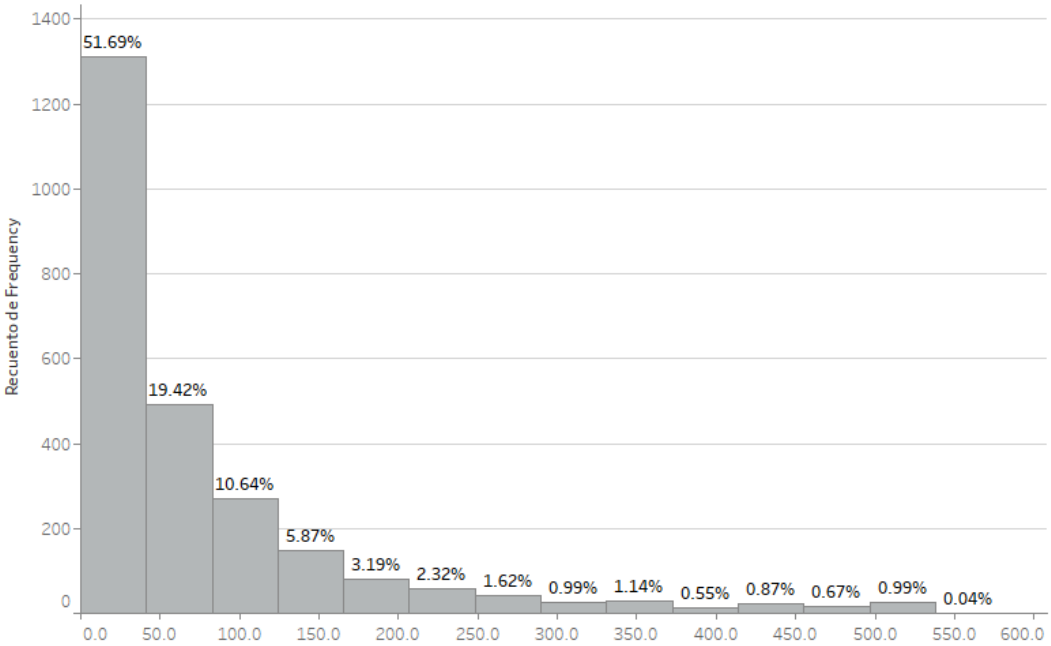
Para ambas variables, se aplicaron pruebas Kolmogórov-Smirnov probando la normalidad de las distribuciones. En la tabla 8 se puede presentan los p-valores de las pruebas de normalidad, donde se determinó que las variables en cuestión no presentaban un comportamiento normal.

3.1.2. Análisis de la variable frecuencia (F)

La variable frecuencia, como se puede observar en la figura 10 a continuación, mostró una alta acumulación de clientes en los niveles más bajos de frecuencia, y menores cantidades de clientes mayor frecuencia.

Figura 10.

Histograma de frecuencia (F)



Nota. Histograma de la variable frecuencia (F), con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

Gráficamente se observó la falta de normalidad de la variable, pues su distribución fue más cercana a una distribución exponencial. También se observó en las medidas de tendencia central de la variable, pues en promedio, los clientes profesionales compraron cada 77.1 días, con una desviación estándar de 99.5 días.

Esto representó un coeficiente de variación de 129.1 %, lo que hizo evidente que utilizar esta variable en su forma básica no sería adecuado.

Tabla 9.

Caracterización de la variable frecuencia (F)

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	77.1	99.5	129.1 %

Nota. Media, desviación estándar y coeficiente de variación para variable frecuencia (F). Elaboración propia, realizado con Word.

Dados los resultados anteriormente presentados, se procedió a realizar una transformación de la variable permitió reducir el error y aproximarse más a la distribución normal.

3.1.2.1. Transformación logarítmica de la variable frecuencia (F)

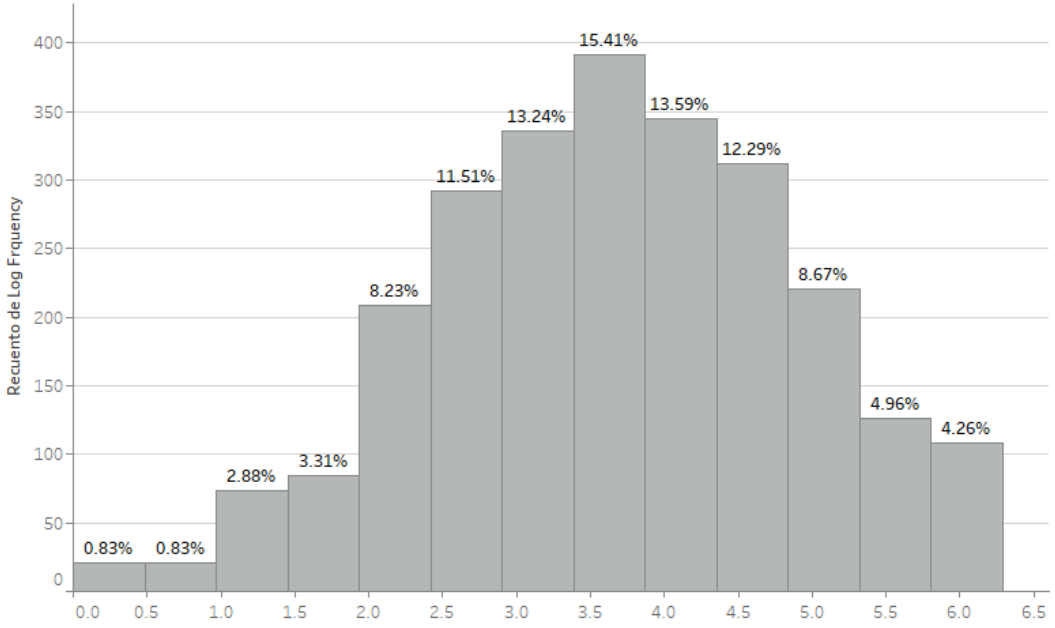
Al igual que con la variable reciencia, la frecuencia también fue transformada de forma logarítmica.

El resultado del histograma presentado en la figura 11 permite identificar gráficamente como la distribución de la variable transformada tuvo una forma similar a la curva normal, con las cantidades de clientes más altas alrededor de las medidas de tendencia central; sin embargo, se nota cierta acumulación de observaciones en los valores más altos de la variable.

Fue necesario corroborar las observaciones del punto anterior utilizando pruebas adecuadas para determinar la normalidad de los datos transformados.

Figura 11.

Histograma de (F) con transformación logarítmica



Nota. Histograma de la variable frecuencia (F) con transformación logarítmica, con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

El comportamiento descrito en el párrafo anterior también se pudo verificar observando las medidas de tendencia central en la tabla 10, donde el promedio de la variable transformada, que tuvo un valor de 3.653, coincidió con el punto de mayor acumulación de clientes del histograma.

Adicionalmente, los extremos de la distribución no presentaron una cantidad considerablemente alta de valores atípicos. Por ello, la transformación de la variable también permitió que la variabilidad relativa a la media se disminuyera a solo 34.0 %, principalmente ocasionado por tener más observaciones alrededor de la media.

Tabla 10.

Caracterización de la variable frecuencia (F) con transformación logarítmica

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	3.653	1.243	34.0 %

Nota. Media, desviación estándar y coeficiente de variación para variable frecuencia (F) con transformación logarítmica. Elaboración propia, realizado con Word.

3.1.2.2. Análisis de normalidad para variable frecuencia (F)

Finalmente, fue necesario utilizar pruebas de normalidad para determinar la versión de la variable más adecuada para utilizarse en los métodos de agrupación por medio de pruebas de normalidad Kolmogórov-Smirnov. Los resultados descritos en la tabla 11 muestran los p-valores de la prueba.

Tabla 11.

P-valores de prueba de normalidad para frecuencia (F)

Variable	P valor de prueba de normalidad	
	Variable original	Variable transformada
Frecuencia	0.000	0.211

Nota. P-valores de las pruebas Kolmogórov-Smirnov para probar la normalidad de frecuencia (F) en su versión original y transformada. Elaboración propia, realizado con Word.

Coincidiendo con el comportamiento gráfico de la variable original, se rechazó la posibilidad de que esta tuviera un comportamiento normal.

Por otro lado, la transformación logarítmica obtuvo un p-valor que el valor de significancia. Los resultados anteriores permitieron concluir que el uso de una transformación logarítmica de la variable frecuencia fue la alternativa adecuada.

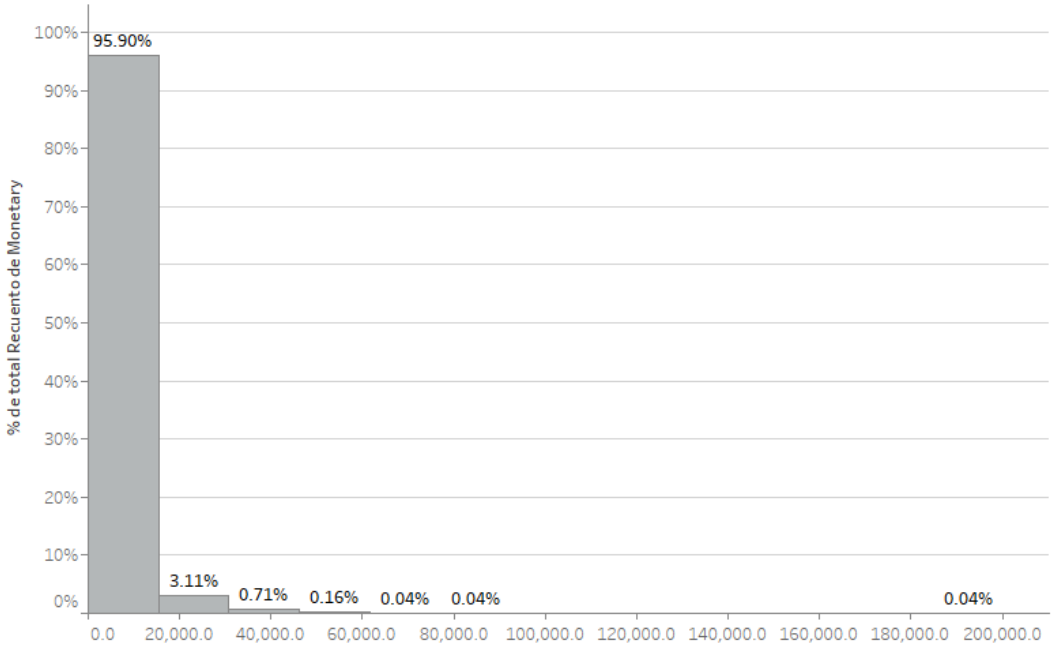
3.1.3. Análisis de la variable monto promedio (M)

En la figura 12, se presenta el histograma de la variable monto promedio. Es destacable que la primera clase del histograma concentra el 95.9 % de toda la clientela, revelando un patrón de distribución altamente sesgado. Este comportamiento atípico se vuelve aún más evidente al considerar la presencia significativa de segmentos sin observaciones.

Este fenómeno sugiere la presencia de un valor atípico que podría distorsionar el análisis, planteando la posibilidad de su exclusión para obtener resultados más precisos y representativos.

Figura 12.

Histograma de monto promedio (M)



Nota. Histograma de la variable monto promedio (M), con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

La tabla 12 muestra la distribución de frecuencias de la variable monto promedio (M), y proporciona una visión detallada de la variabilidad en el monto promedio de clientes. Se observa que la primera clase, con un rango de 0 a 15,402, engloba a la gran mayoría de la clientela, representando el 95.90 % del total. Las clases subsiguientes muestran una disminución significativa en la cantidad de clientes, destacando la presencia de clases con densidades mínimas.

Este patrón, sumado al hecho de que varias clases no contienen observaciones, sugiere una concentración desigual de clientes en distintos

rangos de monto promedio. La inclusión de clases con densidades mínimas indica la posible presencia de valores atípicos, respaldando la necesidad de una evaluación más profunda y la consideración de excluirlos para un análisis más robusto y preciso.

Tabla 12.

Frecuencias de monto promedio de compras (M)

Límite inferior	Límite superior	Cantidad de clientes	Densidad
0	15 402	2 434	95.90 %
15 403	30 804	79	3.11 %
30 805	46 206	18	0.71 %
46 207	61 608	4	0.16 %
61 609	77 011	1	0.04 %
77 012	92 413	1	0.04 %
92 414	107 815	0	0 %
107 816	123 218	0	0 %
123 219	138 620	0	0 %
138 621	154 023	0	0 %
154 024	169 425	0	0 %
169 426	184 827	0	0 %
184 828	200 230	1	0.04 %

Nota. Tabla de distribución de frecuencias para variable monto promedio (M), mostrada también en el histograma anterior. Elaboración propia, realizado con Word.

Como se observa en la tabla 13, la variable monto promedio de compra mostró una media de 4,423.1 con una desviación estándar de 6,938.4; esto generó un coeficiente de variación del 156.9 %.

Tabla 13.

Caracterización de la variable monto promedio

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	4 423.10	6 938.40	156.9 %

Nota. Media, desviación estándar y coeficiente de variación para variable monto promedio (M). Elaboración propia, realizado con Word.

Dicha variabilidad, que fue principalmente ocasionada por los valores extremos mostrados de la distribución, provocó también un alejamiento importante de la normalidad. Ante tal problemática, se tomaron dos posibilidades: eliminar los datos extremos (como la observación en la última clase del histograma), o bien, realizar una transformación para facilitar el análisis de dicha variable.

3.1.3.1. Transformación logarítmica de la variable monto promedio (M)

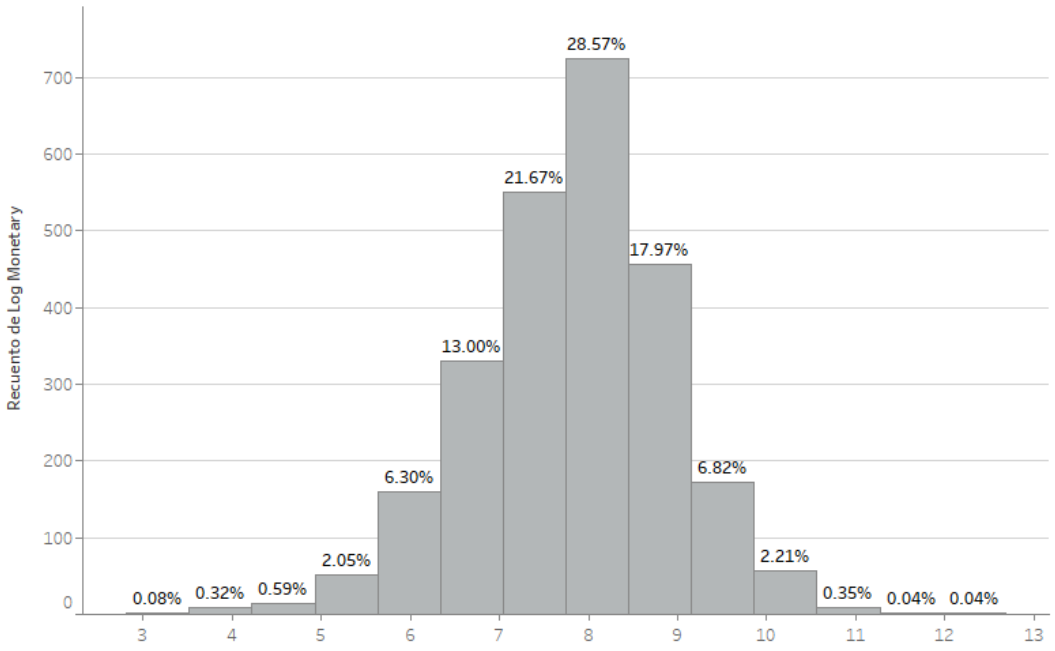
En la misma línea de análisis realizado sobre las variables reciencia y frecuencia, se realizó una transformación logarítmica sobre esta variable. Dada la relación explicada en la tabla de frecuencias, se encontró potencial de mejorar en el comportamiento general de la variable.

Los resultados de la conversión de dicha variable a su forma logarítmica se muestran en el histograma de la figura 13.

En este caso, se observó como la transformación de la variable permitió un comportamiento cercano a la distribución normal. Adicionalmente, permitió reducir el impacto de los casos más extremos de la distribución haciendo que la mayoría de los datos se encuentren agrupados alrededor de las medidas de tendencia central (esto se observa en la acumulación del 94.06 % de los datos en las seis clases centrales del histograma).

Figura 13.

Histograma de monto promedio (M) con transformación logarítmica



Nota. Histograma de la variable monto promedio (M) con transformación logarítmica, con su correspondiente distribución porcentual. Elaboración propia, realizado con Tableau.

De la misma forma que se observó en el histograma, se reflejaron dichos cambios en las medidas de tendencia central, que se muestran en la tabla 14. En

este análisis, se observó como la transformación de la variable redujo el coeficiente de variación a un 14.1 %.

Este cambio en la variable permite que los datos se condensaran, haciéndolos más adecuados para su análisis, así como para la aplicación de algoritmos de agrupación (como el algoritmo de K medias).

Tabla 14.

Caracterización de la variable monto promedio (M) con transformación logarítmica

Tipo de cliente	Media	Desviación estándar	Coeficiente de variación
Cliente profesional	7.82	1.10	14.1 %

Nota. Media, desviación estándar y coeficiente de variación para variable monto promedio (M) con transformación logarítmica. Elaboración propia, realizado con Word.

3.1.3.2. Análisis de normalidad para variable monto promedio (M)

Los p-valores resultantes de las pruebas Kolmogórov-Smirnov aplicadas a las variables originales y transformadas se encuentran en la tabla 15.

Los resultados de estas pruebas de normalidad revelaron, como era esperado, que la variable en su versión original no cumplía con los parámetros para considerarse normal. Sin embargo, la variable transformada tampoco tiene un p-valor suficientemente alto por lo que se rechazó la hipótesis que afirmaba su normalidad.

Tabla 15.

P-Valores de pruebas de normalidad para variable monto promedio

Variable	P valor de prueba de normalidad	
	Variable original	Variable transformada
Frecuencia	0.000	0.001

Nota. P-valores de las pruebas Kolmogórov-Smirnov para probar la normalidad de monto promedio (M) en su versión original y transformada. Elaboración propia, realizado con Word.

3.1.4. Análisis de correlación entre variables reciencia (R), frecuencia (F) y monto promedio (M)

Para garantizar el funcionamiento adecuado de los métodos de agrupamiento (como el algoritmo de K medias) es necesario evaluar a priori la correlación entre las variables tratadas; que puede realizarse desde la perspectiva paramétrica o no paramétrica.

Esto se definió aplicando una prueba de normalidad de Kolmogorov-Smirnov a cada variable en su versión original y también con su transformación logarítmica.

Las pruebas de normalidad revelaron que ninguna de las variables tuvo un comportamiento normal en su versión original; y al aplicar una transformación logarítmica solamente la frecuencia se ajustó a la distribución normal estándar.

Los resultados de dichas pruebas de normalidad, resumidos en el valor de p prueba Kolmogorov-Smirnov se muestran en la tabla 16.

Tabla 16.

Resumen de p-valores de pruebas de normalidad

Variable	P Valor para Variables Originales	P Valor para Variables Transformadas
Frecuencia	0.000	0.211
Monto promedio	0.000	0.001
Reciencia	0.000	0.000

Nota. P-valores de las pruebas Kolmogórov-Smirnov para probar la normalidad de las variables frecuencia (F), monto promedio (M) y reciencia (R) en su versión original y transformada. Elaboración propia, realizado con Word.

La conclusión anterior fue útil para elegir el método de análisis de correlación. Dado que las variables no mostraron una distribución normal, se eligió utilizar el coeficiente de correlación de Spearman.

Adicionalmente, durante el análisis gráfico preliminar utilizando diagramas de dispersión, se notó que utilizar las variables sin transformación existía una fuerte distorsión en los resultados (causado por la diferencia en la escala de la variable monto promedio comparada con la escala de la frecuencia y reciencia de compra).

Los resultados de la matriz de correlación mostrados en la tabla 17, reflejan que sí existe una correlación positiva entre las variables reciencia y monto promedio, aunque el valor del coeficiente de correlación se encuentra en un rango que puede considerarse bajo.

Tabla 17.

Matriz de correlación de Spearman para variables RFM

	Frecuencia	Monto Promedio	Reciencia
Frecuencia	1.0000	0.0553	0.0117
Monto Promedio	0.0553	1.0000	0.2125
Reciencia	0.0117	0.2125	1.0000

Nota. Matriz de coeficientes de correlación de Spearman para la interacción de variables reciencia (R), frecuencia (F) y monto promedio (M) con transformación logarítmica. Elaboración propia, realizado con Tableau.

Por otro lado, los coeficientes entre las variables reciencia y frecuencia y entre frecuencia y monto medio son sustancialmente son más cercanos a cero que el caso anterior.

Para determinar la significancia de estos últimos, se planteó una prueba de hipótesis, planteando como hipótesis nula la inexistencia de correlación entre cada par de variables ($H_0: r_{spm} = 0; H_a: r_{spm} \neq 0$). Los resultados de dichas pruebas de hipótesis se muestran en la tabla 18.

Tabla 18.

P-valores de coeficientes de Spearman en variables RFM

	Frecuencia	Monto Promedio	Reciencia
Frecuencia	1.000	0.005	0.556
Monto Promedio	0.005	1.000	0.000
Reciencia	0.556	0.000	1.000

Nota. Matriz de p-valores para la significancia del coeficiente de correlación de Spearman para la interacción de variables reciencia (R), frecuencia (F) y monto promedio (M) con transformación logarítmica. Elaboración propia, realizado con Tableau.

Al analizar los resultados en forma de p-valores para estas pruebas, no se rechazó la hipótesis nula para el coeficiente de Spearman entre las variables reciencia y frecuencia.

Para los pares de variables restantes no fue posible concluir de esta manera, ya que su p-valor fue menor que un nivel de significancia del 0.05. Con ello, se concluyó con la existencia de correlación significativa entre la reciencia y el monto promedio de compras, y entre la frecuencia y el monto promedio.

A pesar de que la prueba de hipótesis reveló que la variable monto promedio sí se correlaciona con la reciencia y la frecuencia, los valores del coeficiente de Spearman se clasificaron como débiles. Esto implicó que, aunque

sí existió dependencia entre las variables, el efecto que tuvieron entre sí no incrementó la multicolinealidad para la creación del modelo de regresión logística.

3.1.5. Agrupación de clientes utilizando método de percentiles

El primer objetivo implicó la agrupación de los clientes utilizando métodos de agrupación por K medias, sin embargo, la construcción del modelo de regresión logística también tomó en consideración la segmentación tradicional por medio de percentiles. Esto por su capacidad de aportar información acerca de cada variable de forma independiente.

3.1.5.1. Agrupación percentil de variables reciencia (R), frecuencia (F) y monto promedio (M)

Para las variables reciencia, frecuencia y monto promedio de compras se realizaron agrupaciones basadas en la posición relativa de los clientes según cada variable (deciles y percentiles agrupados). Se clasificaron en el grupo 1 a los primeros dos deciles, como grupo 2 a los siguientes dos deciles y, por último, los seis deciles con peor calificación de cada variable como grupo 3.

En la tabla 19 se observan las segmentaciones para cada variable. De dicha tabla se puede destacar que para las variables frecuencia, no existió un patrón claro de mejora en la retención de clientes al tener mayores frecuencias de compra.

Para la reciencia de compras se observó como la tasa de retención de los grupos R1 (77.1 %), R2 (57.0 %) y R3 (31.6 %) disminuyeron a medida que aumentó la posición relativa de estos en la variable.

Este mismo comportamiento se observa para la variable monto promedio de compras en menor escala, pues la tasa de retención tiene un máximo de 55.3 % en el grupo M1, y un mínimo de 41.0 % en el grupo M3.

Tabla 19.

Segmentación por percentiles de variables RFM

Variable	Segmento	Cantidad de clientes	Tasa de retención
Reciencia	R1	510	77.1 %
	R2	507	57.0 %
	R3	1 521	31.6 %
Frecuencia	F1	548	44.7 %
	F2	468	51.7 %
	F3	1 522	44.3 %
Monto promedio	M1	508	55.3 %
	M2	507	50.5 %
	M3	1 523	41.0 %

Nota. Segmentos percentiles para las variables reciencia (R), frecuencia (F) y monto promedio (M) con su correspondiente cantidad de clientes y tasa de retención de clientes. Elaboración propia, realizado con Word.

A pesar de los patrones que se presentaron entre la posición relativa de los clientes y la tasa de retención, no fue posible concluir estadísticamente sobre la influencia de esta segmentación sobre la variable de respuesta partiendo únicamente desde su tasa de retención. Para ello, fue necesario el uso de pruebas de independencia de datos.

3.1.5.2. Interacción entre segmentación percentil

Dado que la segmentación de cada variable en tres diferentes grupos, hay un espacio de veintisiete posibles agrupaciones para el cliente. Desde el punto de vista mercadológico, no es útil manejar una variable de segmentación compuesta por altas cantidades de subgrupos, pues implicaría que en cada campaña se deba utilizar un ángulo de comunicación diferente para cada agrupación. Por ello, en la práctica, se suelen buscar variables de segmentación con menos categorías para facilitar la personalización masiva.

3.1.5.2.1. Interacción entre reciencia y frecuencia (RF)

En la tabla 20 se muestran los resultados obtenidos al combinar las segmentaciones percentiles de las variables reciencia y frecuencia. Al analizar la interacción de estas variables se obtuvieron agrupaciones de clientes más pequeñas y heterogéneas, que manifestaron diferentes niveles de respuesta ante la variable objetivo. Esto se puede observar con la agrupación R1F2, que tuvo la máxima tasa de retención con 86.5 % y un mínimo en R3F1 con 23.3 %.

Tabla 20.

Interacción de variables reciencia (R) y frecuencia (F)

Segmento	Cantidad de clientes	Tasa de retención
R1 F2	104	86.5 %
R1 F1	126	81.7 %
R1 F3	280	71.4 %
R2 F1	109	63.3 %
R2 F3	303	55.4 %
R2 F2	95	54.7 %

Continuación de la tabla 20.

Segmento	Cantidad de clientes	Tasa de retención
R3 F2	269	37.2 %
R3 F3	939	32.7 %
R3 F1	313	23.3 %

Nota. Detalle de segmentos obtenidos por la interacción de las agrupaciones percentiles de reciencia (R) y frecuencia (F), con su correspondiente cantidad de clientes y tasa de retención. Elaboración propia, realizado con Word.

A pesar de los puntos expuestos anteriormente, se debe aclarar que no se obtuvo una secuencia clara en el ordenamiento de la variable frecuencia. Esto fue un indicador de que la frecuencia tiene una menor incidencia en la variable objetivo del modelo.

3.1.5.2.2. Interacción entre reciencia y monto promedio (RM)

La interacción entre la variable reciencia y monto promedio (RM), mostrada en la tabla 21. En dicha tabla se muestran resultados similares a la combinación de reciencia y frecuencia (RF) con una tasa de retención máxima de 85.5 % y mínima de 30.3 %.

En este caso, fue útil observar cómo ambas variables están fuertemente relacionadas con la variable de respuesta, pues los segmentos de la variable reciencia mostraron un ordenamiento claro (los segmentos con mejor valor de reciencia tuvieron mejores tasas de retención), y dentro de estos, hay un ordenamiento claro de la variable monto promedio (los clientes con mejor monto promedio de compra regresaron a comprar en mayor proporción).

Tabla 21.

Interacción de variables reciencia (R) y monto promedio (M)

Segmento	Cantidad de clientes	Tasa de retención
R1 M1	131	85.5 %
R1 M2	128	82.0 %
R1 M3	251	70.1 %
R2 M1	117	67.5 %
R2 M2	115	54.8 %
R2 M3	275	53.5 %
R3 M1	260	34.6 %
R3 M2	264	33.3 %
R3 M3	997	30.3 %

Nota. Detalle de segmentos obtenidos por la interacción de las agrupaciones percentiles de reciencia (R) y monto promedio (M), con su correspondiente cantidad de clientes y tasa de retención. Elaboración propia, realizado con Word.

3.1.5.2.3. Interacción entre frecuencia y monto promedio (FM)

Por último, se analizó la interacción entre variables frecuencia y monto promedio, cuyos resultados se pueden observar en la tabla 22.

En este análisis se obtuvieron resultados menos concluyentes que en el caso de las interacciones que incluyen al menos en una ocasión la variable reciencia. En gran medida, se debe a que la variable reciencia explica una gran parte de la variable de respuesta. Esto se reflejó en el rango observado de la tasa de retención, que tuvo un valor máximo del 63.0 % (segmento F1M1) y un mínimo del 37.2 % segmento (F3M3).

Tabla 22.

Interacción de variables frecuencia (F) y monto promedio (M)

Segmento	Cantidad de clientes	Tasa de retención
F1 M1	100	63.0 %
F1 M2	107	51.4 %
F1 M3	341	37.2 %
F2 M1	105	58.1 %
F2 M2	110	57.3 %
F2 M3	253	46.6 %
F3 M1	303	51.8 %
F3 M2	290	47.6 %
F3 M3	929	40.9 %

Nota. Detalle de segmentos obtenidos por la interacción de las agrupaciones percentiles de frecuencia (F) y monto promedio (M), con su correspondiente cantidad de clientes y tasa de retención. Elaboración propia, realizado con Microsoft Word.

A pesar de que las diferencias entre agrupaciones fueron relativamente pequeñas, sí se observa una ordinalidad entre ellas (la combinación de mejores valores de la variable frecuencia (F1) y de la variable monto promedio (M1) tuvo una tasa de conversión más alta).

3.1.6. Agrupación de clientes utilizando algoritmo de agrupación por K medias

El objetivo del análisis de los clientes desde su comportamiento transaccional fue obtener sus agrupaciones óptimas para predecir las variables de respuesta. Para ello, se aplicó el algoritmo de simulación por K medias, obteniendo las agrupaciones con menor error cuadrático.

Debido a que las escalas entre las variables reciencia, frecuencia y monto promedio son considerablemente diferentes, se vio la necesidad de utilizar la transformación logarítmica que se obtuvo de las variables de interés en la fase de análisis de normalidad.

3.1.6.1. Determinación de la cantidad óptima de centroides

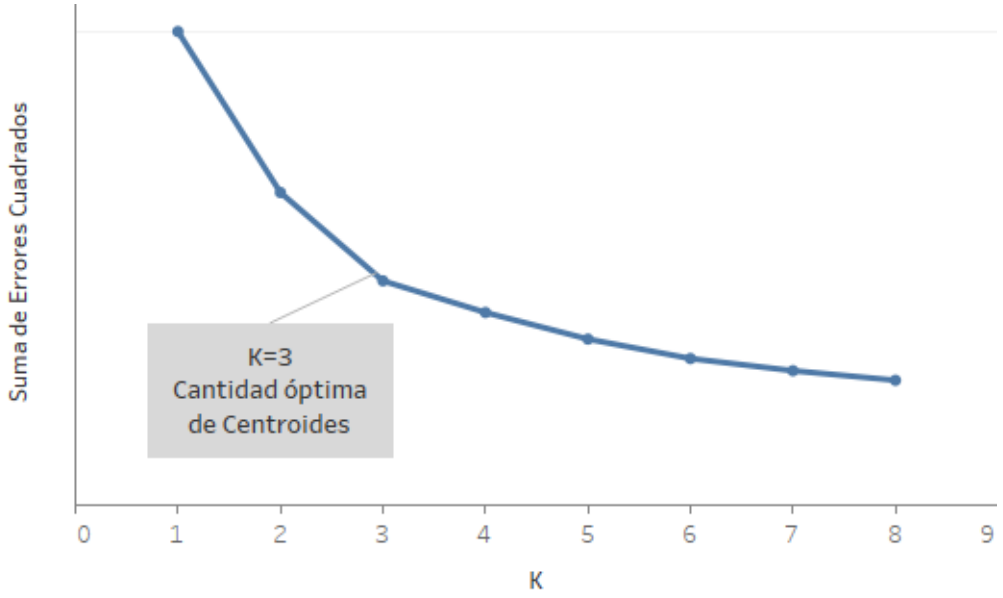
En el proceso de aplicar el algoritmo de agrupación por K medias, la definición precisa de la cantidad de centroides se erige como uno de los parámetros cruciales. Se llevó a cabo un ciclo de iteración del algoritmo de agrupación, explorando valores de K que abarcaban desde 1 hasta 8. Este procedimiento se realizó con mil iteraciones en cada caso, garantizando así un análisis robusto y completo.

La representación visual de los resultados de este proceso se plasma en la figura 14, donde el eje horizontal exhibe los diferentes valores de K aplicados, mientras que el eje vertical muestra la suma de los errores cuadráticos acumulados durante el desarrollo de la agrupación. Como era de prever, al incrementar la cantidad de centroides, se observa una reducción en las distancias entre cada punto y su centro de gravedad, lo que se traduce en una disminución del error total del modelo de agrupamiento.

Sin embargo, resulta notable que valores de K superiores a 3 no conllevan mejoras sustanciales en la reducción del error. Este hallazgo condujo a la conclusión de que la cantidad óptima de centroides es K=3, dado que ofrece un equilibrio eficiente entre la precisión del modelo y la complejidad computacional asociada.

Figura 14.

Comparativa del SSE usando múltiples valores de K



Nota. Comparativa de la suma del error cuadrático (SSE) iterando el algoritmo de agrupación por K medias con valores de K desde 1 hasta 8 utilizando variables reciencia (R), frecuencia (F) y monto promedio (M) con transformación logarítmica. Elaboración propia, realizado con Tableau

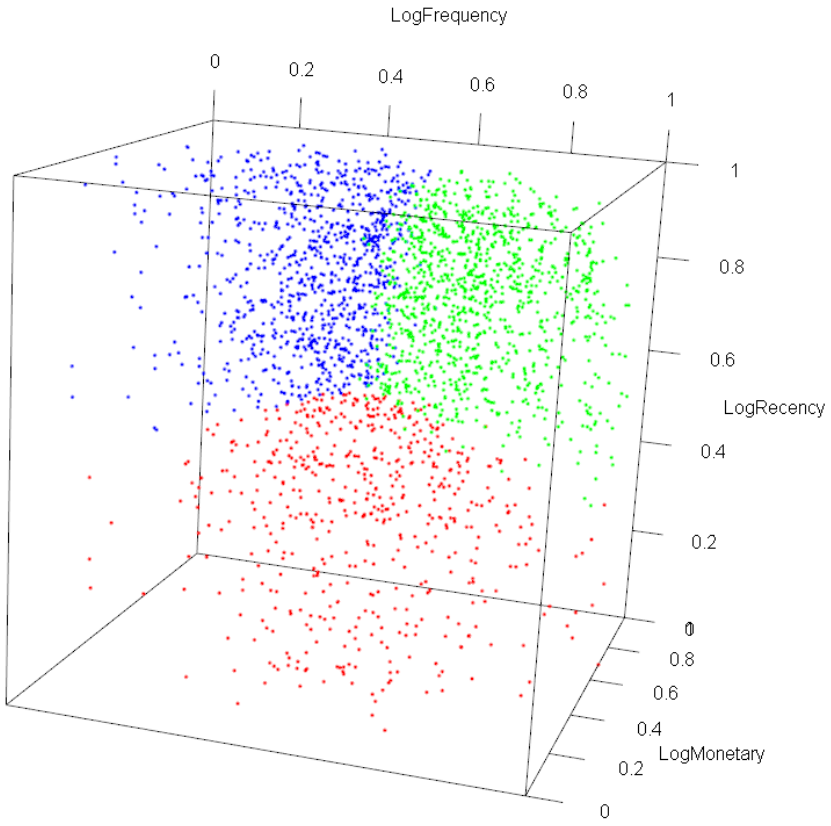
3.1.6.2. Segmentación de clientes por algoritmo de agrupación por K medias

Al realizar la agrupación con tres centroides, se obtuvieron los resultados mostrados en la figura 15, donde cada eje representa las variables frecuencia (LogFrequency), reciencia (LogRecency) y monto promedio (LogMonetary) de compras en su forma transformada (para fines ilustrativos, dichos ejes fueron re escalados para tener un rango de 0 a 1); cada punto representa a un cliente y los colores representan a los segmentos encontrados por medio del algoritmo de agrupación por K medias con K=3.

Como se puede observar gráficamente, el algoritmo de agrupación permitió identificar a aquellos clientes con comportamientos similares utilizando únicamente la posición relativa de sus variables, y generó tres segmentos o agrupaciones diferentes.

Figura 15.

Agrupación de clientes usando algoritmo de K medias



Nota. Segmentos obtenidos utilizando el algoritmo de K medias con 3 centroides en variables reciencia (R), frecuencia (F) y monto promedio (M), todas las variables en su versión transformada. Elaboración propia, realizado con Rstudio.

Estas agrupaciones se pueden observar detalladamente en la tabla 23, donde se denota el resultado de cada segmento en términos de su tasa de retención. Se observa cómo el segmento denominado Segmento B, incluyó a un total de 555 clientes, que tuvieron una tasa de retención del 76.6 %. Es importante observar cómo este segmento de clientes tuvo clientes con compras recientes; pues su reciencia promedio es de 12.5 días.

Los segmentos denominados Segmento C y Segmento A, tuvieron tamaños y tasas de retención similares, pero presentaron diferencias significativas en el valor de frecuencia. En esta variable, el segmento C tuvo una frecuencia de 138.4 días, mientras que el segmento A, mostró una frecuencia de 17.0 días.

Tabla 23.

Resumen de segmentación por K medias

Segmento	Cantidad de clientes	Frecuencia promedio	Monto promedio	Reciencia promedio	Tasa de retención
B	555	57.9	5 743.50	12.50	76.6 %
C	1 069	138.4	4 272.90	167.1	37.0 %
A	914	17.0	3 769.90	199.5	36.2 %

Nota. Cantidad de clientes, media de frecuencia (F), media de monto promedio (M), media de reciencia y tasa de retención para cada segmento obtenido en la agrupación por K medias con 3 centroides. Elaboración propia, realizado con Word.

Estos resultados son consistentes con obtenidos en el análisis individual de las variables reciencia, frecuencia y monto promedio, donde se pudo observar que la tasa de retención de clientes (que es una representación de la variable de

respuesta) se ve alterada principalmente con los cambios en los niveles de la variable reciencia.

3.2. Objetivo 2: identificar las variables de clientes que interfieren en la pérdida o retención de clientes, usando pruebas de independencia y pruebas de correlación

El uso de variables no correlacionadas en la construcción de modelos predictivos tiende a sobre explicar el objeto de estudio, logrando un ajuste artificial. Por esta razón, la elección de las variables fue una de las fases claves en la construcción del modelo de regresión, que fue posible, analizando la correlación o independencia que existió entre cada variable de segmentación y la variable de respuesta (pérdida o retención de cada cliente).

La mayoría de las variables analizadas desde la perspectiva de correlación fueron de tipo categórica, pues se obtuvieron de los datos de segmentación de los clientes.

En los casos donde las variables originales eran de tipo cuantitativo, se buscó crear agrupaciones para convertirlas en datos categóricos. Este tratamiento de datos se fundamentó, primero, en que la variable de respuesta es de tipo dicotómica, denotando la retención o pérdida del cliente; y segundo, que operativamente los equipos de mercadeo en la empresa minorista de materiales de construcción suelen usar agrupaciones antes que valores continuos.

3.2.1. Método de envío preferido

El método de envío preferido se obtuvo de la información transaccional de los clientes existiendo tres opciones: envíos en ruta, recolección en la tienda o

una combinación de estos. La tabla de contingencia resultante se puede observar en la tabla 24.

La aplicación de la prueba de independencia a la tabla de contingencia mostrada registró un p-valor menor a 0.00001. Con ello, fue posible rechazar la hipótesis nula que afirma la independencia entre esta variable y la variable de respuesta.

Tabla 24.

Tabla de contingencia para el método de envío

	Perdido	Retenido	Total
Múltiples	914	924	1 838
Recoge en tienda	401	218	619
Ruta	61	20	81
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con el método de envió preferido. Elaboración propia, realizado con Word.

3.2.2. Condición de pago

De la misma forma, se analiza la independencia entre la variable objetivo y la condición de pago.

La aplicación de la prueba de independencia resultó en un p-valor de 0.0276. Al compararlo con un nivel de significancia de 0.05, se pudo concluir en rechazar la hipótesis nula, afirmando la relación entre la condición de pago y la variable de respuesta.

Sin embargo, bajo valores de significancia más estrictos esta suposición tuvo una conclusión diferente. Por esta razón, la variable de condición de pago, a pesar de mostrar cierta relación, fue candidata a eliminarse durante el análisis de multicolinealidad.

En la tabla 25 se puede observar la tabulación cruzada entre la variable condición de pago y la variable de respuesta.

Tabla 25.

Tabla de contingencia para la condición de pago

	Perdido	Retenido	Total
Crédito 8 días	16	14	30
Crédito 15 días	3	3	6
Crédito 30 días	40	61	101
Contado	1 317	1 084	2 401
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con la condición de pago. Elaboración propia, realizado con Word.

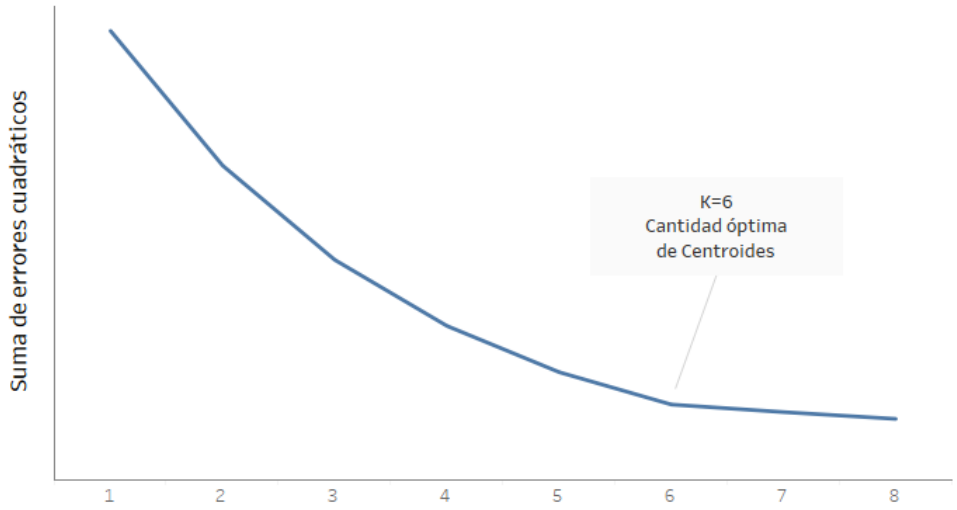
3.2.3. Antigüedad de cotización

La antigüedad de cotización se obtuvo de la diferencia en días entre la fecha de última cotización y la fecha en la que se ejecutaron los cálculos. Posterior al análisis inicial de esta variable, se decidió hacer una agrupación utilizando el algoritmo de K medias.

Como se observa en la figura 16, la cantidad óptima de centroides para esta agrupación es de 6 segmentos.

Figura 16.

Error cuadrático en agrupación de antigüedad de cotización



Nota. Comparativa de la suma del error cuadrático (SSE) iterando el algoritmo de agrupación por K medias con valores de K desde 1 hasta 8 utilizando la variable antigüedad de cotización. Elaboración propia, realizado con Tableau.

El resultado de la tabulación cruzada entre los seis segmentos de antigüedad de cotización y la variable de respuesta se pueden observar en la tabla 26. Cabe mencionar que el nombramiento de dichos segmentos fue aleatorio, por lo que no tiene una ordinalidad respecto a las tasas de retención observadas.

Al analizar la tabla de contingencia por medio de la distribución chi cuadrado, se obtuvo un p-valor menor a 0.0001, con lo que se rechazó la hipótesis nula aceptando que existe asociación entre las variables.

Tabla 26.

Tabla de contingencia para la antigüedad de cotización

	Perdido	Retenido	Total
A1	348	688	1 036
A2	376	120	396
B1	226	69	295
B2	342	248	590
C1	123	21	144
C2	61	16	77
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos de antigüedad de cotización. Elaboración propia, realizado con Word.

3.2.4. Sucursal origen

La sucursal de origen fue obtenida al encontrar la sucursal donde se generó la primera transacción del cliente.

Los resultados de la tabla de contingencia se pueden observar en la tabla 27, donde luego de calcular las frecuencias esperadas y compararlas con las frecuencias reales, se obtuvo un p-valor de 0.0001. Con ello, se determinó que ambas variables sí presentaron relación significativa.

Tabla 27.

Tabla de contingencia para sucursal origen

	Perdido	Retenido	Total
Sucursal 1	1	0	1
Sucursal 2	5	6	11
Sucursal 3	8	41	99
Sucursal 4	45	34	79
Sucursal 5	35	29	64
Sucursal 6	3	0	3
Sucursal 7	131	94	225
Sucursal 8	46	49	95
Sucursal 9	15	23	38
Sucursal 10	29	28	57
Sucursal 11	2	2	4
Sucursal 12	38	36	74
Sucursal 13	51	27	78
Sucursal 14	1	1	2
Sucursal 15	25	38	63
Sucursal 16	71	46	117
Sucursal 17	41	42	83
Sucursal 18	82	99	181
Sucursal 19	48	27	75
Sucursal 20	6	15	21
Sucursal 21	31	35	66
Sucursal 22	70	35	105

Continuación de la tabla 22.

	Perdido	Retenido	Total
Sucursal 23	6	3	9
Sucursal 24	53	57	110
Sucursal 25	14	16	30
Sucursal 26	25	24	49
Sucursal 27	74	77	151
Sucursal 28	8	12	20
Sucursal 29	86	51	137
Sucursal 30	56	21	77
Sucursal 31	9	13	22
Sucursal 32	83	82	165
Sucursal 33	128	99	227
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con la sucursal de origen. Elaboración propia, realizado con Word.

3.2.5. Canal de origen

El canal de origen fue obtenido mediante la agrupación de las sucursales presentadas en la variable anterior. Su tabla de contingencia se puede observar en la tabla 28.

El análisis de independencia entre esta variable y la variable de respuesta reflejó un p-valor de 0.4301; con lo que no fue posible rechazar la hipótesis nula, reflejando que el canal y la variable de respuesta fueron independientes.

Tabla 28.

Tabla de contingencia para variable canal de origen

	Perdido	Retenido	Total
Digital	35	29	64
Proyectos	53	57	110
Retail	1 288	1 076	2 364
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con el canal de origen. Elaboración propia, realizado con Word.

3.2.6. Reciencia (R) percentil

La tabla de contingencia que compara los segmentos obtenidos por agrupación de percentiles de la variable reciencia (R) con la variable de respuesta se muestra en la tabla 29.

Tabla 29.

Tabla de contingencia para variable reciencia (R)

	Perdido	Retenido	Total
R1	117	393	510
R2	218	289	507
R3	1 041	480	1 521
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos percentiles de la variable reciencia (R). Elaboración propia, realizado con Word.

De la tabla de contingencia anterior se puede destacar que el segmento R1 retuvo a la mayoría de sus clientes (77.05 % de clientes retenidos), a diferencia de los segmentos R2 (57.00 %) y R3 (31.56 %).

El comportamiento anteriormente descrito fue un indicio de la existencia de una relación significativa con la variable de respuesta. Esto se confirmó con el resultado mostrado por la prueba de independencia (p-valor<0.0001), que concluyó que sí existe una relación entre la agrupación percentil de la reciencia (R) y la variable de respuesta.

3.2.7. Frecuencia (F) percentil

De la misma forma, se analizó la agrupación percentil de la variable frecuencia de compra. La tabla 30, que muestra su correspondiente tabla de contingencia, reveló cierta homogeneidad al compararse con los resultados obtenidos con la variable reciencia (R).

Tabla 30.

Tabla de contingencia para variable frecuencia (F)

	Perdido	Retenido	Total
F1	303	245	548
F2	226	242	468
F3	847	675	1 522
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos percentiles de la variable frecuencia (F). Elaboración propia, realizado con Word.

La prueba de independencia aplicada sobre esta tabla de contingencia se obtuvo un p-valor de 0.0171; lo que permitió rechazar la hipótesis nula, aceptando la dependencia entre ambas variables.

Sin embargo, esta hipótesis fue contrastada con un nivel de significancia α=0.05, aplicando niveles más estrictos de confianza, el resultado varió. Por ello, esta variable se mantuvo como candidata a eliminarse del modelo en la fase de análisis de multicolinealidad y exactitud.

3.2.8. Monto promedio (M) percentil

En la tabla 31 se muestran los resultados de tabular las agrupaciones percentiles de la variable monto promedio (M) y la variable de respuesta de forma dicotómica. Al igual que lo sucedido con la variable reciencia (R), se observó una menor proporción de clientes retenidos en la agrupación M3 (41.04 %) al compararlo con los segmentos M2 (50.49 %) y M1 (55.31 %).

Tabla 31.

Tabla de contingencia para variable monto promedio (M)

	Perdido	Retenido	Total
M1	227	281	508
M2	251	256	507
M3	898	625	1 526
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos percentiles de la variable monto promedio (M). Elaboración propia, realizado con Word.

La prueba de independencia aplicada sobre estos datos reflejó que la probabilidad de que hubiera independencia entre ambas variables fue inferior a 0.0001. Con ello, se rechazó la hipótesis nula y se determinó que sí existió efecto significativo entre ambas variables.

3.2.9. RF percentil

En la práctica mercadológica de la empresa también puede utilizarse la interacción en pares de las variables reciencia (R), frecuencia (F) y monto promedio (M) como una variable de segmentación.

Tabla 32.

Tabla de contingencia para variables reciencia y frecuencia

	Perdido	Retenido	Total
R1 F1	23	103	126
R1 F2	14	90	104
R1 F3	80	200	280
R2 F1	40	69	109
R2 F2	43	52	95
R2 F3	135	168	303
R3 F1	240	73	313
R3 F2	169	100	269
R3 F3	632	307	939
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos de interacción entre variables reciencia (R) y frecuencia (F). Elaboración propia, realizado con Word.

La tabla 32 presenta los resultados obtenidos de hacer la tabulación cruzada de la interacción entre reciencia (R) y frecuencia (F). Se pudo observar que los segmentos con mejores puntuaciones de ambas variables retuvieron más clientes. Este comportamiento fue similar al del análisis individual de las variables, pero se amplificó por la cantidad de agrupaciones.

El p-valor obtenido en la prueba de independencia realizada a esta interacción fue menor a 0.0001. Con ello, fue posible concluir que sobre una relación de dependencia significativa entre ambas variables.

3.2.10. RM percentil

De la misma manera, se analizó la interacción entre las variables reciencia y monto promedio de compra agrupadas por el método percentil.

De este análisis, se obtuvo un p-valor inferior a 0.0001, con lo que se concluyó que sí existe una relación significativa entre la retención o pérdida de clientes y la segmentación generada por la interacción de estas variables.

Tabla 33.

Tabla de contingencia para variables reciencia y monto promedio de compra

	Perdido	Retenido	Total
R1 M1	19	112	131
R1 M2	23	105	128
R1 M3	75	176	251
R2 M1	38	79	117
R2 M2	52	63	115
R2 M3	128	147	275
R3 M1	170	90	260

Continuación de la tabla 33.

	Perdido	Retenido	Total
R3 M2	176	88	264
R3 M3	695	302	997
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos de interacción entre variables reciencia (R) y monto promedio (M). Elaboración propia, realizado con Word.

3.2.11. FM percentil

A continuación, se realizó el análisis de independencia de datos entre la combinación de la frecuencia (F) y el monto promedio de compra (M). La tabla de contingencia resultante se puede observar en la tabla 34. La aplicación de la prueba de independencia de datos para esta tabla de contingencia reflejó un valor de p menor a 0.0001, con lo que se rechazó la hipótesis que afirmaba la independencia entre ambas variables.

Tabla 34.

Tabla de contingencia para variables frecuencia y monto promedio de compra

	Perdido	Retenido	Total
F1 M1	37	63	100
F1 M2	52	55	107
F1 M3	214	127	341
F2 M1	44	61	105
F2 M2	47	63	110

Continuación de la tabla 34.

	Perdido	Retenido	Total
F2 M3	135	118	253
F3 M1	146	157	303
F3 M2	152	138	290
F3 M3	549	380	929
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos de interacción entre frecuencia (F) y monto promedio (M). Elaboración propia, realizado con Word.

3.2.12. Segmentación RFM por algoritmo de K medias

Por último, el análisis de independencia se realizó también con las segmentaciones obtenidas mediante el algoritmo de agrupación por K medias.. En la tabla 35 se pueden observar los resultados de esta tabulación.

Tabla 35.

Tabla de contingencia para segmentación obtenida con algoritmo de K medias

	Perdido	Retenido	Total
Segmento A	583	331	914
Segmento B	663	406	1 069
Segmento C	130	425	555
Total	1 376	1 162	2 538

Nota. Tabla de contingencia comparando la variable objetivo (retención o pérdida del cliente) con los segmentos obtenidos por el algoritmo de agrupación por K medias. Elaboración propia, realizado con Word.

Al aplicar una prueba de independencia sobre esta tabulación, se obtuvo un p-valor menor a 0.0001, con lo que se concluyó rechazando la hipótesis nula, indicando que sí existe relación entre estas variables.

3.3. Objetivo 3: relacionar las variables cuantitativas y cualitativas de los clientes profesionales, para construir un modelo de regresión logística que permita cuantificar las probabilidades de pérdida y compra de cada cliente profesional

Partiendo de los resultados del análisis de independencia de las variables regresoras, se construyó el modelo de regresión logística para inferir la probabilidad de retención o pérdida de los clientes profesionales de forma iterativa; generando varias combinaciones de dichas variables y comparándolos, utilizando matrices de confusión, criterios de información y criterios de multicolinealidad.

3.3.1. Construcción de modelos de regresión

Las variables examinadas durante el objetivo 2 fueron empleadas en la construcción de diversos modelos de regresión. Cada modelo fue meticulosamente elaborado, explorando las posibles interacciones entre las variables para evaluar su impacto en la predicción de la retención de clientes.

En la tabla 36, se presenta una matriz detallada que expone las distintas combinaciones de variables utilizadas, dando origen a un total de dieciocho modelos únicos.

En un análisis posterior, se profundizó sobre la multicolinealidad de los modelos, priorizando los que mostraron mejores niveles de exactitud.

Tabla 36.

Elección de variables regresoras

Modelo	Segmento	Metodo Envio	Cantidad Categorias	Sucursal	Segmento Cotización	Termino de Pago	Canal	F Percentil	M Percentil	R Percentil	RM	FM	RF
Modelo 1	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Modelo 2	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Modelo 3	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Modelo 4	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗
Modelo 5	✓	✓	✓	✓	✗	✓	✗	✗	✗	✗	✗	✗	✗
Modelo 6	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
Modelo 7	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗
Modelo 8	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗
Modelo 9	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
Modelo 10	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✗	✗
Modelo 11	✗	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✗	✗
Modelo 12	✗	✓	✓	✗	✓	✓	✓	✗	✗	✗	✓	✓	✓
Modelo 13	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
Modelo 14	✓	✓	✓	✗	✓	✓	✗	✓	✓	✓	✗	✗	✗
Modelo 15	✓	✓	✓	✗	✓	✓	✓	✗	✗	✗	✗	✗	✗
Modelo 16	✓	✗	✗	✗	✓	✗	✗	✗	✗	✗	✗	✗	✗
Modelo 17	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✗	✗	✗
Modelo 18	✗	✗	✗	✗	✓	✗	✗	✗	✗	✗	✓	✓	✓

Nota. Se muestran las diferentes combinaciones de variables utilizadas en la construcción de modelos de regresión logísticos. Elaboración propia, realizado con Tableau.

3.3.2. Análisis de exactitud

Los modelos obtenidos en el punto anterior fueron utilizados para obtener los valores predichos para cada modelo utilizando los datos del modelo de

entrenamiento, y estos se compararon con los resultados reales para hacer una matriz de confusión del modelo.

Los resultados de los modelos de predicción generan valores comprendidos entre 0 y 1, representando la probabilidad de retención de cada cliente. Para simplificar la interpretación, se contrastaron con un valor constante de 0.5, transformándolos en una variable dicotómica.

Para ello, se categorizaron los predichos, asignando un valor de retención cuando la probabilidad de retención superaba el umbral del 0.5, mientras que aquellos casos que no cumplían con esta regla se clasificaban como predicciones de pérdida.

Tabla 37.

Matriz de confusión para modelo 7

		Valores Reales		Total
		Perdido	Retenido	
Valores predichos	Perdido	1,058	463	1,521
	Retenido	318	699	1,017
Total		1,376	1,162	2,538

Nota. Matriz de confusión del modelo 7; comparando los valores reales de retención y los valores predichos del modelo. Elaboración propia, realizado con Word.

En la tabla 37 se muestran los resultados de la matriz de confusión para el modelo 7, que mostró un total de 1058 clientes con predicción de pérdida correctos y 699 clientes con predicción de retención correctos. Estos resultados se pueden interpretar como una precisión en la predicción del 69.39 %.

3.3.3. Elección de modelos

Al realizar el análisis de exactitud de todos los modelos fue posible hacer una comparativa que permitió identificar los modelos de mejor ajuste a los resultados reales. También se utilizó el criterio de información de Akaike (AIC) como una segunda guía para establecer la validez del modelo.

Tabla 38.

Resumen de exactitud y ajuste de los modelos de predicción

Modelo	AIC	Exactitud de predicción
Modelo 7	3,053	69.39 %
Modelo 8	3,052	38.99 %
Modelo 14	3,034	68.44 %
Modelo 17	3,054	68.28 %
Modelo 11	3,038	68.28 %
Modelo 12	3,033	68.20 %
Modelo 18	3,049	68.16 %
Modelo 10	3,037	68.16 %
Modelo 6	3,139	68.01 %
Modelo 13	3,034	67.97 %
Modelo 16	3,145	67.61 %
Modelo 15	3,123	67.30 %
Modelo 9	3,145	67.26 %
Modelo 4	3,433	59.85 %
Modelo 5	3,434	59.65 %
Modelo 2	3,449	59.34 %
Modelo 3	3,434	55.52 %
Modelo 1	3,504	54.22 %

Nota. Resumen de exactitud y criterio de información de Akaike (AIC) para cada modelo construido. Elaboración propia, realizado con Word.

En la tabla 38 se muestran los resultados de los criterios de información de Akaike (AIC) y la exactitud de predicción obtenida por matrices de confusión para cada modelo probado.

En dicha tabla, se determinó que los modelos 1,2,3,4 y 5 mostraron un ajuste deficiente a los datos reales comparado con el resto de los modelos Por esta razón, se descartaron dichos modelos para las secciones siguientes.

Por otro lado, los modelos siguientes presentaron resultados similares en su exactitud de predicción. Al analizar nuevamente la tabla 36, que muestra la elección de variables regresoras, fue posible identificar que los modelos 1, 2, 3, 4 y 5 no utilizaron el segmento de antigüedad de cotización como una de las variables regresoras, por lo que se debe destacar la importancia de esta variable como una predictiva de la retención o pérdida de los clientes.

Los modelos restantes mostraron valores similares en ambas mediciones, por lo que fue necesario tener una comparativa de modelos que tomara en cuenta ambas. La figura 17 muestra en su eje vertical los valores del criterio de información de Akaike (AIC), y en su eje horizontal el valor de la exactitud que se obtuvo con cada modelo.

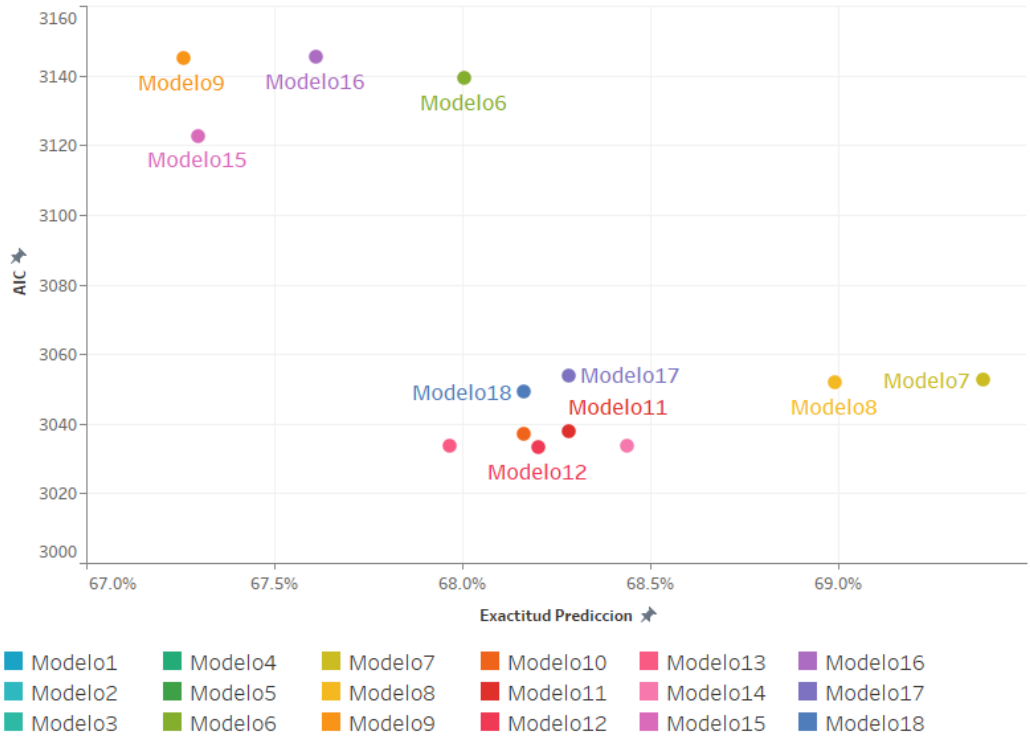
Como se puede observar, en el gráfico, al excluir los modelos de menor ajuste (modelos 1, 2, 3, 4 y 5), los resultados de los modelos restantes presentaron valores similares de exactitud (67.26 % a 69.39 %) y de criterio de información de Akaike (3033 a 3145).

Sin embargo, el modelo con mejor ajuste a los datos fue el modelo 7, pues tuvo una exactitud superior al resto de modelos, y mostró un criterio de

información de Akaike suficientemente bajo. Por ello, este fue el modelo elegido para el análisis de multicolinealidad.

Figura 17.

Comparativa de la exactitud de los modelos



Nota. Gráfico comparativo de exactitud y ajuste de modelos; en el eje C se muestra la exactitud de predicción, y en el eje Y el criterio de información de Akaike (AIC). Elaboración propia, realizado con Tableau.

3.3.4. Análisis de multicolinealidad

Se analizó la multicolinealidad del modelo elegido mediante el factor de inflación de la varianza (VIF). En la tabla 39 se muestran los valores de los grados

de libertad (Df) y del factor de inflación de la varianza (VIF) para las variables regresoras del modelo 7.

Tabla 39.

Análisis de multicolinealidad para el modelo 7

Variable	Grados de libertad (Df)	VIF
Cantidad de categorías	1	1.07
Método de envío	2	1.31
RM	8	3.88
Segmento de cotización	5	1.97
Segmento RFM	1	2.13
Sucursal	32	2.30
Término pago	3	1.29

Nota. Análisis de multicolinealidad para las variables del modelo 7, incluyendo los grados de libertad (Df) y el factor de inflación de la varianza (VIF). Elaboración propia, realizado con Word.

En este análisis se determinó que las dos variables que provocaron un efecto de inflación sobre la variabilidad del modelo en mayor medida fueron la variable RM percentil y la variable sucursal.

También es importante mencionar que los grados de libertad mostrados por la variable sucursal (32) incrementaron la variabilidad de los residuos, sin embargo, dado que el tamaño de la población es de 2,538 individuos, dicho efecto se consideró irrelevante.

Tal como se muestra en la tabla 40, se generaron tres modelos alternos al modelo 7: el primero eliminando la variable RM percentil, el segundo excluyó la variable sucursal, y el tercero se ajustó eliminando ambas variables.

Tabla 40.

Variables regresoras para modelos alternos al modelo 7

Modelo	Segmento	Metodo Envio	Cantidad Categorias	Sucursal	Segmento Cotización	Termino de Pago	Canal	F Percentil	M Percentil	R Percentil	RM	FM	RF
Modelo 7	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✓	✗	✗
Modelo 7A	✓	✓	✓	✓	✓	✓	✗	✗	✗	✗	✗	✗	✗
Modelo 7B	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✓	✗	✗
Modelo 7C	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗

Nota. Se muestran las diferentes combinaciones de variables utilizadas en la construcción de modelos de regresión logísticos alternos del modelo 7. Elaboración propia, realizado con Tableau.

Al iterar estos modelos y analizar su exactitud, se obtuvieron los resultados que se muestran en la figura 18.

En este gráfico, se pudo determinar que los modelos alternativos 7A y 7C, no mejoraron los resultados del modelo original, por lo que se dedujo que la inclusión de la variable de interacción RM fue clave para garantizar la exactitud y ajuste del modelo.

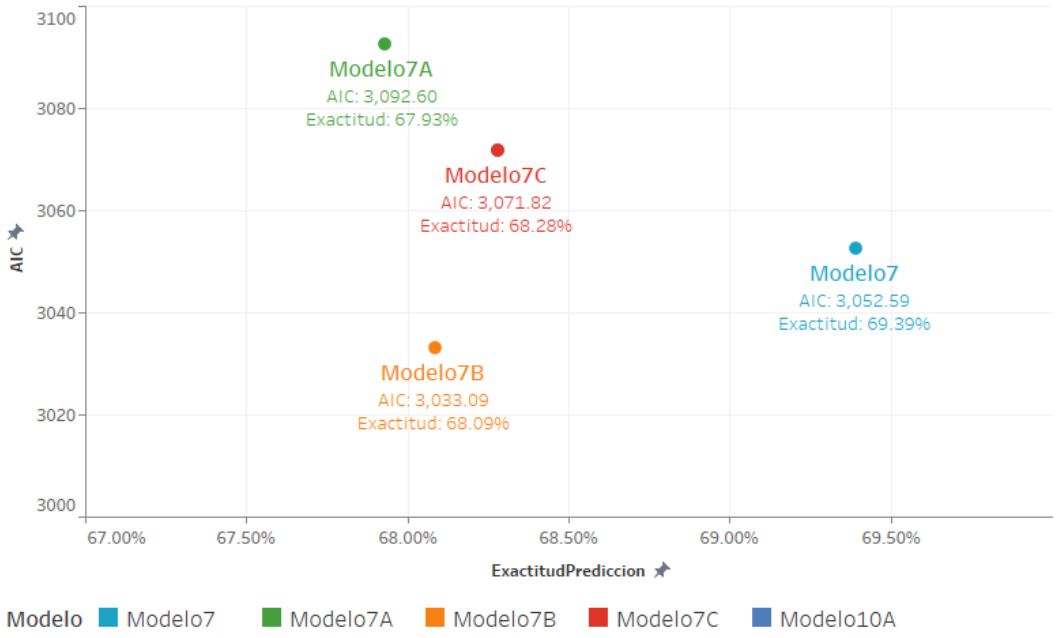
Por otro lado, el modelo 7B reflejó una mejora en el criterio de información de Akaike (AIC), sin embargo, esto fue a costa de una reducción en la exactitud general del modelo. Este resultado particular mostró que la inclusión de la variable sucursal incrementa levemente la exactitud, pero su cantidad de

tratamientos (que son introducidas en el modelo como valores *dummy*) incrementan su criterio de información de Akaike (AIC).

Por las razones anteriormente descritas, se determinó que el modelo de mejor ajuste para estimar la probabilidad de retención de clientes profesionales fue el modelo 7.

Figura 18.

Comparativa de la exactitud de modelos alternos al modelo 7



Nota. Gráfico comparativo de exactitud y ajuste de modelos; en el eje C se muestra la exactitud de predicción, y en el eje Y el criterio de información de Akaike (AIC). Elaboración propia, realizado con Tableau.

3.4. Objetivo general: construir un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales, en una empresa minorista de materiales de construcción en Guatemala

Los resultados del análisis de las variables reciencia, frecuencia y monto promedio de compras (incluyendo su agrupación percentil y por K medias), el análisis de independencia entre las variables regresoras y objetivo, así como la construcción iterativa de modelos de predicción permitieron identificar el modelo denominado como modelo 7, como el de mejor ajuste a los datos utilizados en el modelo de entrenamiento con una exactitud de 69.39 %.

El modelo de mejor ajuste está definido por la siguiente combinación de variables en su notación informática:

$$\begin{aligned} Variable\ Objetivo \sim\ & Segmento\ RFM \\ & + Metodo\ Envio \\ & + Cantidad\ Categorias \\ & + Sucursal \\ & + Termino\ Pago \\ & + Segmento\ Cotización \\ & + RM \end{aligned} \qquad \text{(Ec.33)}$$

La ecuación anterior fue utilizada para automatizar en tiempo real la predicción de la pérdida o retención de los clientes profesionales; generando así una base para el manejo de estrategias comerciales en la empresa minorista de materiales de construcción.

4. DISCUSIÓN DE RESULTADOS

El objetivo de la investigación fue construir un modelo de regresión logístico que permitiera estimar la probabilidad de retener o perder a los clientes, con el fin de implementar un sistema de seguimiento y servicio estratégico.

4.1. Análisis interno

La primera fase de la investigación tuvo como fin en el análisis descriptivo y agrupación de observaciones de las variables reciencia (R), frecuencia (F) y monto promedio de compra (M).

Tabla 41.

Coeficientes de variación y p-valores de normalidad de variables RFM

Variable	Coeficiente de variación		P valor de normalidad	
	Datos originales	Datos transformados	Datos originales	Datos transformados
Monto Promedio	129.10 %	34.00 %	0.000	0.211
Reciencia	156.90 %	14.10 %	0.000	0.001
Frecuencia	93.40 %	33.90 %	0.000	0.000

Nota. Coeficientes de variación y valores p de prueba de normalidad de variables reciencia, frecuencia y monto promedio de compras. Elaboración propia, realizado con Word.

Como se puede observar en la tabla 41, el análisis descriptivo de estas variables permitió identificar que las tres variables en su forma original mostraron altos coeficientes de variación (129.10 % para la frecuencia, 156.90 % para el monto promedio y 93.4 % para la reciencia) y un nulo ajuste a la distribución normal. Con ello, se concluyó que la transformación de estas variables de forma logarítmica fue adecuada, pues redujo la variabilidad (coeficientes de 34.00 % para la frecuencia, 14.10 % para el monto promedio y 33.90 % para la reciencia) aunque únicamente la frecuencia mostró un comportamiento normal al transformarse.

El uso de las variables transformadas en el algoritmo de agrupación por K medias mostró ser una forma eficiente de agrupar a los individuos del estudio, pues se obtuvieron tres segmentos heterogéneos en tamaño y comportamiento, como se hace referencia en la fase de presentación de resultados. A pesar de ello, también se obtuvo información valiosa sobre el comportamiento de los clientes utilizando el método de segmentación clásica de percentiles.

La finalidad del segundo objetivo de la investigación fue realizar un análisis de correlación entre las variables anteriormente descritas y otras variables de segmentación de los clientes. Dado que las variables fueron tratadas de forma categórica, se utilizó una prueba de independencia para cada variable. De dichas pruebas se destaca el resultado de las variables canal (p valor de 0.4301), condición de pago (p valor de 0.0276) y la agrupación de frecuencia en forma percentil (p valor de 0.0171).

Al analizar la falta de relación entre las variables condición de pago y canal con la variable de respuesta se detectó que la cantidad de observaciones en cada nivel de las variables regresoras provocó dicha independencia. Dicha restricción de negocio provocó que la variable canal fuera excluida del análisis y la variable

condición de pago no aportase mucha información en la construcción de modelos.

Por otro lado, la agrupación de la variable frecuencia en forma percentil demostró tener dependencia débil con la variable de respuesta. Sin embargo, dado que esta agrupación es una categorización de la variable frecuencia, cabe la posibilidad de buscar una forma diferente de categorizar dichas variables para buscar explicar más dicho fenómeno; esto no fue posible en el caso práctico mostrado, ya que es una definición actual del negocio.

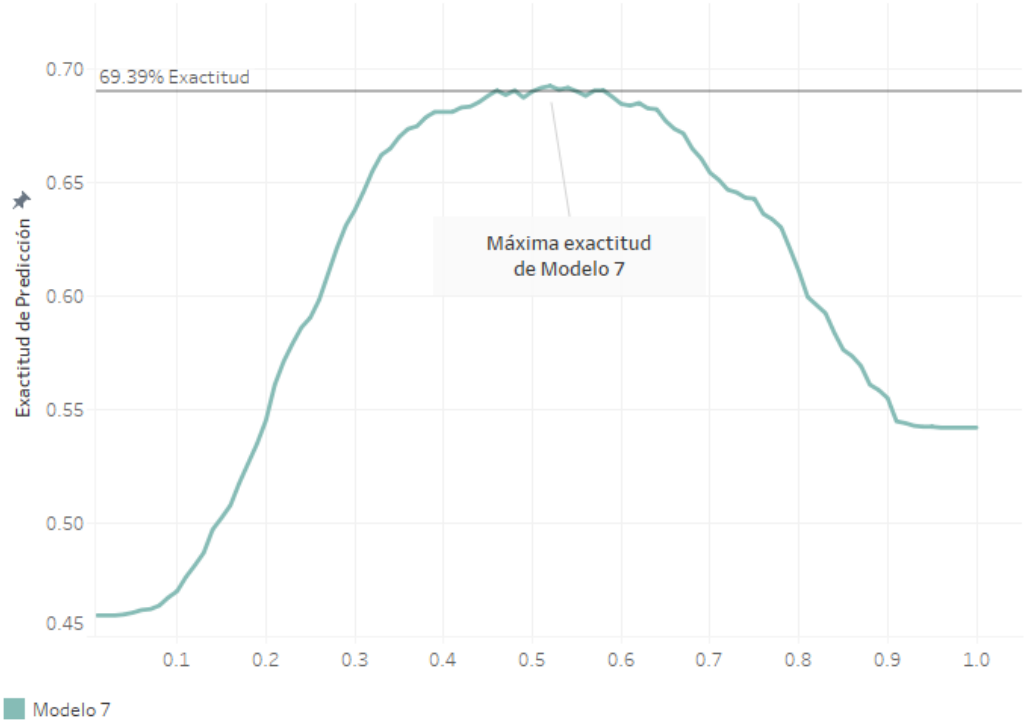
Por último, la construcción de modelos logísticos para predecir la compra o pérdida de los clientes permitió identificar la combinación de variables que explican el problema de estudio.

La medición y comparativa de dichos modelos se realizó principalmente a través de matrices de confusión; convirtiendo la variable de respuesta (probabilidad de retención del cliente) en una variable dicotómica. Para ello, si el modelo dio un resultado mayor a 0.5, dicha observación se categorizó como predicción de retención, por el contrario, se categorizó como una predicción de pérdida.

La generalización descrita fue revisada para descartar que el valor utilizado como referencia fuera un factor de inflación sobre la exactitud medida de los modelos. Como se muestra en la figura 19, la exactitud de predicción es similar al utilizar valores de referencia entre 0.45 y 0.55, por lo que el valor de 0.5 es un valor insesgado para determinar la exactitud de predicción.

Figura 19.

Exactitud del modelo 7 con diferentes valores de referencia



Nota. Gráfico de exactitud de precisión al utilizar diferentes valores de referencia para convertir los predichos en variables dicotómicas; en el eje X se muestran los valores de referencia, y en el eje Y se muestra la exactitud de predicción. Elaboración propia, realizado con Tableau.

Cabe mencionar que la extracción de resultados tomó una definición importante de negocio, pues se asumió que el ciclo de vida de los clientes no debe ser mayor a seis meses en el caso de los clientes profesionales. Una manera de mejorar la exactitud de predicción de los resultados sería profundizar sobre dicha suposición para encontrar estadísticamente, cuál es el ciclo de vida promedio de los clientes profesionales.

En la aplicación de los modelos de predicción, se debe diferenciar el tratamiento de los errores, pues un falso negativo (predicción de pérdida de un cliente que se retuvo) implica la aplicación de técnicas de servicio al cliente más personalizadas, por lo cual no representa un riesgo para la empresa. Por el contrario, los falsos positivos (predicción de retención de un cliente que se perdió) sí generan riesgo para la empresa.

En la tabla 42 se muestra nuevamente la matriz de confusión del modelo elegido.

Tabla 42.

Matriz de confusión para modelo de mejor ajuste

		Valores reales		Total
		Perdido	Retenido	
Valores predichos	Perdido	1 058	463	1 521
	Retenido	318	699	1 017
Total		1 376	1 162	2 538

Nota. Matriz de confusión del modelo de mejor ajuste; comparando los valores reales de retención y los valores predichos del modelo. Elaboración propia, realizado con Word.

Al analizar nuevamente los resultados de dicho modelo, se observa un total de 318 falsos positivos, lo que representa un riesgo de 12.53 %.

4.2. Análisis externo

El estudio realizado por Yoseph y AlMalaily (2019) propone que la mejor manera de segmentar a los clientes según su comportamiento es utilizando las

variables de frecuencia y monto promedio de compras, excluyendo el efecto de la reciencia. Sin embargo, los resultados de la presente investigación demuestran que la variable de frecuencia está menos vinculada con la retención o pérdida de clientes, y que la inclusión de la reciencia como una variable explicativa ha arrojado mejores resultados en la predicción de la retención. A pesar de esta discrepancia, es necesario analizar el contexto empresarial aplicado en el estudio de los autores mencionados antes de generalizar estos resultados a otros entornos.

Por otro lado, Cuadros, Gonzales y Jiménez (2017) establecieron en su investigación que la segmentación de clientes basada en su comportamiento es más detallada cuando se incluyen variables adicionales que describen su comportamiento. Esta afirmación se confirmó en el presente estudio, el cual demostró que la inclusión de la variable antigüedad de última cotización incrementó la precisión de los modelos de predicción. Esto también se puede interpretar como una segmentación más detallada de los clientes utilizando estas variables.

Es importante mencionar que en la investigación mencionada anteriormente se seleccionaron variables adicionales que se consideraron ajenas al estudio, como el margen promedio. Sería pertinente realizar un análisis de la influencia de esta variable utilizando la perspectiva del análisis multivariado para determinar si estas variables son útiles para segmentar a los clientes en el contexto de la investigación actual.

Jain, Khunteta & Srivastava (2020) mencionan en los resultados de su estudio una exactitud de predicción de 85.23 % para su mejor modelo de regresión logística. En contraste, la mejor iteración de modelos de la presente investigación presentó una exactitud máxima de 69.23 %. Dicha diferencia en la

exactitud presentada puede ser explicada desde dos perspectivas; la primera implica el análisis del mercado y de los servicios ofrecidos por cada escenario, y la segunda la revisión de las variables regresoras.

El caso de los autores previamente mencionados es una aplicación de dicho método en una empresa de telecomunicaciones, donde los clientes, por tener un esquema de contratación, suelen tener un comportamiento más predecible. Por ello, se puede inferir que el comportamiento de los clientes es más aleatorio que el caso de los autores.

Por otro lado, el estudio de Hargreaves (2019), que también fue aplicado al sector de telecomunicaciones, presentó una exactitud de predicción de 76.1 % utilizando una regresión logística binaria utilizando hasta 20 variables regresoras. A diferencia del autor, el presente estudio utilizó únicamente 7 variables regresoras para generar resultados similares (69.39 % de exactitud de predicción), manteniendo el principio de parsimonia estadística. La utilización de un número menor de variables predictoras en el presente estudio contribuye a la robustez del modelo, evitando un sobreajuste causado por la inclusión de variables no necesariamente correlacionadas con los datos.

CONCLUSIONES

1. El proceso de agrupación de los clientes profesionales en segmentos similares, utilizando el método de simulación por K Medias basado en las variables de reciencia, frecuencia y monto de compras transformadas, así como su segmentación clásica por percentiles, permitió obtener segmentos heterogéneos de clientes. Esto facilitó el análisis y la comprensión del comportamiento de los clientes, lo cual es fundamental para la implementación de estrategias de retención y personalización masiva.

2. A través de las pruebas de independencia, se identificaron las variables que influyen en la pérdida o retención de clientes. Se encontró una correlación significativa entre la reciencia, el monto promedio (tanto en su segmentación por K medias como en su agrupación clásica por percentiles) y la antigüedad de última cotización, mientras que variables como el canal, frecuencia y la condición de pago no mostraron una relación fuerte con la retención de clientes.

3. El modelo de regresión logística construido permitió cuantificar las probabilidades de pérdida y compra de cada cliente con una exactitud máxima de 69.39 %. Este modelo proporciona una herramienta valiosa para la toma de decisiones estratégicas, ya que permite identificar los factores clave que influyen en la retención de clientes y, por lo tanto, dirigir los esfuerzos y recursos hacia aquellos clientes con mayor probabilidad de permanecer en la empresa.

4. Se construyó el modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales en una empresa minorista de materiales de construcción en Guatemala. Este enfoque analítico ofrece a la empresa una base sólida para implementar un sistema de seguimiento y servicio estratégico, orientado a retener a los clientes de mayor valor y tomar decisiones informadas para maximizar la satisfacción y fidelidad del cliente.

RECOMENDACIONES

1. Utilizar la segmentación obtenida con el algoritmo de K medias para personalizar la comunicación con los clientes profesionales, pues automáticamente se obtienen tres segmentos suficientemente heterogéneos. A pesar de ello, se recomienda también continuar con la segmentación clásica de las variables RFM para algunas aplicaciones específicas.

2. Explorar otras variables y segmentaciones para comprender la percepción de la empresa en cada cliente. Además de los datos transaccionales utilizados en este estudio, sería importante considerar aspectos como el *Net Promoter Score* (NPS), encuestas u otros indicadores de satisfacción para mejorar la precisión de los modelos de predicción.

3. Profundizar en el análisis del ciclo de vida del cliente para identificar el momento en que un cliente se considera perdido. Esto permitirá mejorar en gran medida las suposiciones tomadas en el proceso de modelado de la retención de clientes profesionales.

4. Considerar la ampliación del modelo de regresión desarrollado en este estudio hacia otros segmentos de clientes para mejorar sus estrategias de servicio al cliente y fidelización. Este modelo proporcionará información valiosa para tomar decisiones informadas y adaptar las acciones de retención a los clientes más valiosos.

REFERENCIAS

Aleksandrova, Y. (2018). Application of machine learning for churn prediction based on transactional data (RFM analysis). [Aplicación de machine learning para predicción de churn basado en datos transaccionales (análisis RFM).] *SGEM International Multidisciplinary Scientific GeoConference EXPO Proceedings*. Sofia, Bulgaria. https://doi.org/10.5593/sgem2018/2.1/s07.016

Anitha, P. & Patil, M. (2022). RFM model for customer purchase behavior using K-Means algorithm [Modelo RFM para el comportamiento de compra del cliente utilizando el algoritmo K-Means]. *Journal of King Saud University - Computer and Information Sciences*, *34*(5), 1785-1792. https://doi.org/10.1016/j.jksuci.2019.12.011

Badii, M. H., Guillen, A., Lugo, O. P. & Aguilar, J. J. (2014). Correlación no-paramétrica y su aplicación en la investigaciones científica. *International Journal of Good Conscience*, *9*(2), 31-40. http://www.spentamexico.org/v9-n2/A5.9%282%2931-40.pdf

Bewick, V., Cheek, L. & Ball, J. (2003). Statistics review 7: Correlation and regression [Revisión estadística 7: Correlación y regresión.]. *Critical Care*, *7*(6), 451. https://ccforum.biomedcentral.com/articles/10.1186/cc2401

Boateng, E., & Abaye, D. (2019). A Review of the Logistic Regression Model with Emphasis on Medical Research [Una revisión del modelo de regresión

logística con énfasis en la investigación médica]. *Journal of Data Analysis and Information Processing*, *7*(4), 190–207. https://www.scirp.org/journal/paperinformation.aspx?paperid=95655

Castillo, S., y Damian, E. (2007). Q-Q Plot Normal. Los puntos de posición gráfica. *Iniciación a la Investigación*, *1*(2),1-20. https://revistaselectronicas.ujaen.es/index.php/ininv/issue/view/26

Cuadros, L., Gonzales, C., y Jiménez, P. (2017). Análisis multivariado para segmentación de clientes basada en RFM. *Tecnura, 21*(54), 41–51. https://doi.org/10.14483/22487638.12957

Del Castillo, R. y Salazar, R. (2018). *Fundamentos básicos de estadística*. Del Castillo Galarza, Raúl Santiago.

Dietrichson, A. (2019). *Métodos cuantitativos.* Bookdown.Org. https://bookdown.org/dietrichson/metodos-cuantitativos/test-de-normalidad.html

Dogan, O., Aycin, E., & Bulut, Z. (2018). Customer segmentation by using RFM model and clustering methods: a case study in retail industry [Segmentación de clientes mediante el modelo RFM y métodos de clustering: un estudio de caso en la industria minorista]. *International Journal of Contemporary Economics and Administrative Sciences, 8*(1), 1–19. https://doi.org/10.5930/issn.1925-4423

Flores, C., y Flores, K. (2021). Pruebas para comprobar la normalidad de datos en procesos productivos: Anderson - Darling, Ryan - Joiner, Shapiro -

Wilk y Kolmogórov - Smirnov. *Societas*, *23*(2), 83–106. https://matriculapre.up.ac.pa/index.php/societas/article/view/2302

Gajanova, L., Nadanyiova, M. & Moravcikova, D. (2019). The use of demographic and psychographic segmentation to creating marketing strategy of brand loyalty [El uso de la segmentación demográfica y psicográfica para crear una estrategia de marketing de lealtad a la marca]. *Scientific Annals of Economics and Business, 66*(1), 65-84. https://doi.org/10.2478/saeb-2019-0005

Hargreaves, C. A. (2019). A machine learning algorithm for churn reduction & revenue maximization: an application in the telecommunication industry [Un algoritmo de aprendizaje automático para reducir la deserción y maximizar los ingresos: una aplicación en la industria de las telecomunicaciones]. *International Journal of Future Computer and Communication,* *8*(4), 109–113. https://doi.org/10.18178/ijfcc.2019.8.4.550

Hernández, J., Espinosa, J., Penaloza, M., Díaz, E., Bautista, M., Riaño, M. & Chacón, O. (2018). Sobre el uso adecuado del coeficiente de correlación de Pearson: verificación de supuestos mediante un ejemplo aplicado a las ciencias de la salud. *Archivos Venezolanos de Farmacología y Terapéutica,* *37*(5), 552-570. https://www.redalyc.org/journal/559/55963207020/55963207020.pdf

Hernandez, J., Tello, E., Marin, H. & Romero, G. (2019). Aplicación de técnicas de aprendizaje no supervisado para la agrupación de trazas en el dominio de minería de procesos. *Pistas Educativas,* *41*(133), 356-374. http://pistaseducativas.celaya.tecnm.mx/index.php/pistas

Jain, H., Khunteta, A., & Srivastava, S. (2020). Churn prediction in telecommunication using logistic regression and logit boost [Predicción de abandono en telecomunicaciones mediante regresión logística y impulso logístico]. *Procedia Computer Science, 167*, 101–112. https://doi.org/10.1016/j.procs.2020.03.187

Lopez, P. & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Universitat Autònoma de Barcelona. https://ddd.uab.cat/pub/caplli/2015/131469/metinvsoccuan_cap3-6a2015.pdf

May, J., & Looney, S. (2022). On sample size determination when comparing two independent Spearman or Kendall coefficients [Sobre la determinación del tamaño de la muestra al comparar dos coeficientes independientes de Spearman o Kendall]. *Open Journal of Statistics*, *12*(02), 291-302. https://doi.org/10.4236/ojs.2022.122020

Möllering, L. (2018). *A customer segmentation sequence for B2B markets based on levels of market orientation of firms* [Una secuencia de segmentación de clientes para mercados B2B basada en los niveles de orientación al mercado de las empresas]. [Tesis de maestría, Universidad de Twente de Paises Bajos). Archivo digital. https://essay.utwente.nl/75340/

Murad, H. (2021). *Marketing Automation Customers Segmentation* [Tesis de maestría, Instituto de Tecnología de Rochester de Estados Unidos). Archivo digital. https://scholarworks.rit.edu/theses/11065/

Oshaki, M., Wang, P., Matsuda, K., Katagiri, S., Watanabe, H. & Ralescu, A. (2017). Confusion-Matrix-Based Kernel Logistic Regression for

Imbalanced Data Classification [Regresión logística del kernel basada en matriz de confusión para clasificación de datos desequilibrada]. *IEEE Transactions on Knowledge and Data Engineering*, *29*(9), 1806-1819. https://ieeexplore.ieee.org/document/7879291

Prykhodko, S., Prykhodko, N., Makarova, L., Kudin, O., Smykodub, T. & Prykhodko, A. (2017). Detecting bivariate outliers on the basis of normalizing transformations for non-Gaussian data [Detección de valores atípicos bivariados sobre la base de transformaciones de normalización para datos no gaussianos]. *The Vth international conference of Advanced Information Systems and Technologies, AIST 2017.* Sumy, Ucrania. http://essuir.sumdu.edu.ua/handle/123456789/55754

Rahim, M., Mushafiq, M., Khan, S. & Arain, Z. (julio, 2021). RFM-based repurchase behavior for customer classification and segmentation [Comportamiento de recompra basado en RFM para clasificación y segmentación de clientes]. *Journal of Retailing and Consumer Services, 61*, 102566. https://doi.org/10.1016/j.jretconser.2021.102566

Real Statistics Using Excel. (diciembre, 2022a). *Durbin-Watson*. https://real-statistics.com/statistics-tables/durbin-watson-table/

Real Statistics Using Excel. (diciembre, 2022b). *Kolmogorov-Smirnov Table*. https://real-statistics.com/statistics-tables/shapiro-wilk-table/

Real Statistics Using Excel. (diciembre, 2022c). *Shapiro–Wilk table*. https://real-statistics.com/statistics-tables/shapiro-wilk-table/

Sagaró, N. & Zamora, L. (2019). Análisis estadístico implicativo versus Regresión logística binaria para el estudio de la causalidad en salud. *Multimed, 23*(6), http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1028-48182019000601416

Sánchez, A., & Borges, Á. (2003). Transformación Z de Fisher para la determinación de intervalos de confianza del coeficiente de correlación de Pearson. *Psicothema*, *17*(1), 148-153. https://www.psicothema.com/pdf/3079.pdf

Senaviratna, N. y Cooray, T. (2019). Diagnosing Multicollinearity of Logistic Regression Model [Diagnóstico de multicolinealidad del modelo de regresión logística]. Asian *Journal of Probability and Statistics*, *5*(2), 1-9. https://www.researchgate.net/publication/336213669_Diagnosing_Multicollinearity_of_Logistic_Regression_Model

Serna, J. (2019). *Comparación de algunas estimaciones del τ de Kendall para datos bivariados con censura a intervalo.* [Tesis de maestría, Universidad Nacional de Colombia). Archivo digital. https://repositorio.unal.edu.co/handle/unal/76871?show=full

Ullah, I., Raza, B., Malik, A., Imran, M., Islam, S., & Kim, S. (2019). A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector [Un modelo de predicción de abandono que utiliza bosque aleatorio: análisis de técnicas de aprendizaje automático para la predicción de abandono e identificación de factores en el sector de las telecomunicaciones]. *IEEE Access, 7*, 60134–60149. https://doi.org/10.1109/access.2019.2914999

Van, M., Moons, K. G., de Groot, J. A., Collins, G. S., Altman, D. G., Eijkemans, M. J., & Reitsma, J. B. (2018). Sample size for binary logistic prediction models: Beyond events per variable criteria [Tamaño de muestra para modelos de predicción logística binaria: más allá de los eventos por criterio de variable]. *Statistical Methods in Medical Research, 28*(8), 2455–2474. https://doi.org/10.1177/0962280218784726

Yoseph, F., y AlMalaily, M. (2019). New market segmentation methods using enhanced (RFM), CLV, modified regression and clustering methods [Nuevos métodos de segmentación de mercado utilizando métodos mejorados (RFM), CLV, regresión modificada y clustering]. *International Journal of Computer Science and Information Technology, 11*(1), 43–60. https://aircconline.com/ijcsit/V11N1/11119ijcsit04.pdf

APÉNDICES

Apendice 1.

Coeficientes del modelo de regresión logística

Variable	Coeficiente
Intercepto	-12.4006607
Segmento RFM	-0.03421036
MetodoEnvio_Recoge.en.Tienda	-0.20986047
MetodoEnvio_Ruta	-0.75801984
CantidadCategorias_UniCategoria	-0.66705096
Sucursal_Sucursal2	13.34486380
Sucursal_Sucursal3	12.78457910
Sucursal_Sucursal4	12.82664720
Sucursal_Sucursal5	12.80871190
Sucursal_Sucursal6	0.51250090
Sucursal_Sucursal7	13.02195250
Sucursal_Sucursal8	13.25368060
Sucursal_Sucursal9	13.50033250
Sucursal_Sucursal10	12.99335920
Sucursal_Sucursal11	12.09747290
Sucursal_Sucursal12	13.42552610
Sucursal_Sucursal13	12.93918300
Sucursal_Sucursal14	11.60846710
Sucursal_Sucursal15	13.09915450
Sucursal_Sucursal16	12.76540590
Sucursal_Sucursal17	12.66164650

Continuación del apéndice 1.

Variable	Coeficiente
Sucursal_Sucursal18	13.33734110
Sucursal_Sucursal19	12.75886730
Sucursal_Sucursal20	14.31762380
Sucursal_Sucursal21	13.55060900
Sucursal_Sucursal22	12.74760410
Sucursal_Sucursal23	12.10161660
Sucursal_Sucursal24	12.91118860
Sucursal_Sucursal25	13.11781960
Sucursal_Sucursal26	13.35430340
Sucursal_Sucursal27	13.24866010
Sucursal_Sucursal28	13.20741510
Sucursal_Sucursal29	12.81877740
Sucursal_Sucursal30	12.52272260
Sucursal_Sucursal31	13.25657630
Sucursal_Sucursal32	13.18125200
Sucursal_Sucursal33	12.70187870
TerminoPago_30.Dias.Credito	0.00781756
TerminoPago_8.Dias.Credito	-0.91159605
TerminoPago_Contado	-0.35932838
SegmentoCoti_A2	0.55448379
SegmentoCoti_B1	0.94408252
SegmentoCoti_B2	1.61080135
SegmentoCoti_C1	2.47547231
SegmentoCoti_C2	1.09342065
RM_R1.M2	-0.16884229
RM_R1.M3	-0.76550399

Continuación del apéndice 1.

Variable	Coeficiente
RM_R2.M1	-0.58130817
RM_R2.M2	-1.12470589
RM_R2.M3	-1.11488995
RM_R3.M1	-1.64752881
RM_R3.M2	-1.66508561
RM_R3.M3	-1.68059856
RM_R3.M3	-1.68059856

Nota. Coeficientes del modelo de regresión de mejor ajuste. Elaboración propia, realizado con Microsoft Word y Rstudio.

Apendice 2.

Matriz de coherencia

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA			
Tipo de objetivo	Problema de investigación	Preguntas de investigación	Objetivos
Objetivo general	No hay información confiable sobre el comportamiento de los clientes profesionales	¿Cómo se comporta la probabilidad de retención de los clientes profesionales en función de sus variables cuantitativas y cualitativas?	Construir un modelo de regresión logística para estimar la probabilidad de retención de clientes profesionales en una empresa minorista de materiales de construcción en Guatemala
Objetivos específicos	No se conoce cuáles son los grupos óptimos de las variables de RFM en clientes profesionales	¿Cuáles son las agrupaciones óptimas de los clientes en función de sus variables RFM?	Agrupar a los clientes profesionales en segmentos similares basado en las variables de reciencia, frecuencia y monto de compras aplicando métodos de simulación por K Medias
	Se desconoce qué segmentos de mercado tienen una frecuencia de compra diferente	¿Qué variables de los clientes profesionales provocan una variación significativa en la retención o pérdida de los clientes?	Identificar las variables de clientes que sí infieren en la pérdida o retención de clientes usando pruebas de independencia y pruebas de correlación.
	No se conoce la probabilidad de compra de los clientes profesionales	¿Cuál es el modelo óptimo para describir la probabilidad de retención o pérdida de los clientes profesionales?	Relacionar las variables cuantitativas y cualitativas de los clientes profesionales construyendo un modelo de regresión logística que permita cuantificar las probabilidades de pérdida y compra de cada cliente profesional

Continuación del apéndice 2.

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA		
Tipo de objetivo	Procedimiento y técnicas	Resultados
Objetivo general	Extracción de datos de clientes para el análisis del comportamiento desde sus variables de RFM.	Se automatizó la extracción de datos para obtener las variables RFM en conjunto con otras variables de segmentación, así como la construcción del modelo de regresión para predecir la retención de clientes con una exactitud de 69.39 %.
Objetivos específicos	Aplicar algoritmo de simulación por K Medias y método de agrupación por percentiles a variables RFM.	Se segmentó el espacio de clientes usando variables RFM y el método clásico por percentiles, obteniendo grupos heterogéneos de las variables RFM.
	Pruebas de correlación e independencia de datos.	Se aplicaron pruebas de independencia a variables de segmentación obteniendo las variables que están relacionadas con la retención de clientes.
	Aplicación de regresión logística bivariada, criterios de información de modelos, matrices de confusión.	Se iteró para construir modelos de regresión logística en los clientes profesionales, midiendo su exactitud por medio de matrices de confusión.

Continuación del apéndice 2.

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA	
Tipo de objetivo	Conclusiones
Objetivo general	El modelo de regresión logística desarrollado en esta investigación proporciona una herramienta efectiva para estimar la probabilidad de retención de clientes profesionales en una empresa minorista de materiales de construcción en Guatemala. Este enfoque analítico ofrece a la empresa una base sólida para implementar un sistema de seguimiento y servicio estratégico, orientado a retener a los clientes de mayor valor y tomar decisiones informadas para maximizar la satisfacción y fidelidad del cliente
Objetivos específicos	El proceso de agrupación de los clientes profesionales en segmentos similares, utilizando el método de simulación por K Medias basado en las variables de reciencia, frecuencia y monto de compras transformadas, así como su segmentación clásica por percentiles, permitió obtener segmentos heterogéneos de clientes. Esto facilitó el análisis y la comprensión del comportamiento de los clientes, lo cual es fundamental para la implementación de estrategias de retención y personalización masiva.
	A través de las pruebas de independencia, se identificaron las variables que influyen en la pérdida o retención de clientes. Se encontró una correlación significativa entre la reciencia, el monto promedio (tanto en su segmentación por K medias como en su agrupación clásica por percentiles) y la antigüedad de última cotización, mientras que variables como el canal, frecuencia y la condición de pago no mostraron una relación fuerte con la retención de clientes
	El modelo de regresión logística construido, que relaciona variables cuantitativas y cualitativas de los clientes profesionales, permitió cuantificar las probabilidades de pérdida y compra de cada cliente con una exactitud máxima de 69.39 %. Este modelo proporciona una herramienta valiosa para la toma de decisiones estratégicas, ya que permite identificar los factores clave que influyen en la retención de clientes y, por lo tanto, dirigir los esfuerzos y recursos hacia aquellos clientes con mayor probabilidad de permanecer en la empresa

Continuación del apéndice 2

CONSTRUCCIÓN DE UN MODELO DE REGRESIÓN LOGÍSTICA PARA ESTIMAR LA PROBABILIDAD DE RETENCIÓN DE CLIENTES PROFESIONALES, EN UNA EMPRESA MINORISTA DE MATERIALES DE CONSTRUCCIÓN DE GUATEMALA	
Tipo de objetivo	Recomendaciones
Objetivo general	A la empresa minorista de materiales de construcción, considerar la ampliación del modelo de regresión desarrollado en este estudio hacia otros segmentos de clientes para mejorar sus estrategias de servicio al cliente y fidelización. Este modelo proporcionará información valiosa para tomar decisiones informadas y adaptar las acciones de retención a los clientes más valiosos.
Objetivos específicos	En el contexto de la medición del comportamiento de los clientes profesionales, se recomienda a la empresa utilizar la segmentación obtenida con el algoritmo de K medias para personalizar la comunicación con los clientes profesionales, pues automáticamente se obtienen tres segmentos suficientemente heterogéneos. A pesar de ello, se recomienda también continuar con la segmentación clásica de las variables RFM para algunas aplicaciones específicas.
	Sería beneficioso que la empresa minorista de materiales de construcción explore otras variables y segmentaciones para comprender la percepción de la empresa en cada cliente. Además de los datos transaccionales utilizados en este estudio, sería importante considerar aspectos como el Net Promoter Score (NPS), encuestas u otros indicadores de satisfacción para mejorar la precisión de los modelos de predicción
	Se sugiere a la empresa minorista de materiales de construcción profundizar en el análisis del ciclo de vida del cliente para identificar el momento en que un cliente se considera perdido. Esto permitirá implementar estrategias más efectivas para recuperar clientes o evitar que lleguen a ese punto crítico.

Nota. Matriz de coherencia del proceso de investigación. Elaboración propia.

Apendice 3.

Tabla de distribución normal estándar

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
-3.00	0.0013	0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018
-2.90	0.0019	0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025
-2.80	0.0026	0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034
-2.70	0.0035	0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045
-2.60	0.0047	0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060
-2.50	0.0062	0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080
-2.40	0.0082	0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104
-2.30	0.0107	0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136
-2.20	0.0139	0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174
-2.10	0.0179	0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222
-2.00	0.0228	0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281
-1.90	0.0287	0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351
-1.80	0.0359	0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436
-1.70	0.0446	0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537
-1.60	0.0548	0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655
-1.50	0.0668	0.0681	0.0694	0.0708	0.0721	0.0735	0.0749	0.0764	0.0778	0.0793
-1.40	0.0808	0.0823	0.0838	0.0853	0.0869	0.0885	0.0901	0.0918	0.0934	0.0951
-1.30	0.0968	0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131
-1.20	0.1151	0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335
-1.10	0.1357	0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562
-1.00	0.1587	0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814
-0.90	0.1841	0.1867	0.1894	0.1922	0.1949	0.1977	0.2005	0.2033	0.2061	0.2090
-0.80	0.2119	0.2148	0.2177	0.2206	0.2236	0.2266	0.2296	0.2327	0.2358	0.2389
-0.70	0.2420	0.2451	0.2483	0.2514	0.2546	0.2578	0.2611	0.2643	0.2676	0.2709
-0.60	0.2743	0.2776	0.2810	0.2843	0.2877	0.2912	0.2946	0.2981	0.3015	0.3050
-0.50	0.3085	0.3121	0.3156	0.3192	0.3228	0.3264	0.3300	0.3336	0.3372	0.3409
-0.40	0.3446	0.3483	0.3520	0.3557	0.3594	0.3632	0.3669	0.3707	0.3745	0.3783
-0.30	0.3821	0.3859	0.3897	0.3936	0.3974	0.4013	0.4052	0.4090	0.4129	0.4168
-0.20	0.4207	0.4247	0.4286	0.4325	0.4364	0.4404	0.4443	0.4483	0.4522	0.4562
-0.10	0.4602	0.4641	0.4681	0.4721	0.4761	0.4801	0.4840	0.4880	0.4920	0.4960

Continuación del apéndice 3.

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.00	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.10	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.20	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.30	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.40	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.50	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.60	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.70	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.80	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.90	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.00	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.10	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.20	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.30	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.40	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.50	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.60	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.70	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.80	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.90	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.00	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.10	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.20	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.30	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.40	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.50	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.60	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.70	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.80	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.90	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

Nota. Tabla de distribución normal estándar. Elaboración propia.

ANEXOS

Anexo 1.

Tabla de valores críticos para test de Kolmogorov-Smirnov

			α		
n	**0.01**	**0.05**	**0.1**	**0.15**	**0.2**
1	0.995	0.975	0.950	0.925	0.900
2	0.929	0.842	0.776	0.726	0.684
3	0.828	0.708	0.642	0.597	0.565
4	0.733	0.624	0.564	0.525	0.494
5	0.669	0.565	0.510	0.474	0.446
6	0.618	0.521	0.470	0.436	0.410
7	0.577	0.486	0.438	0.405	0.381
8	0.543	0.457	0.411	0.381	0.358
9	0.514	0.432	0.388	0.360	0.339
10	0.490	0.410	0.368	0.342	0.322
11	0.468	0.391	0.352	0.326	0.307
12	0.450	0.375	0.338	0.313	0.295
13	0.433	0.361	0.325	0.302	0.284
14	0.418	0.349	0.314	0.292	0.274
15	0.404	0.338	0.304	0.283	0.266
16	0.392	0.328	0.295	0.274	0.258
17	0.381	0.318	0.286	0.266	0.250
18	0.371	0.309	0.278	0.259	0.244
19	0.363	0.301	0.272	0.252	0.237
20	0.356	0.294	0.264	0.246	0.231
25	0.320	0.270	0.240	0.220	0.210
30	0.290	0.240	0.220	0.200	0.190
35	0.270	0.230	0.210	0.190	0.180
40	0.250	0.210	0.190	0.180	0.170

Continuación del anexo 1.

			α		
n	**0.01**	**0.05**	**0.1**	**0.15**	**0.2**
45	0.240	0.200	0.180	0.170	0.160
50	0.230	0.190	0.170	0.160	0.150

Nota. Tabla de valores de estadístico de Kolmogorov – Smirnov con diferentes niveles de significancia (α). Obtenido de Real Statistics Using Excel (2022b). *Kolmogorov-Smirnov Table* [Tabla Kolmogorov-Smirnov] (https://real-statistics.com/statistics-tables/kolmogorov-smirnov-table/), consultado el 14 de agosto de 2023. De dominio público.

Anexo 2.

Coeficientes a_i para pruebas de normalidad n≤30

n =	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
a1	0.7071	0.6872	0.6646	0.6431	0.6233	0.6052	0.5888	0.5739	0.5601	0.5475	0.5359	0.5251
a2	-	0.1677	0.2413	0.2806	0.3031	0.3164	0.3244	0.3291	0.3315	0.3325	0.3325	0.3318
a3	-	-	-	0.0875	0.1401	0.1743	0.1976	0.2141	0.2260	0.2347	0.2412	0.2460
a4	-	-	-	-	-	0.0561	0.0947	0.1224	0.1429	0.1586	0.1707	0.1802
a5	-	-	-	-	-	-	-	0.0399	0.0695	0.0922	0.1099	0.1240
a6	-	-	-	-	-	-	-	-	-	0.0303	0.0539	0.0727
a7	-	-	-	-	-	-	-	-	-	-	-	0.0240

n =	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**
a1	0.5150	0.5056	0.4968	0.4886	0.4808	0.4734	0.4643	0.4590	0.4542	0.4493	0.4450	0.4407
a2	0.3306	0.3290	0.3273	0.3253	0.3232	0.3211	0.3185	0.3156	0.3126	0.3098	0.3069	0.3043
a3	0.2495	0.2521	0.2540	0.2553	0.2561	0.2565	0.2578	0.2571	0.2563	0.2554	0.2543	0.2533
a4	0.1878	0.1939	0.1988	0.2027	0.2059	0.2085	0.2119	0.2131	0.2139	0.2145	0.2148	0.2151
a5	0.1353	0.1447	0.1524	0.1587	0.1641	0.1686	0.1736	0.1764	0.1787	0.1807	0.1822	0.1836
a6	0.0880	0.1005	0.1109	0.1197	0.1271	0.1334	0.1399	0.1443	0.1480	0.1512	0.1539	0.1563
a7	0.0433	0.0593	0.0725	0.0837	0.0932	0.1013	0.1092	0.1150	0.1201	0.1245	0.1283	0.1316
a8	-	0.0196	0.0359	0.0496	0.0612	0.0711	0.0804	0.0878	0.0941	0.0997	0.1046	0.1089
a9	-	-	-	0.0163	0.0303	0.0422	0.0530	0.0618	0.0696	0.0764	0.0823	0.0876
a10	-	-	-	-	-	0.0140	0.0263	0.0368	0.0459	0.0539	0.0610	0.0672
a11	-	-	-	-	-	-	-	0.0122	0.0228	0.0321	0.0403	0.0476
a12	-	-	-	-	-	-	-	-	0.0000	0.0107	0.0200	0.0284

Continuación del anexo 2.

n =	**27**	**28**	**29**	**30**	**31**	**32**	**33**	**34**	**35**	**36**	**37**	**38**
a1	0.4366	0.4328	0.4291	0.4254	0.4220	0.4188	0.4156	0.4127	0.4096	0.4068	0.4040	0.4015
a2	0.3018	0.2992	0.2968	0.2944	0.2921	0.2898	0.2876	0.2854	0.2834	0.2813	0.2794	0.2774
a3	0.2522	0.2510	0.2499	0.2487	0.2475	0.2463	0.2451	0.2439	0.2427	0.2415	0.2403	0.2391
a4	0.2152	0.2151	0.2150	0.2148	0.2145	0.2141	0.2137	0.2132	0.2127	0.2121	0.2116	0.2110
a5	0.1848	0.1857	0.1864	0.1870	0.1874	0.1878	0.1880	0.1882	0.1883	0.1883	0.1883	0.1881
a6	0.1584	0.1601	0.1616	0.1630	0.1641	0.1651	0.1660	0.1667	0.1673	0.1678	0.1683	0.1686
a7	0.1346	0.1372	0.1395	0.1415	0.1433	0.1449	0.1463	0.1475	0.1487	0.1496	0.1505	0.1513
a8	0.1128	0.1162	0.1192	0.1219	0.1243	0.1265	0.1284	0.1301	0.1317	0.1331	0.1344	0.1356
a9	0.0923	0.0965	0.1002	0.1036	0.1066	0.1093	0.1118	0.1140	0.1160	0.1179	0.1196	0.1211
a10	0.0728	0.0778	0.0822	0.0862	0.0899	0.0931	0.0961	0.0988	0.1013	0.1036	0.1056	0.1075
a11	0.0540	0.0598	0.0650	0.0697	0.0739	0.0777	0.0812	0.0844	0.0873	0.0900	0.0924	0.0947
a12	0.0358	0.0424	0.0483	0.0537	0.0585	0.0629	0.0669	0.0706	0.0739	0.0770	0.0798	0.0824
a13	0.0178	0.0253	0.0320	0.0381	0.0435	0.0485	0.0530	0.0572	0.0610	0.0645	0.0677	0.0706
a14	0.0000	0.0084	0.0159	0.0227	0.0289	0.0344	0.0395	0.0441	0.0484	0.0523	0.0559	0.0592
a15	-	-	0.0000	0.0076	0.0144	0.0206	0.0262	0.0314	0.0361	0.0404	0.0444	0.0481
a16	-	-	-	-	0.0000	0.0068	0.0131	0.0187	0.0239	0.0287	0.0331	0.0372
a17	-	-	-	-	-	-	0.0000	0.0062	0.0119	0.0172	0.0220	0.0264
a18	-	-	-	-	-	-	-	-	0.0000	0.0057	0.0110	0.0158
a19	-	-	-	-	-	-	-	-	-	-	0.0000	0.0053

Continuación del anexo 2.

n =	39	40	41	42	43	44	45	46	47	48	49	50
a1	0.3989	0.3964	0.3940	0.3917	0.3894	0.3872	0.3850	0.3830	0.3808	0.3789	0.3770	0.3751
a2	0.2755	0.2737	0.2719	0.2701	0.2684	0.2667	0.2651	0.2635	0.2620	0.2604	0.2589	0.2574
a3	0.2380	0.2368	0.2357	0.2345	0.2334	0.2323	0.2313	0.2302	0.2291	0.2281	0.2271	0.2260
a4	0.2104	0.2098	0.2091	0.2085	0.2078	0.2072	0.2065	0.2058	0.2052	0.2045	0.2038	0.2032
a5	0.1880	0.1878	0.1876	0.1874	0.1871	0.1868	0.1865	0.1862	0.1859	0.1855	0.1851	0.1847
a6	0.1689	0.1691	0.1693	0.1694	0.1695	0.1695	0.1695	0.1695	0.1695	0.1693	0.1692	0.1691
a7	0.1520	0.1526	0.1531	0.1535	0.1539	0.1542	0.1545	0.1548	0.1550	0.1551	0.1553	0.1554
a8	0.1366	0.1376	0.1384	0.1392	0.1398	0.1405	0.1410	0.1415	0.1420	0.1423	0.1427	0.1430
a9	0.1225	0.1237	0.1249	0.1259	0.1269	0.1278	0.1286	0.1293	0.1300	0.1306	0.1312	0.1317
a10	0.1092	0.1108	0.1123	0.1136	0.1149	0.1160	0.1170	0.1180	0.1189	0.1197	0.1205	0.1212
a11	0.0967	0.0986	0.1004	0.1020	0.1035	0.1049	0.1062	0.1073	0.1085	0.1095	0.1105	0.1113
a12	0.0848	0.0870	0.0891	0.0909	0.0927	0.0943	0.0959	0.0972	0.0986	0.0998	0.1010	0.1020
a13	0.0733	0.0759	0.0782	0.0804	0.0824	0.0842	0.0860	0.0876	0.0892	0.0906	0.9190	0.0932
a14	0.0622	0.0651	0.0677	0.0701	0.0724	0.0745	0.0765	0.0783	0.0801	0.0817	0.0832	0.0846
a15	0.0515	0.0546	0.0575	0.0602	0.0628	0.0651	0.0673	0.0694	0.0713	0.0731	0.0748	0.0764
a16	0.0409	0.0444	0.0476	0.0506	0.0534	0.0560	0.0584	0.0607	0.0628	0.0648	0.0667	0.0685
a17	0.0305	0.0343	0.0379	0.0411	0.0442	0.0471	0.0497	0.0522	0.0546	0.0568	0.0588	0.0608
a18	0.0203	0.0244	0.0283	0.0318	0.0352	0.0383	0.0412	0.0439	0.0465	0.0489	0.0511	0.0532
a19	0.0101	0.0146	0.0188	0.0227	0.0263	0.0296	0.0328	0.0357	0.0385	0.0411	0.0436	0.0459
a20	0.0000	0.0049	0.0094	0.0136	0.0175	0.0211	0.0245	0.0277	0.0307	0.0335	0.0361	0.0386
a21	-	-	0.0000	0.0045	0.0087	0.0126	0.0163	0.0197	0.0229	0.0259	0.0288	0.0314
a22	-	-	-	-	0.0000	0.0042	0.0081	0.0118	0.0153	0.0185	0.0215	0.0244
a23	-	-	-	-	-	-	0.0000	0.0039	0.0076	0.0111	0.0143	0.0174
a24	-	-	-	-	-	-	-	-	0.0000	0.0037	0.0071	0.0104

Nota. Tabla de coeficientes a_i para pruebas de normalidad Shapiro-Wilk. Obtenido de Real Statistics Using Excel (2022c). *Shapiro-Wilk Table* [Tabla Shapiro-Wilk] (https://real-statistics.com/statistics-tables/shapiro-wilk-table/), consultado el 14 de agosto de 2023. De dominio público.

Anexo 3.

Valores p para prueba de Shapiro Wilk

					p valor				
n	**0.01**	**0.02**	**0.05**	**0.1**	**0.5**	**0.9**	**0.95**	**0.98**	**0.99**
3	0.753	0.756	0.767	0.789	0.959	0.998	0.999	1.000	1.000
4	0.687	0.707	0.748	0.792	0.935	0.987	0.992	0.996	0.997
5	0.686	0.715	0.762	0.806	0.927	0.979	0.986	0.991	0.993
6	0.713	0.743	0.788	0.826	0.927	0.974	0.981	0.986	0.989
7	0.730	0.760	0.803	0.838	0.928	0.972	0.979	0.985	0.988
8	0.749	0.778	0.818	0.851	0.932	0.972	0.978	0.984	0.987
9	0.764	0.791	0.829	0.859	0.935	0.972	0.978	0.984	0.986
10	0.781	0.806	0.842	0.869	0.938	0.972	0.978	0.983	0.986
11	0.792	0.817	0.850	0.876	0.940	0.973	0.979	0.984	0.986
12	0.805	0.828	0.859	0.883	0.943	0.973	0.979	0.984	0.986
13	0.814	0.837	0.866	0.889	0.945	0.974	0.979	0.984	0.986
14	0.825	0.846	0.874	0.895	0.947	0.975	0.980	0.984	0.986
15	0.835	0.855	0.881	0.901	0.950	0.975	0.980	0.984	0.987
16	0.844	0.863	0.887	0.906	0.952	0.976	0.981	0.985	0.987
17	0.851	0.869	0.892	0.910	0.954	0.977	0.981	0.985	0.987
18	0.858	0.874	0.897	0.914	0.956	0.978	0.982	0.986	0.988
19	0.863	0.879	0.901	0.917	0.957	0.978	0.982	0.986	0.988
20	0.868	0.884	0.905	0.920	0.959	0.979	0.983	0.986	0.988
21	0.873	0.888	0.908	0.923	0.960	0.980	0.983	0.987	0.989
22	0.878	0.892	0.911	0.926	0.961	0.980	0.984	0.987	0.989
23	0.881	0.895	0.914	0.928	0.962	0.981	0.984	0.987	0.989
24	0.884	0.898	0.916	0.930	0.963	0.981	0.984	0.987	0.989
25	0.888	0.901	0.918	0.931	0.964	0.981	0.985	0.988	0.989
26	0.891	0.904	0.920	0.933	0.965	0.982	0.985	0.988	0.989
27	0.894	0.906	0.923	0.935	0.965	0.982	0.985	0.988	0.990
28	0.896	0.908	0.924	0.936	0.966	0.982	0.985	0.988	0.990
29	0.898	0.910	0.926	0.937	0.966	0.982	0.985	0.988	0.990
30	0.900	0.912	0.927	0.939	0.967	0.983	0.985	0.988	0.990
31	0.902	0.914	0.929	0.940	0.967	0.983	0.986	0.988	0.990
32	0.904	0.915	0.930	0.941	0.968	0.983	0.986	0.988	0.990
33	0.906	0.917	0.931	0.942	0.968	0.983	0.986	0.989	0.990
34	0.908	0.919	0.933	0.943	0.969	0.983	0.986	0.989	0.990

Continuación del anexo 3.

	p valor								
n	**0.01**	**0.02**	**0.05**	**0.1**	**0.5**	**0.9**	**0.95**	**0.98**	**0.99**
35	0.910	0.920	0.934	0.944	0.969	0.984	0.986	0.989	0.990
36	0.912	0.922	0.935	0.945	0.970	0.984	0.986	0.989	0.990
37	0.914	0.924	0.936	0.946	0.970	0.984	0.987	0.989	0.990
38	0.916	0.925	0.938	0.947	0.971	0.984	0.987	0.989	0.990
39	0.917	0.927	0.939	0.948	0.971	0.984	0.987	0.989	0.991
40	0.919	0.928	0.940	0.949	0.972	0.985	0.987	0.989	0.991
41	0.920	0.929	0.941	0.950	0.972	0.985	0.987	0.989	0.991
42	0.922	0.930	0.942	0.951	0.972	0.985	0.987	0.989	0.991
43	0.923	0.932	0.943	0.951	0.973	0.985	0.987	0.990	0.991
44	0.924	0.933	0.944	0.952	0.973	0.985	0.987	0.990	0.991
45	0.926	0.934	0.945	0.953	0.973	0.985	0.988	0.990	0.991
46	0.927	0.935	0.945	0.953	0.974	0.985	0.988	0.990	0.991
47	0.928	0.936	0.946	0.954	0.974	0.985	0.988	0.990	0.991
48	0.929	0.937	0.947	0.954	0.974	0.985	0.988	0.990	0.991
49	0.929	0.939	0.947	0.955	0.974	0.985	0.988	0.990	0.991
50	0.930	0.938	0.947	0.955	0.974	0.985	0.988	0.990	0.991

Nota. Valores p para estadísticos de prueba Shapiro-Wilk. Obtenido de Real Statistics Using Excel (2022c). *Shapiro-Wilk Table* [Tabla Shapiro-Wilk] (https://real-statistics.com/statistics-tables/shapiro-wilk-table/), consultado el 14 de agosto de 2023. De dominio público.

Anexo 4.

Tabla de valores críticos Durbin Watson para α=0.01

n	k = 1		k = 2		k = 3		k = 4		k = 5		k = 6		k = 7	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0.390	1.142	-	-	-	-	-	-	-	-	-	-	-	-
7	0.435	1.036	0.294	1.676	-	-	-	-	-	-	-	-	-	-
8	0.497	1.003	0.345	1.489	0.229	2.102	-	-	-	-	-	-	-	-
9	0.554	0.998	0.408	1.389	0.279	1.875	0.183	2.433	-	-	-	-	-	-
10	0.604	1.001	0.466	1.333	0.340	1.733	0.230	2.193	0.150	2.690	-	-	-	-
11	0.653	1.010	0.519	1.297	0.396	1.640	0.286	2.030	0.193	2.453	0.124	2.892	-	-
12	0.697	1.023	0.569	1.274	0.449	1.575	0.339	1.913	0.244	2.280	0.164	2.665	0.105	3.053
13	0.738	1.038	0.616	1.261	0.499	1.526	0.391	1.826	0.294	2.150	0.211	2.490	0.140	2.838
14	0.776	1.054	0.660	1.254	0.547	1.490	0.441	1.757	0.343	2.049	0.257	2.354	0.183	2.667
15	0.811	1.070	0.700	1.252	0.591	1.465	0.487	1.705	0.390	1.967	0.303	2.244	0.226	2.530
16	0.844	1.086	0.738	1.253	0.633	1.447	0.532	1.664	0.437	1.901	0.349	2.153	0.269	2.416
17	0.873	1.102	0.773	1.255	0.672	1.432	0.574	1.631	0.481	1.847	0.393	2.078	0.313	2.319
18	0.902	1.118	0.805	1.259	0.708	1.422	0.614	1.604	0.522	1.803	0.435	2.015	0.355	2.238
19	0.928	1.133	0.835	1.264	0.742	1.416	0.650	1.583	0.561	1.767	0.476	1.963	0.396	2.169
20	0.952	1.147	0.862	1.270	0.774	1.410	0.684	1.567	0.598	1.736	0.515	1.918	0.436	2.110
21	0.975	1.161	0.889	1.276	0.803	1.408	0.718	1.554	0.634	1.712	0.552	1.881	0.474	2.059
22	0.997	1.174	0.915	1.284	0.832	1.407	0.748	1.543	0.666	1.691	0.587	1.849	0.510	2.015
23	1.017	1.186	0.938	1.290	0.858	1.407	0.777	1.535	0.699	1.674	0.620	1.821	0.545	1.977
24	1.037	1.199	0.959	1.298	0.881	1.407	0.805	1.527	0.728	1.659	0.652	1.797	0.578	1.944
25	1.055	1.210	0.981	1.305	0.906	1.408	0.832	1.521	0.756	1.645	0.682	1.776	0.610	1.915
26	1.072	1.222	1.000	1.311	0.928	1.410	0.855	1.517	0.782	1.635	0.711	1.759	0.640	1.889
27	1.088	1.232	1.019	1.318	0.948	1.413	0.878	1.514	0.808	1.625	0.738	1.743	0.669	1.867
28	1.104	1.244	1.036	1.325	0.969	1.414	0.901	1.512	0.832	1.618	0.764	1.729	0.696	1.847
29	1.119	1.254	1.053	1.332	0.988	1.418	0.921	1.511	0.855	1.611	0.788	1.718	0.723	1.830
30	1.134	1.264	1.070	1.339	1.006	1.421	0.941	1.510	0.877	1.606	0.812	1.707	0.748	1.814

Nota. Tabla de valores críticos Durbin Watson para α=0.01. Obtenido de Real Statistics Using Excel (2022a). *Durbin-Watson Table* [Tabla Durbin-Watson] (https://real-statistics.com/statistics-tables/durbin-watson-table/), consultado el 14 de agosto de 2023. De dominio público.