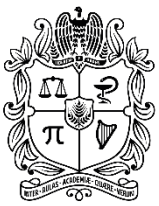

UNIVERSIDAD
NACIONAL
DE COLOMBIA

Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación

Davinson Mosquera González

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2022

Método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación

Davinson Mosquera González

Tesis de investigación presentada como requisito parcial para optar al título de:

Magíster en Ingeniería - Analítica

Director:

Ph.D. John Willian Branch Bedoya

Universidad Nacional de Colombia

Facultad de Minas, Departamento de Ciencias de la Computación y de la Decisión

Medellín, Colombia

2022

Dedicatoria

A Dios por permitirme culminar esta etapa de mi vida; a mi madre y abuela por sus oraciones y apoyo incondicional; a mi familia y amigos por su acompañamiento.

Resumen

La presente tesis de investigación, tiene como objetivo proponer un método para la segmentación de clientes, incorporando la predicción del valor monetario del cliente como una variable de segmentación, para tal fin, se propone una metodología cuantitativa, en la que los datos a utilizar corresponden a las transacciones de una tienda en línea de regalos para toda ocasión de Reino Unido, denominada “Online Retail II”, que consta de un total de 5.833 clientes y 1.067.371 registros; a partir de los cuales se realiza un proceso de caracterización de los datos, seguido de la predicción del valor monetario de cada cliente utilizando técnicas estadísticas y de aprendizaje de máquinas, que posteriormente se incluye como variable en el proceso de segmentación. Finalmente, se hace un comparativo entre los resultados de segmentar clientes sin incorporar la predicción del valor monetario y la segmentación de clientes incorporando la predicción del valor monetario; con lo que se concluye que el método propuesto, utilizando el algoritmo de Vecinos más cercanos para la predicción del valor monetario del cliente, al incorporarlo en la segmentación de clientes, logra un desempeño económico entre 10% y 20% mejor que segmentar sin incorporar esta variable.

Palabras clave: segmentación de clientes, valor monetario, aprendizaje de máquinas, modelos paramétricos

Abstract

Method for customer segmentation incorporating the prediction of the customer's monetary value as a segmentation variable

This research thesis aims to propose a method for customer segmentation, incorporating the prediction of the customer's monetary value as a segmentation variable, for this purpose, a quantitative methodology is proposed, in which the data to be used correspond to the transactions of an online all-occasion gifts store in the United Kingdom, called “Online Retail II”, consisting of a total of 5.833 customers and 1.067.371 registrations; from which a data characterization process is carried out, followed by the prediction of the monetary value of each client using statistical and machine learning techniques, which is later included as a variable in the segmentation process. Finally, a comparison is made between the results of segmenting customers without incorporating the prediction of the monetary value and the customer segmentation incorporating the prediction of the monetary value; with which it is concluded that the proposed method, using the Nearest Neighbors algorithm for the prediction of the monetary value of the client, when incorporating it into the client segmentation, achieves an economic performance between 10% and 20% better than segmenting without incorporating this variable.

Keywords: customer segmentation, customer lifetime value (CLV), machine learning, parametric models

Contenido

Lista de figuras

Pág.

Lista de tablas

Pág.

1. Introducción

En los últimos años, las empresas han experimentado un contexto cada vez más competitivo, donde solo aquellas empresas que aporten verdadero valor a sus clientes, con base en relaciones sólidas y rentables en el tiempo, sobrevivirán. En este sentido, las empresas están en una constante búsqueda para identificar, atraer y retener cliente (Tsai et al., 2019). Es así como, en la formulación de estrategias de relacionamiento con los clientes (CRM), se encuentra la segmentación de clientes. Esta consiste en el agrupamiento de los diferentes clientes de una organización, en subgrupos más pequeños que el total, donde se tiene como objetivo que clientes con características similares (sociodemográficas o transaccionales) se encuentren en el mismo grupo; y que a partir de allí se puedan hacer estrategias segmentadas o dirigidas que permitan la optimización de recursos (Chatterjee et al., 2021).

Durante más de un siglo, las empresas han creado, administrado y utilizado datos sobre los clientes para mejorar las relaciones con los clientes, optimizar la rentabilidad y segmentar clientes (Oblander et al., 2020). Tradicionalmente, una de las técnicas más utilizada para la segmentación de clientes es el modelo de Recencia, Frecuencia, Monto, del inglés *Recency, Frequency, Monetary* (RFM), en el que a partir de datos transaccionales de los clientes, los agrega en grupos con comportamientos de compra similares (Rathi & Ravi, 2017; Heldt et al., 2021). Si bien es una técnica ampliamente utilizada, presenta la limitación de estar basada en momentos puntuales del tiempo y de restringir su capacidad de segmentación al comportamiento pasado de los clientes, sin contemplar variables adicionales o pronósticos del futuro (Yoseph & Heikkila, 2019).

En consecuencia, diferentes autores han propuesto realizar estudios de segmentación de clientes utilizando la predicción del comportamiento futuro de los clientes, para ello se ha propuesto el Valor Monetario del Cliente o *Customer Lifetime Value (CLV)*; este puede ser definido como el valor monetario de la relación entre la organización y el cliente, basado en las contribuciones pasadas y valor presente del cliente, proyectados a una

relación futura con la empresa, es decir, es una predicción de las ganancia que un cliente llevaría a una organización, en su relación futura (Rathi & Ravi, 2017; Sifa et al., 2018; Channa, 2019).

Para cumplir los fines de predecir el Valor Monetario del Cliente, *Customer Lifetime Value (CLV),* para su uso en la segmentación de clientes, se ha optado por técnicas que se agrupan principalmente en dos enfoques: i) enfoque estadístico o paramétrico, en los cuales a partir de reglas de comportamiento de consumidor y supuestos estadísticos, se estiman parámetros que permiten la predicción del CLV (Jasek et al., 2019) y ii) enfoque de aprendizaje de máquinas, en los cuales se utilizan técnicas supervisadas, principalmente de regresión, para la predicción del CLV (Chen et al., 2019; Bauer & Jannach, 2021). En la Figura 1-1 se representa un esquema de los dos enfoques enunciados.

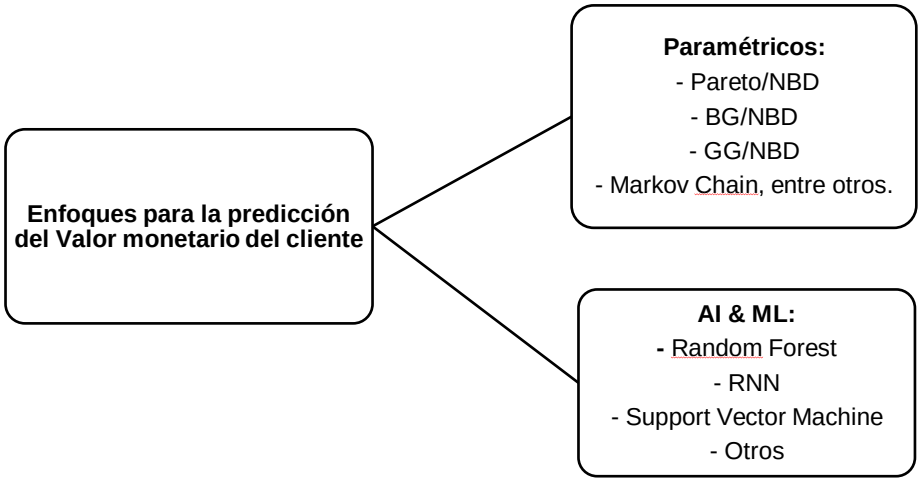

Figura 1-1: Enfoques para la predicción del valor monetario del cliente (CLV)
Fuente: elaboración propia

Es así como la presente tesis de maestría tiene como objetivo principal proponer un método para la segmentación de clientes, que incorpore la predicción del valor monetario del cliente como una variable de segmentación. Para lo que se plantea una estructura del documento dividida en 7 capítulos y finalmente las Conclusiones y recomendaciones: 1) Introducción; 2) Marco teórico; 3) Revisión sistemática de la literatura; 4) Caracterización de los datos; 5) Selección de métodos para la predicción del valor monetario; 6) Método para la segmentación de clientes incorporando la predicción del valor monetario; 7)

Validación de los resultados del método de segmentación propuesto; y las Conclusiones y recomendaciones.

1.1 Motivación

Esta tesis de investigación está motivada desde la importancia y pertinencia de la utilización de la analítica de datos en la segmentación de clientes, dado que esto contribuye en dos de los objetivos principales de las empresas, la rentabilidad y su sostenibilidad en el tiempo. En principio, en las dinámicas de competitividad empresarial actuales, se identificó que muchas empresas, en el contexto colombiano, implementan estrategias de identificación, atracción y retención de clientes, sin soporte profundo en los datos o utilizando reglas subjetivas. Así mismo, la identificación de empresas que se basan en el comportamiento pasado o técnicas tradicionales como el modelo RFM, para la segmentación de cliente, donde corren el riesgo de no lograr una correcta optimización del recurso o presupuesto de mercadeo.

Al ejemplificar datos sobre las consecuencias de incorrecta gestión de los clientes de las empresas, se tiene que, siguiendo el principio de Pareto, una pequeña parte de los clientes de las empresas contribuyen más a los ingresos de la empresa que el resto (Srivastava, 2017); una organización, del sector retail o minorista, pierde el 50% de sus clientes cada 5 años (Guadarrama & Rosales, 2015); y finalmente, la captación de un cliente nuevo requiere un esfuerzo cinco veces mayor que conservar uno existente (Alipour, 2016).

Es así, como se plantea la ventaja y potencial que tiene la generación de conocimiento a partir de los datos, para aportar a los objetivos de las empresas, de cara a poder soportar la toma de decisiones sobre las estrategias a implementar con sus clientes, a partir de los datos. En ese sentido, la importancia y pertinencia del uso de la analítica de datos para estos objetivos y contribuir con la optimización de recursos en las empresas, al invertir sus recursos de mercadeo en los clientes que agregan verdadero valor.

1.2 Trabajos previos

En la literatura, la segmentación de clientes con enfoque de una predicción de valor monetario del cliente ha sido desarrollada principalmente a partir del uso de métodos

paramétricos, lo que se le denomina métodos tradicionales, en este sentido, Jasek et al. (2019), en su artículo "*Comparative Analysis of Selected Probabilistic Customer Lifetime Value Models In Online Shopping",* realizan un amplio comparativo de métodos paramétricos para la predicción del valor monetario de los clientes en empresas tipo B2C (Negocio a cliente), y concluyen que el modelo BG/NBD y Pareto NBD logran los mejores resultados para la predicción del valor monetario.

Chen et al. (2019) plantean en su estudio "*Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models",* proponen el uso de Redes Neuronales Convolucionales para la predicción del valor monetario del cliente, así mismo en este artículo se realiza una comparativa con los modelos tradicionales más utilizados para este fin, donde concluyen que el modelo de aprendizaje profundo propuesto logra mejores resultados que el modelo Pareto/NBD.

Win & Bo (2020), en su investigación titulada "*Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm",* cambian el enfoque tradicional del valor monetario del cliente, en el cual pasan de un problema de regresión a un problema de clasificación, en el cual predicen la clase de los clientes para un año siguiente a partir del valor del CLV.

Finalmente, uno de los trabajos más recientes y relevantes es el presentado por Bauer & Jannach (2021), titulado "*Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-based Models",* en el cual proponen un método de Redes Neuronales Recurrentes para la predicción del valor monetario del cliente, donde obtienen como resultados que arrojan mejores resultados que los modelos paramétricos.

1.3 Problema de investigación

En la literatura formal, la mayoría de estudios se enfocan en la segmentación de clientes utilizando métodos tradicionales como los modelos Recencia, Frecuencia, Monto; o en el uso de técnicas tradicionales para la predicción del valor monetario del cliente; donde recientemente diferentes autores han propuesto nuevos métodos para la predicción del valor monetario del cliente, utilizando técnicas de aprendizaje de máquinas e inteligencia artificial; sin embargo, generalmente, estos estudios se quedan hasta la predicción del

valor del cliente, pero no se expresa el paso para incorporarlo en una segmentación final de clientes; así mismo, las técnicas tradicionales de predicción del valor monetario de clientes son costosas computacionalmente y limitan su uso a partir de premisas o supuestos estadísticos que no siempre se cumplen en los datos. En este sentido, se identifica la necesidad de indagar o proponer métodos de segmentación de clientes que tomen como insumos las predicciones realizadas del valor monetario del cliente y que sea más flexible en cuanto a supuestos del comportamiento original de los datos.

Con lo anterior, se propone como pregunta principal de investigación: ¿Cómo mejorar los resultados en la segmentación de clientes al incorporar la predicción del valor monetario del cliente como una variable de segmentación?

1.4 Objetivos

1.4.1 Objetivo general

Proponer un método para la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación.

1.4.2 Objetivos específicos

- Caracterizar los datos transaccionales históricos de los clientes mediante el uso de técnicas de limpieza, exploración y descripción de los datos

- Seleccionar métodos paramétricos y de aprendizaje de máquinas para la predicción del valor monetario del cliente.

- Diseñar un método para la segmentación de clientes que incorpore la predicción del valor monetario del cliente como una variable de segmentación.

- Validar los resultados en la segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable de segmentación.

1.5 Contribuciones

Con los resultados de la presente tesis de investigación se logran aportes pertinentes al campo de conocimiento relacionado con la segmentación de clientes y la predicción del valor monetario de los clientes, a continuación, se enuncian las principales contribuciones:

- La revisión sistemática de la literatura del capítulo 3, se realizó soportada en un análisis bibliométrico, el cual contribuye al soporte de futuras investigaciones, a partir de las temáticas emergentes identificadas, así mismo, contribuye al crecimiento del campo de conocimiento en Colombia, donde solo se cuenta con 3 publicaciones en revistas de alto impacto indexadas a Scopus.

- En la literatura revisada, no se encontraron estudios que presentaran de una manera abierta la forma en la que se realiza la comparativa entre técnicas estadísticas y técnicas de aprendizaje de máquinas, para evaluar el desempeño en la predicción del valor monetario del cliente, dado que son “caja negra”. Esta tesis, describe detalladamente un método para realizar este tipo de comparativas.

- En la revisión de literatura realizada, se evidenciaron pocos casos en los cuales posterior a realizar la predicción del valor monetario del cliente, lo incorporaran como insumo en un proceso de segmentación de clientes. Esta tesis de investigación demuestra cómo con el uso de esta variable en el proceso de segmentación se mejoran los resultados de este.

- El método propuesto de segmentación de clientes de la presente tesis demuestra que, con los datos utilizados, se valida la hipótesis principal del presente estudio, relacionada a que al realizar segmentación de clientes incorporando la variable de predicción del valor monetario del cliente, se contribuye a una mejor optimización de presupuesto en relación con el mercadeo dirigido a sus clientes más valiosos. Encontrando que representa la identificación de contribuciones económicas de un 12% por encima del uso de métodos tradicionales.

2. Marco teórico

2.1 Segmentación de clientes

La segmentación de clientes brinda una buena comprensión de la necesidad de los clientes y ayuda a identificar a los clientes potenciales de la empresa. La división de los clientes en segmentos también aumenta los ingresos de la empresa. Se cree que retener a los clientes es más importante que encontrar nuevos clientes (Christy et al., 2018). Así mismo, el proceso de segmentar a los clientes con comportamientos similares en el mismo segmento y con diferentes patrones en diferentes segmentos se denomina segmentación de clientes (Kansal et al., 2018).

2.2 Valor monetario del cliente

El término ampliamente utilizado en literatura es *Customer Lifetime Value (CLV)*; el cual es definido como el valor monetario de la relación entre la organización y el cliente, basado en las contribuciones pasadas y valor presente del cliente, proyectados a una relación futura con el mismo, es decir, es una predicción de las ganancia que un cliente llevaría a una organización, en su relación futura (Rathi & Ravi, 2017; Sifa et al., 2018; Channa, 2019).

Existen diferentes fórmulas para calcular el CLV, Siguiendo a Win & Bo (2020), el *Customer Lifetime Value* y el *Customer Value,* se calculan según lo presentado en la Ecuación 2-1 y Ecuación 2-2; respectivamente. Donde *Churn Rate* es la tasa de deserción del cliente, *Profit margin* el margen de ganancia de la empresa, *Average Order Value* el valor promedio de las compras del cliente y *Purchase Frequency* la frecuencia de compra de cada cliente.

$$\textit{Customer Lifetime Value} = \left(\frac{Customer\ Value}{Churn\ Rate}\right) \times Profit\ margin \qquad (2\text{-}1)$$

$$\textit{Customer Value} = Average\ Order\ Value \times Purchase\ Frequency \quad (2\text{-}2)$$ [1]

Para la predicción del valor Monetario del Cliente, Customer Lifetime Value (CLV), para su uso en la segmentación de clientes, se ha optado por técnicas que se agrupan principalmente en dos enfoques: i) enfoque estadístico o paramétrico, en los cuales a partir de reglas de comportamiento de consumidor y supuestos estadísticos, se estiman parámetros que permiten la predicción del CLV (Jasek et al., 2019) y ii) enfoque de aprendizaje de máquinas, en los cuales se utilizan técnicas supervisadas, principalmente de regresión, para la predicción del CLV (Chen et al., 2019; Bauer & Jannach, 2021).

2.3 Métodos para la predicción del valor monetario de clientes

2.3.1 Métodos paramétricos para la predicción del valor monetario de clientes

Los modelos paramétricos, son los más usados en la predicción del valor monetario de los clientes, donde se han desarrollado paquetes o librerías en lenguajes de programación como R (BTYD y BTYD Plus) y en Python, denominada *Lifetimes.* Jasek et al. (2019), plantea los siguientes como los métodos paramétricos más utilizados en la predicción del valor monetario del cliente:

- Negative Binominal Distribution (Ehrenberg, 1959) l NBD;
- Pareto/NBD (Schmittlein et al., 1987) l Pareto/NBD;
- Beta-geometric/NBD (Fader et al., 2005ª) l BG/NBD;
- Modified Beta-geometric/NBD (Batislam et al., 2007) l MBG/NBD;
- Beta-geometric/NBD with Fixed Regularity (Platzer, 2016) l BG/CNBD-k;
- Modified Beta-geometric/NBD (Platzer, 2016) lMBG/CNBD-k;
- Hierarchical Bayes Pareto/NBD (Ma & Liu, 2007) l Pareto/NBD (HB);
- Variante de Abe del Pareto/NBD (Abe, 2009) l Pareto/NBD (Abe);
- Pareto/Gamma-Gamma-Gamma (Platzer & Reutterer, 2016) l Pareto/GGG.

[1] En el presente estudio, se calcula el valor monetario del cliente, utilizando la fórmula del *Customer Value* propuesta por Win & Bo (2020)

2.3.2 Métodos de aprendizaje de máquinas para la predicción del valor monetario de clientes

El aprendizaje de máquinas, también conocido como aprendizaje automático, machine learning o por sus siglas ML, es una rama de la Inteligencia Artificial, que permite a un sistema aprender de los datos en lugar de a través de la programación explícita (Simeone, 2018; Sodhi, Awasthi y Sharma, 2019). En consecuencia, Hurwitz y Kirsch (2018) sugieren que un modelo de aprendizaje de máquinas es un proceso de inducción de conocimiento a partir del aprendizaje automático generado por el entrenamiento con datos, para generar una salida o modelo, por ejemplo, un algoritmo predictivo creará un modelo predictivo. Luego, cuando proporciona datos del modelo predictivo, recibirá una predicción basada en los datos que entrenaron al modelo.

En el caso de los métodos de aprendizaje de máquinas, se han usado diferentes algoritmos para la predicción del valor monetario del cliente, Win & Bo (2020), en su investigación titulada "*Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm",* cambian el enfoque tradicional del valor monetario del cliente, en el cual pasan de un problema de regresión a un problema de clasificación, en el cual predicen la clase de los clientes para un año siguiente a partir del valor del CLV. Así mismo, Bauer & Jannach (2021), titulado "*Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-based Models",* en el cual proponen un método de Redes Neuronales Recurrentes para la predicción del valor monetario del cliente, donde obtienen como resultados que arrojan mejores resultados que los modelos paramétricos.

3. Revisión sistemática de la literatura

3.1 Introducción

Las revisiones de literatura juegan un papel esencial en la investigación académica para recopilar el conocimiento existente y examinar el estado de un campo. Sin embargo, en diferentes ocasiones, los investigadores de diferentes disciplinas revisiones superficiales y narrativas que carecen de una investigación sistemática de la literatura. Por lo que se propone la elaboración de estudios bibliométricos para realizar revisiones bibliográficas de manera reproducible y científica (Linnenluecke et al., 2020).

Para la revisión de la literatura se tienen en cuenta las siguientes preguntas de investigación:

- ¿Cuáles son las técnicas y modelos usados para predecir el valor monetario del cliente?
- ¿Cuáles son las tendencias en la medición y predicción del valor monetario del cliente?
- ¿Cómo incorporar la predicción del valor monetario del cliente en la segmentación de clientes?

3.2 Metodología

La revisión de la literatura de la presente tesis de investigación fue soportada en la realización de un estudio bibliométrico, técnica que utiliza métodos matemáticos y estadísticos para medir la cantidad y calidad de la producción científica sobre un campo del conocimiento (Cadavid et al., 2012; Valérie & Pierre, 2010). De esta manera, se realiza la evaluación del rendimiento de la actividad científica alrededor de un campo del conocimiento, así mismo, los indicadores o índices bibliométricos proporcionan

información tanto cuantitativa como cualitativa sobre la producción científica, es decir sobre el impacto de esa producción (Velasco et al., 2012).

En el análisis bibliométrico deben ser consideradas bases de datos confiables y rigurosas de conocimiento, como es el caso de las publicaciones académicas reportadas en Scopus (Herrera-Franco et al., 2020), la cual cumple criterios de cantidad de citaciones y accesibilidad, convirtiéndola en una de las bases de datos más usadas en este tipo de análisis de literatura (Dhamija & Bag, 2020; Hall, 2011).

Para la obtención de los datos, se realizó una búsqueda en Scopus a corte de 31 de diciembre de 2021, en la que se utilizó la siguiente ecuación de búsqueda: *(TITLE-ABS-KEY ({customer lifetime value} OR {Customer life time value}) OR TITLE ({CLV}) OR TITLE ({CLTV}) AND NOT TITLE ({CLV-WUS})) AND (TITLE-ABS-KEY (segmentation) OR TITLE-ABS-KEY (cluster*) OR TITLE-ABS-KEY (predict*) OR TITLE-ABS-KEY (model*) OR TITLE-ABS-KEY ({artificial intelligence}) OR TITLE-ABS-KEY ({machine learning}) OR TITLE-ABS-KEY ("neural network*") OR TITLE-ABS-KEY ("23eep learning*") OR TITLE-ABS-KEY ("analytic*")) AND (LIMIT-TO (PUBSTAGE,"final")) AND (LIMIT-TO (DOCTYPE,"ar") OR LIMIT-TO (DOCTYPE,"cp") OR LIMIT-TO (DOCTYPE,"ch") OR LIMIT-TO (DOCTYPE,"re")) AND (LIMIT-TO (LANGUAGE,"English") OR LIMIT-TO (LANGUAGE,"Spanish"))*

Los criterios de inclusión y exclusión utilizados, con el fin de realizar una depuración en la obtención de los datos, son los relacionados a continuación:

Criterios de Inclusión:

- Estar escrito en inglés o español
- Documentos en estado de publicación final
- Estudios con enfoque en mercadeo o clientes/consumidores
- Estudios que utilicen modelos paramétricos o de aprendizaje de máquinas

Criterios de exclusión:

- Las temáticas abordadas no se relacionan a ninguna de las preguntas de investigación
- No se refiere a CLV como Customer Lifetime Value

En consecuencia, se obtuvo como resultado un total de 499 documentos científicos de la base de datos de Scopus. Posteriormente, se realizó un procesamiento e interpretación de los datos con ayuda de la herramienta de código abierto Bibliometrix, el cual es un paquete implementado en el entorno de programación de R, que permite la investigación cuantitativa en cienciometría y bibliometría (Aria & Cuccurullo, 2017). Siguiendo a Camps (2008); Peralta et al. (2015) y Gutiérrez-Salcedo et al. (2018;), los indicadores bibliométricos pueden ser de cantidad o productividad, que permiten visualizar el estado real de la ciencia al contabilizar el total de publicaciones de los autores, revistas, etc.; de calidad, que están orientados a valorar la incidencia, visibilidad e impacto de autores, trabajos y/o revistas en un área o disciplina específica; así mismo, están los indicadores de estructura, que miden las redes y conexiones entre los diferentes autores, publicaciones y áreas de conocimiento (Villa et al., 2018; Nita, 2019).

Finalmente, el análisis de los resultados se realizó a nivel de autores, revistas, países; donde se contempló la productividad anual, citaciones, estructuras conceptuales e intelectuales; así como la identificación, evolución y presentación de los temas de tendencia en el área de conocimiento.

3.3 Resultados

Para el análisis de los resultados, se tiene como insumo los 499 documentos científicos obtenidos de la base de datos de Scopus, los cuales datan desde el año 1987 hasta el año 2021. En la Tabla 3-1 se relaciona un resumen de los datos obtenidos, en la que se observa, entre otros, que las publicaciones son realizadas en un total de 304 fuentes distintas, que el tipo de documento más publicado es artículos con un total de 312 respecto al total, así mismo, se relacionan 1031 autores y en promedio cada autor publica 0,484 artículos.

Tabla 3-1: Resumen de los datos obtenidos en Scopus

Descripción	**Resultados**
Información principal sobre los datos	
Periodo de los datos	1987:2021
Fuentes (revistas, libros, etc.)	304

Descripción	Resultados
Documentos	499
Promedio de años desde la publicación	9,380
Promedio citaciones por documentos	28,390
Promedio citaciones por año por documento	2,217
Referencias	16440
Autores	1031
Tipos de documentos	
Artículos	312
Capítulos de libro	27
Documentos de conferencias	146
Revisiones	14
Colaboración de autores	
Documentos de único autor	60
Documentos por autor	0,484
Autores por documento	2,070
Co-autores por documentos	2,710
Índice de colaboración	2,230

Fuente: elaboración propia a partir de los resultados de Bibliometrix

Se presenta a continuación, un análisis de los indicadores propuestos en la metodología, a saber: indicadores de productividad, indicadores de impacto e indicadores de estructura.

3.3.1 Indicadores de productividad

En primera instancia, se analizan los indicadores de cantidad o productividad alrededor del campo de la segmentación de clientes enfocada en el análisis del *customer lifetime value*. En este sentido, como se presenta en la Figura 3-1, respecto al comportamiento histórico de la producción científica anual, se observa una tendencia creciente en el campo de estudio, reportando una tasa de crecimiento anual de 13,56%, donde el inicio del comportamiento creciente más representativo corresponde al año 2003, así mismo, el año 2011 corresponde al de mayor producción científica, con un total de 42 documentos publicados.

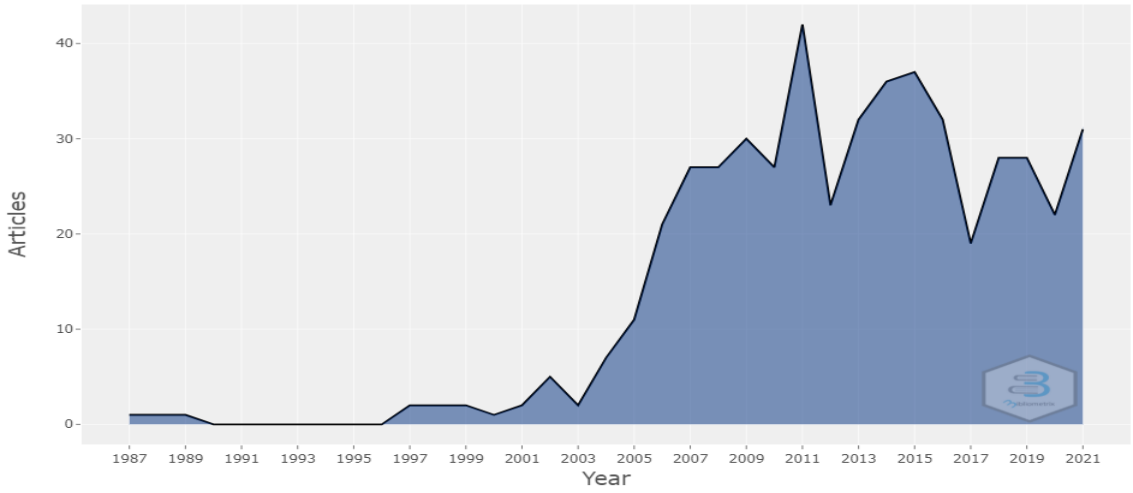

Figura 3-1: Producción científica anual histórica.

Fuente: elaboración propia a partir de la herramienta Bibliometrix

A nivel de autores, la figura 3-2 ilustra los 15 autores más importantes, según el total de producción científica en el periodo de análisis, el gráfico se amplía hasta el número 23 dado los valores iguales en los últimos 9 autores, donde se identifica que el investigador Viswanathan Kumar lidera el listado con un total de 18 publicaciones, seguido del científico Peter S. Fader con 10 publicaciones y en tercera posición el autor Sunil Gupta con 9 publicaciones.

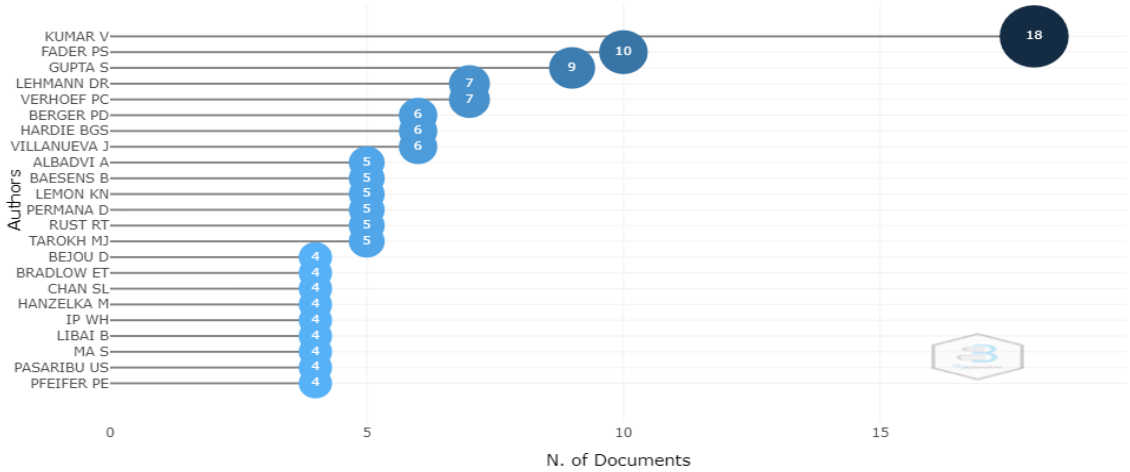

Figura 3-2: Principales autores en cantidad de publicaciones

Fuente: elaboración propia a partir de la herramienta Bibliometrix

Para analizar la productividad de los autores, se utiliza la Ley de Lotka, introducida originalmente por Alfred Lotka en 1926, esta ley enuncia que existe una relación

cuantitativa entre los autores y las contribuciones producidas en un campo dado a lo largo de un periodo de tiempo y que esta relación discreta resulta desigual, puesto que la mayoría de los autores publican pequeñas cantidades de artículos, y que la mayor parte de los artículos, provienen de una pequeña porción de autores altamente productivos (Ardanuy, 2012). En ese sentido, se presenta en la figura 3-3 los resultados de la aplicación de la Ley de Lotka, en el que se observa que el comportamiento empírico de los datos es cercano al comportamiento teórico esperado, donde se presenta que, del total de 1031 autores, el 81,38% presenta una única publicación en el periodo de análisis.

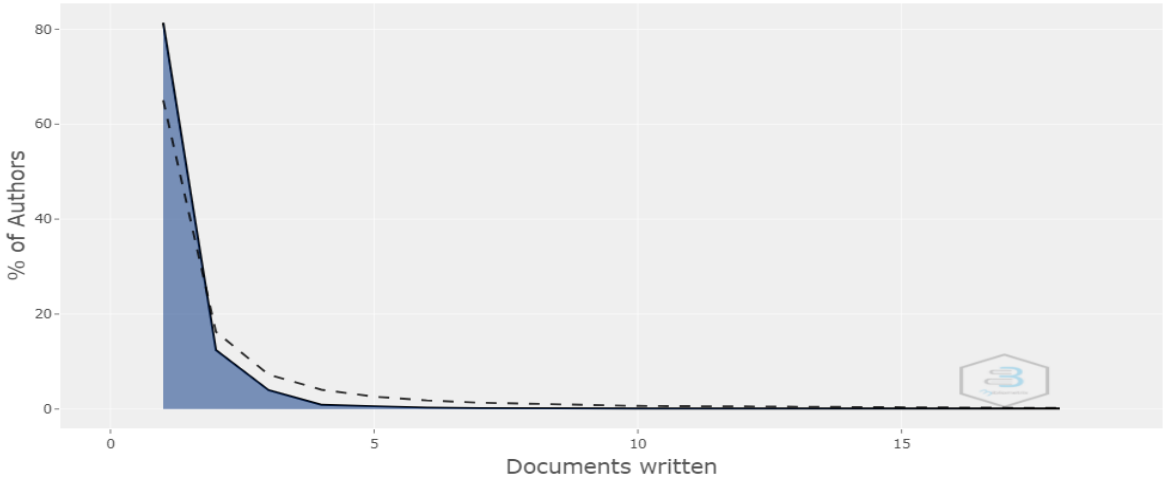

Figura 3-3: Ley de Lotka de la productividad de los autores
Fuente: elaboración propia a partir de la herramienta Bibliometrix

Respecto a las afiliaciones a las cuales pertecen los autores, se presenta en la Figura 3-4 las 15 principales instituciones, en las que se observa que la *Islamic Azad University* de Irán lidera la lista con 15 publicaciones, seguido de *Georgia State University* y *University of Pennsylvania*, ambas con 14 publicaciones, y en tercera posición Tarbiat Modares University y University of Economics, ambas con 13 publicaciones. Al explorar los autores y las afiliaciones, se observa que en algunas publicaciones los autores no han reportado afiliación, o han reportado diferentes afiliaciones en distintas publicaciones.

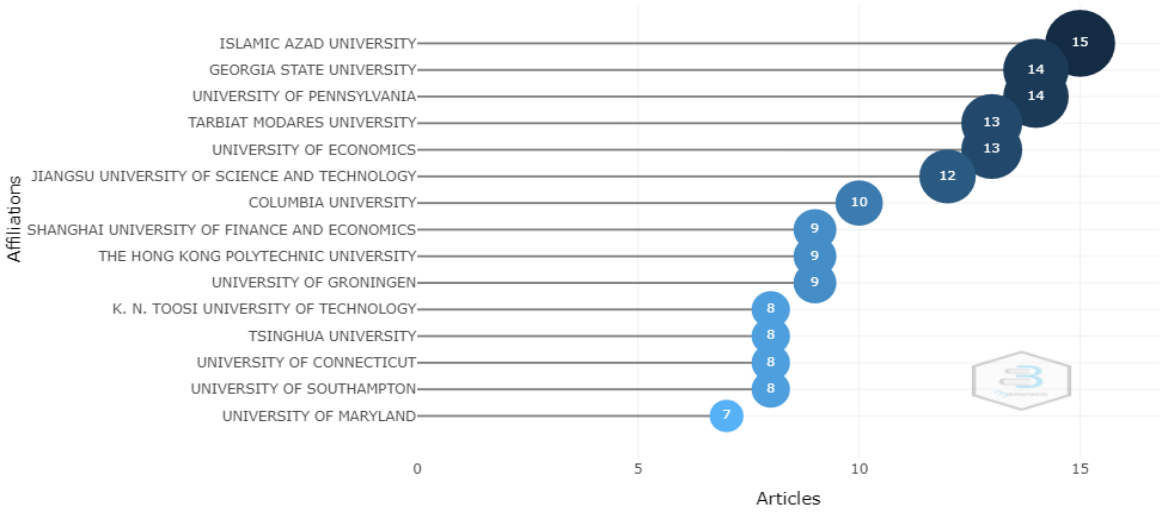

Figura 3-4: Principales afiliaciones en cantidad de publicaciones
Fuente: elaboración propia a partir de la herramienta Bibliometrix

En la Figura 3-5 se relacionan las 15 principales fuentes en las que se publican la mayor cantidad de documentos científicos que conforman la base de datos de análisis, en esta se observa que *Marketing Science* lidera el listado con un total de 19 publicaciones, seguida de *Journal of Business Research* con un total de 15 publicaciones y en tercera posición *Journal of Interactive Marketing* con 13 publicaciones. Se evidencia la presencia de revistas de diferentes áreas, entre ellas, el mercadeo, investigación de operaciones y sistemas y ciencias de la computación; en esta última, se destacan fuentes como *Expert Systems With Applications*, *Association for Computing Machinery International Conference Proceeding Series* y *Procedia Computer Science*.

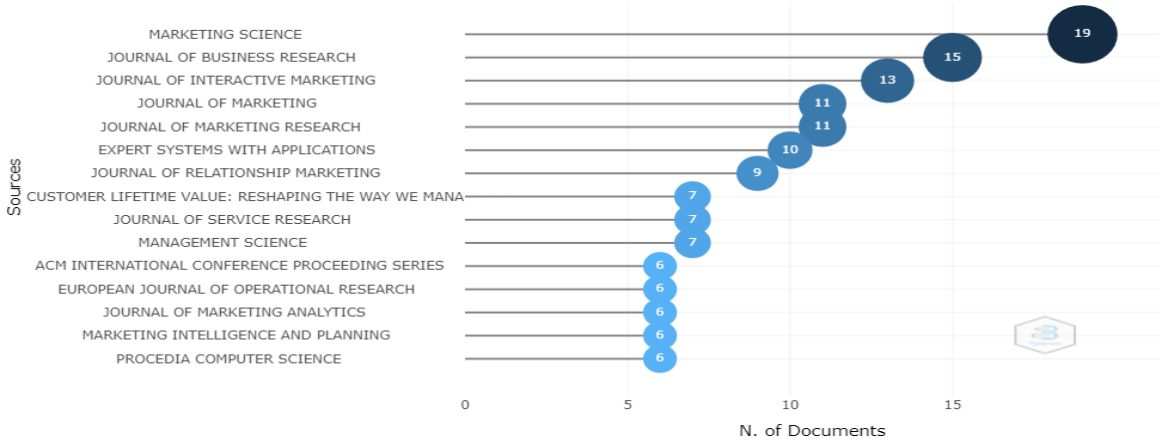

Figura 3-5: Principales fuentes en cantidad de publicaciones
Fuente: elaboración propia a partir de la herramienta Bibliometrix

Finalmente, en relación con los países, se presentan en la Tabla 3-2 los principales 15, teniendo en cuenta su aparición en publicaciones científicas. La base total corresponde a

52 países y 1299 apariciones en publicaciones, en ese sentido, Estados Unidos lidera el listado con 296 apariciones, que corresponden al 22,79% del total, seguido de China con 205 apariciones en publicaciones, que equivalen al 15,78% del total, y en tercera posición Irán con 95 publicaciones, representando el 7,31% del total. Ahora, al analizar el porcentaje acumulado de publicaciones, se evidencia una concentración de la producción científica en pocos países, dado que el 79,45% de los documentos es aportado por el 28,85% de los países, lo cual se encuentra en los márgenes para afirmar que existe un comportamiento que cumple con el Principio o Ley de Pareto (Pareto, 1896). Adicionalmente, se identifica que, el primer país de América Latina en aparecer en la lista es Brasil en la posición 23, con 15 apariciones; así mismo, Colombia se encuentra en la posición 40 con 3 apariciones.

Tabla 3-2: Principales países en aparición de publicaciones científica

País	**Publicaciones**	**Porcentaje**	**Porcentaje acumulado**
Estados Unidos	296	22,79%	22,79%
China	205	15,78%	38,57%
Irán	95	7,31%	45,88%
Alemania	65	5,00%	50,89%
Indonesia	52	4,00%	54,89%
Reino Unido	51	3,93%	58,81%
India	43	3,31%	62,12%
España	33	2,54%	64,67%
Corea del Sur	31	2,39%	67,05%
República Checa	29	2,23%	69,28%
Japón	29	2,23%	71,52%
Canadá	28	2,16%	73,67%
Países Bajos	26	2,00%	75,67%
Turquía	25	1,92%	77,60%
Italia	24	1,85%	79,45%

Fuente: elaboración propia

3.3.2 Indicadores de impacto

Siguiendo con la metodología propuesta, se tienen los indicadores de impacto o de calidad, es así como en la Tabla 3-3 se relacionan los principales 15 autores según su impacto en la producción científica, el orden de la lista es generado con la variable Índice h como primera regla y como segunda regla el total de citaciones. El índice h fue propuesto por George Hirsch y es un indicador en función de la cantidad de citas que han recibido los artículos científicos de un autor, calculado al ordenar de mayor o menor los artículos científicos según el número de citas recibidas, siendo el índice h el número en el que coinciden el número de orden con el número de citas (Hirsch, 2005). Como principal hallazgo se tiene que los 3 primeros en términos de impacto concuerdan con los 3 primeros en términos de productividad, a saber, Viswanathan Kumar con un Índice h de 16, Peter S. Fader con un Índice h de 8 y Sunil Gupta con un Índice h de 6, en ese orden.

Tabla 3-3: Principales autores en indicadores de impacto

Autor	**Índice h**	**Documentos publicados**	**Total citaciones**	**Año de inicio**
Kumar V	16	18	1911	2002
Fader PS	8	10	587	2005
Gupta S	6	8	1188	2003
Verhoef PC	6	7	999	2001
Rust RT	5	5	1146	1999
Berger PD	5	6	920	1998
Lehmann DR	5	6	509	2002
Villanueva J	5	6	502	2007
Hardie BGS	5	6	499	2005
Baesens B	5	5	222	2009
Libai B	4	4	291	2002
Bradlow ET	4	4	125	2006
Pfeifer PE	4	4	66	2004
Albadvi A	4	5	26	2011
Bolton RN	3	3	1282	1998

Fuente: elaboración propia

En la Tabla 3-4 se relacionan las 15 principales fuentes, respecto al impacto generado por las publicaciones que se realizan en éstas. La ordenación de los datos se presenta con el Índice h como variable principal y en segunda instancia el total de citaciones de las fuentes. Es así, como se observa que el listado es liderado por *Marketing Science,* categorizado en Q1 en *Scimago Journal & Country Rank,* con un Índice h de 13, que es interpretado como el total de publicaciones de esa revista que presentan 13 o más citaciones. En segunda posición se encuentra *Journal of Marketing,* Q1 en *Scimago Journal & Country Rank* y un Índice h de 11, y, en tercer lugar, se encuentra *Journal of Marketing Research,* con un Índice h de 10 y Q1 en *Scimago Journal & Country Rank.*

Tabla 3-4: Principales fuentes en impacto de publicaciones

Fuente	**Índice h**	**Documentos publicados**	**Total citaciones**	**Año de inicio**
Marketing Science	13	17	2618	1998
Journal of Marketing	11	11	1028	2006
Journal of Marketing Research	10	11	1376	2005
Journal of Interactive Marketing	10	13	1155	1998
Journal of Business Research	9	14	755	2006
Expert Systems With Applications	8	10	803	2004
Journal of Service Research	7	7	1282	2002
Management Science	7	7	337	2006
Procedia Computer Science	6	6	201	2011
European Journal of Operational Research	6	6	176	2007
Journal of Relationship Marketing	6	9	121	2006
International Journal of Research In Marketing	5	5	124	2007
Journal of Retailing	4	4	448	2001
European Management Journal	4	4	229	1997
Quantitative Marketing and Economics	3	3	131	2007

Fuente: elaboración propia

Finalmente, con relación a los documentos mayor cantidad de citas locales, es decir, las citas que se reportan entre los mismos 499 publicaciones que hacen parte del presente análisis, se relacionan en la Tabla 3-5 los principales 5 documentos, el autor principal, año de publicación y citaciones locales reportadas, donde se observa que el artículo titulado *Customer Lifetime Value: Marketing Models And Applications,* escrito por Paul Berger encabeza el listado con un total de 155 citaciones locales, es de resaltar, que Berger aparece en la posición 6, tanto en el indicador de productividad de autores como de impacto de los autores.

Tabla 3-5: Documentos publicados de mayor impacto en citaciones locales

Documento	**Autor principal**	**Año**	**Citaciones locales**
Customer Lifetime Value: Marketing Models And Applications	Berger PD	1998	155
Modeling Customer Lifetime Value	Gupta S	2006	93
RFM and CLV: Using Iso-Value Curves for Customer Base Analysis	Fader PS	2005	72
Customers as assets	Gupta S	2003	63
An LTV model and customer segmentation based on customer value: a case study on the wireless telecommunication industry	Hwang H	2004	54

Fuente: elaboración propia

3.3.3 Indicadores de estructura

En esta sección, se presentan los resultados relacionados a los indicadores de estructura, los cuales evalúan la conectividad entre las publicaciones, autores y las áreas de conocimiento, para lo que es usual la construcción y análisis de redes sociales (Rueda et al., 2007). En ese sentido, se analiza en primera instancia la estructura conceptual, mediante el uso de redes de co-ocurrencias, lo que permite indagar sobre las conexiones que ocurren entre los mismos términos que aparecen en una colección específica. Siguiendo a Forliano et al. (2021), cuantas más coocurrencias se identifican, más cerca del centro aparecen las palabras en el mapa de la red, cuantas más palabras clave son utilizadas al mismo tiempo por los autores, mayor es su proximidad, lo que resulta en

enlaces más cercanos y robustos, cuanto más se utiliza una palabra clave por parte de los académicos, mayor es su burbuja.

Se presenta en la Figura 3-6 los resultados de las co-ocurrencias generadas a partir de las palabras claves de Scopus, denominadas *keywords plus,* donde se utilizó una configuración de parámetros fijada en 30 nodos y método de segmentación de Louvain, es así como se evidencia la presencia de 3 grupos:

i) Gestión del cliente y medición del *customer lifetime value*: visualizado en color rojo y con la palabra clave *customer lifetime value* como nodo principal, enlaza distintos términos alrededor de la gestión del relacionamiento con los clientes y términos relacionados con técnicas usadas para medición del CLV.

ii) Segmentación de clientes: visualizado en color verde, con la palabra clave *customer segmentation* como nodo principal, se destaca el término k-means clustering, como una de las técnicas más utilizadas para la segmentación de clientes.

iii) Técnicas y métodos para la segmentación y predicción del CLV: en color azul y con la palabra clave data mining como nodo principal, relaciona el área de algunos métodos usados para la predicción del CLV en la segmentación de clientes.

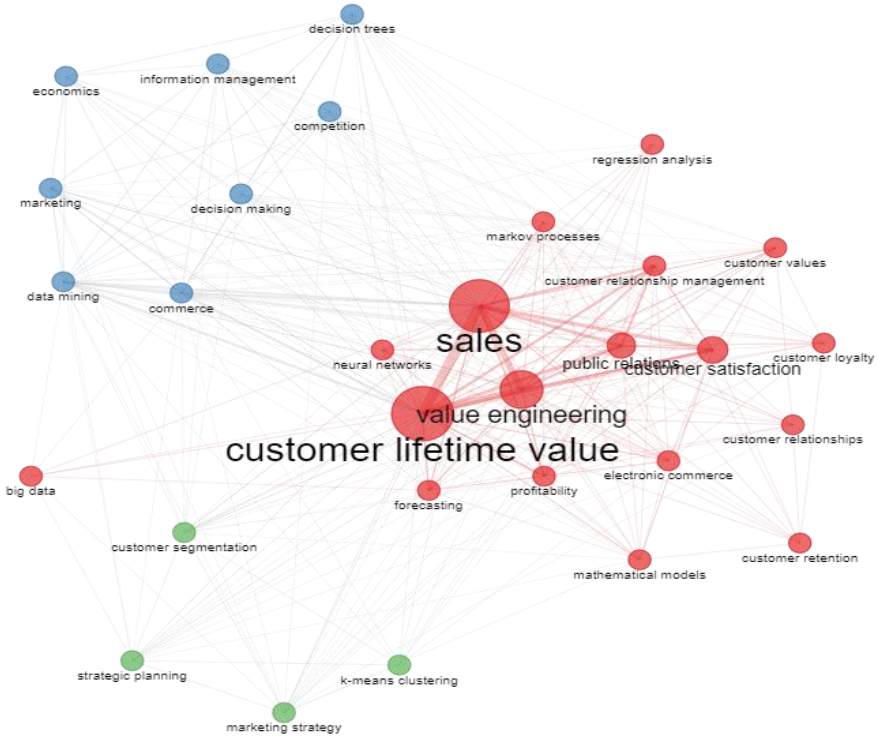

Figura 3-6: Redes de co-ocurrencia de la estructura conceptual
Fuente: elaboración propia utilizando la herramienta biblioshiny

A nivel de estructura intelectual, se tienen las redes de co-citación, la cual se presentan en la Figura 3-7, con una cantidad de 20 nodos, se evidencia que existe un grupo, color rojo, robusto conformado por los principales autores del campo de conocimiento, los cuales fueron identificados desde los análisis de productividad e impacto, tales como Gupta, como nodo principal, donde se suman Kumar, Berger y Fader, entre otros. Un segundo grupo, de color azul, con Kim como nodo principal, reflejando un enlace fuerte con los otros grupos. Finalmente, el grupo 3, de color verde, en el que se identifica a Bolton como nodo principal. En consecuencia, el comportamiento de la Figura 3-7 aporta evidencia a favor de que el campo de conocimiento en estudio es consistente en la generación de nuevos conocimientos a partir del conocimiento ya creado.

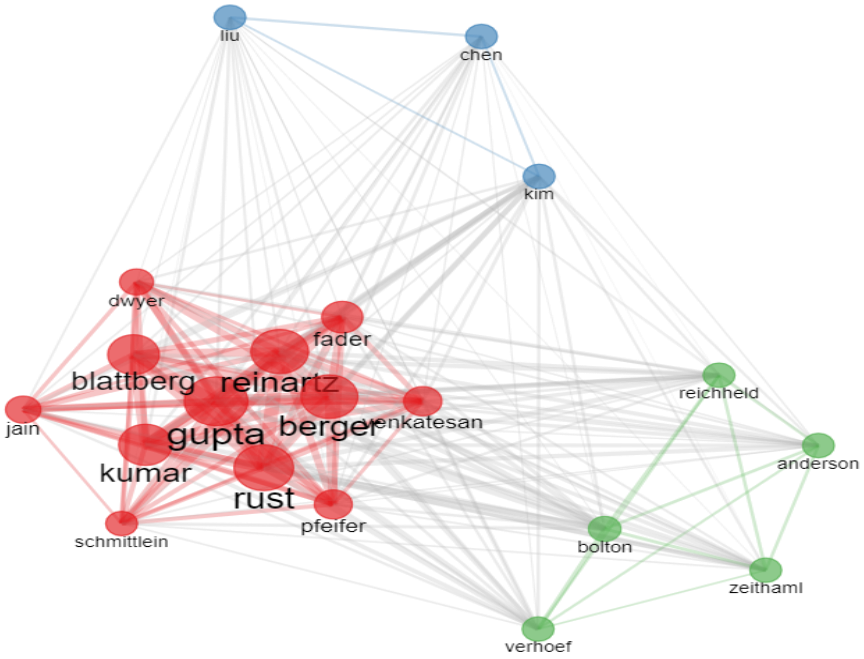

Figura 3-7: Redes de co-citación de la estructura intelectual
Fuente: elaboración propia utilizando la herramienta biblioshiny

Finalmente, la Tabla 3-6, muestra el resumen de los artículos más relevantes de los últimos 3 años en la revisión de la literatura, a partir de las preguntas de investigación propuestas en el análisis bibliométrico, número de citas y la pertinencia acorde a los objetivos de la presente tesis.

Tabla 3-6: Resumen de artículos relevantes de la revisión sistemática de literatura

Autor	Título	Año	Número de citas	Tipo de modelo	Modelo o método	Área de aplicación
Chen PP; Guitart A; Del Ro AF; Periez A	Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models	2019	15	Paramétrico Aprendizaje de máquinas	-Deep neural network -Pareto/NBD	Video games
Jasek P; Vrana L; Sperkova L; Smutny Z; Kobulsky M	Comparative analysis of selected probabilistic customer lifetime value models in online shoping	2019	7	Paramétricos	BG/NBD Pareto/NBD	Online retail
Win TT; Bo KS	Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm	2020	1	Aprendizaje de máquinas	Random Forest	Online retail

Autor	Título	Año	Número de citas	Tipo de modelo	Modelo o método	Área de aplicación
Heldt R; Silveira CS; Luce FB	Predicting customer value per product: From RFM to RFM/P	2021	9	Paramétricos	RFMP	Financial services company
Bauer J; Jannach D	Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-based Models	2021	0	Aprendizaje de máquinas	Recurrent Neural Networks	Online retail

Fuente: elaboración propia

3.4 Conclusiones

En este capítulo, se logró el cumplimiento del primer objetivo específico de la tesis, donde se identifica un campo del conocimiento en crecimiento, con oportunidades del uso del aprendizaje de máquinas para la predicción del valor monetario de los clientes, y así mismo su uso en la segmentación de clientes. En la revisión de la literatura, se identificó una línea de investigación en la cual autores proponen abordar la predicción del valor monetario del cliente mediante técnicas de aprendizaje de máquinas, sin embargo, no se encontró de manera detallada cómo comparar estos resultados con los métodos paramétricos o estadísticos.

Se observa una baja participación de países latinoamericanos en los principales países del campo, y se identifica que las publicaciones son de interés tanto en el campo de las revistas de negocios y ciencias sociales, como en revistas relacionadas a las ciencias de la computación. Así mismo se concluye que es un campo del conocimiento que presenta ciclos de aproximadamente 5 años, en los cuales se renuevan los intereses alrededor del campo.

4. Caracterización de los datos

4.1 Introducción

En este capítulo se desarrolla el proceso utilizado para la caracterización de los datos de la presente tesis, donde se tomó como insumo una base de datos abierta, denominada Online Retail II, que se encuentra en el repositorio abierto de la UCI. Así mismo, el presente apartado permite dar cumplimiento al objetivo específico número 2 de la presente tesis.

4.2 Metodología

Para la caracterización de los datos, se utilizó un proceso metodológico discriminado en 4 etapas: i) obtención de los datos y entendimiento de los datos; ii) exploración y preprocesamiento de los datos; y iii) preparación de datos a nivel de cliente; tal y como se ilustran en la Figura 4-1.

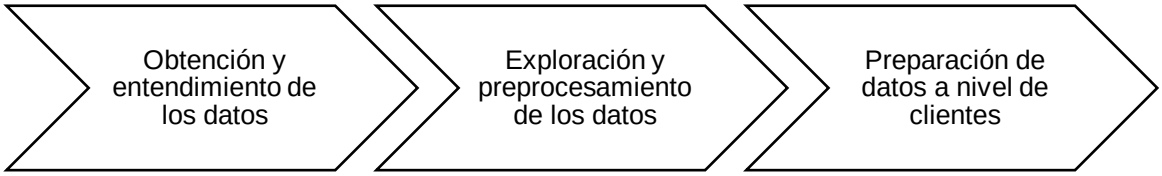

Figura 4-1: Metodología propuesta para caracterización de los datos
Fuente: elaboración propia

En la etapa i) obtención de los datos y entendimiento de los datos, se tuvo en cuenta los siguientes criterios:

- Corresponder a una base de datos de clientes
- Disponer de histórico de comportamiento de compras de mínimo 1 año
- Registro transaccional de compras con variable de identificación única del cliente
- Marcación de fecha de cada uno de los registros de compras o transacciones
- Variable monto o valor de cada una de las compras

Para el entendimiento de los datos, se realizó una interpretación de cada una de las variables de la base de datos obtenida, así como una marcación del tipo de variable y descripción o explicación de su significado.

En la etapa ii) exploración y preprocesamiento de los datos, se tuvo como principal objetivo la limpieza de transacciones o registros de compra atípicos o de comportamiento anómalo. Es así como primero se realiza una depuración de los datos aplicando reglas de negocio y limpieza manual, luego se realiza una detección de datos atípicos utilizando algoritmos para este fin, en este caso *Local Outlier Factor* (Breunig et al., 2000) e *Isolation Forest* (Liu et al., 2008)*;* ambos disponibles en la librería de Scikit Learn de Python (Pedregosa et al., 2011), dado que, siguiendo a Chen et al. (2012), algunos datos, principalmente extremos, pueden ser transacciones válidas respecto al negocio, sin embargo se consideran atípicas respecto a los datos.

Finalmente, en la etapa iii) preparación de datos a nivel de cliente, se realizó la definición de los periodos a utilizar para modelado y testeo, la partición de los datos de acuerdo con los conjuntos requeridos para las etapas de modelado, y la transformación de los datos transaccionales (por fecha de compra) a una base de datos agregada a nivel de clientes.

4.3 Resultados

4.3.1 Obtención y entendimiento de los datos

A continuación, se presentan los resultados del proceso de obtención y entendimiento de los datos, siguiendo la metodología y criterios propuestos para la búsqueda y obtención de los datos, a saber: corresponder a una base de datos de clientes; disponer de histórico de comportamiento de compras de mínimo 1 año; registro transaccional de compras con variable de identificación única del cliente; marcación de fecha de cada uno de los registros de compras o transacciones; y variable monto o valor de cada una de las compras. Se encontró que la base de datos denominada *Online Retail II Data Set,* donada por Daqing Chen, de la *School of Engineering, London South Bank University, y* que se encuentra disponible en el repositorio libre *UCI Machine Learning Repository,* de la Universidad de California, Irvine (Chen, 2019); cumple con los requisitos enunciados.

La base de datos *Online Retail II Data Set* cuenta con datos de venta de una empresa minorista en línea de Reino Unido, la cual fue fundada en 1981, su canal de venta

principal es Amazon y dispone de 80 empleados. Este conjunto de datos contiene todas las transacciones o compras que se produjeron entre el 12 de diciembre del 2009 y el 9 de diciembre del 2011, donde se reporta un total 1.067.371 registros transaccionales, 8 variables y un total de 5.833 clientes. La empresa vende principalmente artículos de regalo para toda ocasión, y cuenta con la presencia de clientes minoristas y mayoristas, no diferenciados en los datos del negocio. En la Tabla 4-1 se muestra un resumen del conjunto de datos.

Tabla 4-1: Resumen del conjunto de datos *Online Retail II*

Característica	**Descripción**
Sector	Minorista o *retail*
Fundación	1981, Reino Unido
Canal de ventas	En línea, Amazon
Empleados	80
Periodo	01/12/2009 – 09/12/2011
Número de registros	1.067.371
Total variables	8
Total clientes	5.833

Fuente: elaboración propia a partir de Chen (2019)

El conjunto de datos cuenta con 8 variables, *InvoiceNo, StockCode, Description, Quantity, InvoiceDate, UnitPrice y CustomerID.* En la Tabla 4-2 se presenta cada una de ellas, con el tipo de variable al que corresponde y una descripción sobre el significado de la variable. Como se puede evidenciar, se cuenta con 3 variables tipo cuantitativas y 5 variables tipo cualitativas.

Tabla 4-2: Información de variables del conjunto de datos

Variable	**Tipo de variable**	**Descripción**
InvoiceNo	Cualitativa nominal	Corresponde al código único que identifica a cada transacción
StockCode	Cualitativa nominal	Corresponde al código asignado a cada producto del inventario de la empresa
Description	Cualitativa nominal	Nombre de cada uno de los productos del inventario de la tienda

Variable	Tipo de variable	Descripción
Quantity	Cuantitativa discreta	Cantidad adquirida por el cliente de cada producto en cada transacción
InvoiceDate	Cuantitativa continua	Fecha y hora en la que se realiza el registro de compra o transacción
UnitPrice	Cuantitativa continua	Precio unitario asociado al producto en cada transacción, expresado en libras esterlinas
CustomerID	Cualitativa nominal	Identificador único asignado a cada cliente
Country	Cualitativa nominal	País de residencia del cliente

Fuente: elaboración propia

4.3.2 Exploración y preprocesamiento de los datos

En esta sección, se realiza una exploración inicial de los datos con el fin de efectuar una limpieza y preprocesamiento de estos. Primero, se presenta en la Tabla 4-3 el encabezado de los 5 primeros registros de los datos originales obtenidos.

Tabla 4-3: Encabezado del conjunto de datos original

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	Price	Customer ID	Country
489434	85048	15CM CHRISTMAS GLASS BALL 20 LIGHTS	12	1/12/2009 7:45	6,95	13085,0	United Kingdom
489434	79323P	PINK CHERRY LIGHTS	12	1/12/2009 7:45	6,75	13085,0	United Kingdom
489434	79323W	WHITE CHERRY LIGHTS	12	1/12/2009 7:45	6,75	13085,0	United Kingdom
489434	22041	RECORD FRAME 7" SINGLE SIZE	48	1/12/2009 7:45	2,10	13085,0	United Kingdom
489434	21232	STRAWBERRY CERAMIC TRINKET BOX	24	1/12/2009 7:45	1,25	13085,0	United Kingdom

Fuente: elaboración propia

Para un mejor entendimiento de las variables que conforman el conjunto de datos y lograr una mejor consistencia con el resto de documento, se hace una traducción al español de los nombres de las 8 variables, En la Tabla 4-4 se presenta la transformación realizada,

Tabla 4-4: Conversión en nombre de variables

Nombre original de la variable	Nombre traducido
InvoiceNo	Código_transacción
StockCode	Código_producto
Description	Descripción
Quantity	Cantidad
InvoiceDate	Fecha
Price	Precio
CustomerID	ID_Cliente
Country	País

Fuente: elaboración propia

En la Figura 4-2 se ilustra el comportamiento histórico de las compras o transacciones, teniendo en cuenta el horizonte temporal de análisis, es así como se observa la presencia de ciclos, siendo enero, tanto en 2010 como en 2011, meses de bajos niveles de ventas; luego un leve crecimiento que se refleja en una nueva caída para el mes de agosto; finalmente se evidencia la presencia de un nuevo pico o crecimiento de las compras en el último trimestre del año,

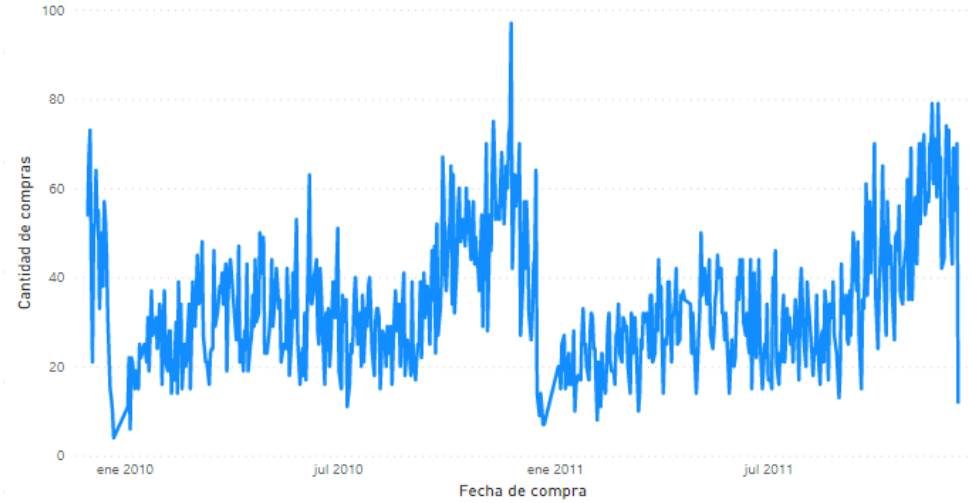

Figura 4-2: Comportamiento histórico de las compras
Fuente: elaboración propia

En esta etapa del proceso, se realiza una inspección de la presencia de registros nulos en el conjunto de datos, donde se encuentra que del total de 1.067.371 transacciones, 243.007 presentan *ID_Cliente* nulos, para efectos de esta tesis, se hace indispensable identificar el cliente asociado a los registros de compra, puesto que el objetivo es realizar análisis de segmentación de clientes, en ese sentido, se eliminan los registros de transacciones asociadas a *ID_Cliente* nulos, dejando disponibles un total 824.364 registros con un *ID_Cliente* asociado, La otra variable que refleja presencia de datos nulos es *Descripción*, mostrando 1.062.989 registros con valor asociado, sin embargo, en este caso no tiene afectación de cara a los objetivos del estudio.

Sobre la base de 824.364 registros válidos, en términos de la completitud de los datos, se realiza un análisis de percentiles sobre las 2 variables cuantitativas diferentes del campo *Fecha*, es decir, sobre las variables *Cantidad y Precio,* Los resultados se visualizan en la Tabla 4-5, en la que se observa la presencia de valores negativos en el mínimo y percentil 1% de la variable *Cantidad,* así como un incremento con indicios anómalos en el percentil 99% y valor máximo de esta variable, Así mismo, se observa en la variable *Precio* valores en 0 y un salto con indicios anómalos en el percentil 99% y el valor máximo, Siguiendo reglas de negocio, estos datos pueden corresponder a devoluciones, cancelaciones, errores humanos en la facturación, entre otros, los cuales serán analizados más adelante en este capítulo.

Tabla 4-5: Métricas y percentiles del conjunto de datos

Métricas y percentiles	**Cantidad**	**Precio**
Conteo	824364,000	824364,000
Media	12,414	3,677
Desviación estándar	188,976	70,241
Mín	-80995,000	0,000
1%	-2,000	0,290
5%	1,000	0,420
10%	1,000	0,550
25%	2,000	1,250
50%	5,000	1,950
75%	12,000	3,750
90%	24,000	6,750
95%	36,000	8,500
99%	120,000	14,950

Métricas y percentiles	Cantidad	Precio
Máx,	80995,000	38970,000

Fuente: elaboración propia

En lo que concierne a las variables cualitativas del conjunto de datos, una de ellas es el *País,* este corresponde al país origen del *ID_Cliente* que realiza la compra o transacción, En la Tabla 4-6, se presentan los porcentajes correspondientes a esta variable, en sus 10 primeras posiciones, donde se identifica que el 89,92% de las compras, están asociadas a clientes de *United Kingdom.* Esta es una variable sociodemográfica que puede impactar en un comportamiento heterogéneo de los clientes respecto a sus hábitos de compra o transaccionales, en el sentido que la empresa tiene sede principal en *United Kingdom,* y que las compras de otros países pueden contener cargas adicionales o factores externos que sesguen el presente estudio. Por lo anterior, y teniendo en cuenta que esta investigación se enfoca en segmentación de clientes a partir de un análisis transaccional y de comportamiento de compras, se limitará el conjunto de datos a los clientes correspondientes a *United Kingdom*. Es así como se pasa de 824.364 registros a 741.301.

Tabla 4-6: Proporción de registros por País

País	Porcentaje
United Kingdom	89,92%
Germany	2,14%
EIRE	1,96%
France	1,72%
Netherlands	0,62%
Spain	0,46%
Belgium	0,38%
Switzerland	0,37%
Portugal	0,30%
Australia	0,23%

Fuente: elaboración propia

Retomando las variables cuantitativas, *Precio* y *Cantidad*, Se procede a realizar un análisis a detalle, para indagar respecto a los hallazgos enunciados en la Tabla 4-5. Es así como se inicia la exploración en la variable *Precio,* en esta se identifican valores de

precios en 0, así como la presencia de valores altos que marcan un salto entre el percentil 99% con un valor de 14,950 y el dato máximo con un valor de 38.970,000. Al realizar una revisión manual se encuentra que en valores de *Precio* 0, existen datos con la marcación de *Descripción* de producto “Manual” o “This is a test product”. Así mismo, al explorar datos por encima del percentil 99% se observan valores de “Manual”, “CRUK Commision”, “POSTAGE”, “DOTCOM POSTAGE”, “CARRIAGE”, “Bank Charges”, “Discount” y “Adjustment by […]” en la variable *Descripción*. En la Tabla 4-7 se presenta una muestra del conjunto de datos con estos hallazgos.

Tabla 4-7: Muestra de hallazgos transacciones anómalas

Código_transacción	Código_producto	Descripción	Cantidad	Fecha	Precio	ID_Cliente	País
496115	M	Manual	1	29/01/2010 11:04	8985.60	17949.0	United Kingdom
C576338	CRUK	CRUK Commission	-1	14/11/2011 15:27	1038.75	14096.0	United Kingdom
576339	DOT	DOTCOM POSTAGE	1	14/11/2011 15:27	1500.36	14096.0	United Kingdom
490727	M	Manual	1	7/12/2009 16:38	0.0	17231.0	United Kingdom
497819	TEST001	This is a test product.	5	12/02/2010 14:58	0.0	14103.0	United Kingdom
C489538	POST	POSTAGE	-1	1/12/2009 12:18	9.580	15796.0	United Kingdom
C490943	BANK CHARGES	Bank Charges	-1	8/12/2009 14:08	15.000	16703.0	United Kingdom
490998	C2	CARRIAGE	1	8/12/2009 17:24	50.000	16253.0	United Kingdom
C495737	ADJUST	Adjustment by john on 26/01/2010 16	-1	26/01/2010 16:23	10.500	16154.0	United Kingdom

Fuente: elaboración propia

Ahora, se realiza una exploración respecto al significado de los campos enunciados como extraños o aparentemente anómalos, donde se aclara que, en bases de datos transaccionales se guarda un registro de cada acción realizada, por lo que es frecuente encontrar este tipo registros que requieren limpieza. En la Tabla 4-8 se relacionan las definiciones equivalentes a las descripciones identificadas como anómalas, un total de 1.426 registros presentan estos valores. A este punto, una vez acotado los datos a *País United Kingdom* y realizada la limpieza de los registros que contienen los valores

relacionados en la Tabla 4-8, en el campo *Descripción,* se tiene que el nuevo conjunto de datos consta de 739.875 registros.

Tabla 4-8: Definiciones de valores de indicios anómalos en la variable *Descripción*

Descripción	**Definición**
Manual	Transacciones que son diligenciadas manualmente por los cajeros o facturadores, algunas de ellas pueden presentar errores de digitación
This is a test producto	Corresponden a registros de prueba en la base de datos
CRUK Commision	Iniciativa que paga parte de las ventas a la Cancer Research UK (Cancer Research UK, 2022)
Postage	Corresponde a gastos de envío postal o de sellos en envíos postales
Dotcom Postage	Corresponde a gastos de envío postal o de sellos en envíos postales
Carriage	Corresponde a gastos de transporte en envíos
Bank Charges	Cargos bancarios que se realizan al momento de la transacción
Discount	Descuentos que se realizan en algunas transacciones o compras
Adjustment by	Ajustes que realizan cajeros o facturadores, normalmente son ajustes manuales

Fuente: elaboración propia

Una vez finalizada la exploración de datos extraños en la variable *Precio,* se procede a realizar la exploración en la variable *Cantidad.* En esta se encontró que todas las transacciones con valores negativos inician con la letra “C” en el campo *Código_transacción,* como se observa en la Tabla 4-7, presentada anteriormente; adicionalmente, se identifica que el valor mínimo y máximo en esta variable es el mismo, uno en negativo y otro en positivo, es decir, -80995 como valor mínimo y 80995 como valor máximo, con lo cual se tienen indicios para validar la existencia o no de que estos registros corresponden a ajustes contables, devoluciones, cancelaciones o movimientos

similares. En esta validación, se encontró que en total existen 15.377 transacciones con cantidades negativas, de las cuales 9.854 presentan otro registro de transacción equivalente, tipo contrapartida, en positivo, es decir, donde los valores de todas las variables coinciden, a excepción del signo en la variable *Cantidad.* En la Tabla 4-9 se presenta una muestra de un caso en el que el registro con cantidad negativa presenta en contrapartida otro registro en positivo. Este hecho permite afirmar que estos registros corresponden a ajustes contables, por lo que no solo se debe eliminar el registro negativo, sino también su respectiva contrapartida en positivo.

Tabla 4-9: Muestra de datos variable *Cantidad* en negativo

Código_transacción	Código_producto	Descripción	Cantidad	Fecha	Precio	ID_Cliente	País
581483	23843	PAPER CRAFT, LITTLE BIRDIE	80995	9/12/2011 9:15	2.08	16446.0	United Kingdom
581484	23843	PAPER CRAFT, LITTLE BIRDIE	-80995	9/12/2011 9:27	2.08	16446.0	United Kingdom

Fuente: elaboración propia

En consecuencia, se hace la eliminación los registros con valores negativos en *Cantidad*, sus respectivas contrapartidas, así como la eliminación de 40 registros adicionales que aún conservaban la variable *Precio en 0,* esto genera un total de 25.271 registros que se depuraron del conjunto de datos transaccionales. Finalmente, se presenta en la Tabla 4-10 los resultados obtenidos.

Tabla 4-10: Resumen métricas y percentiles con limpieza manual

Métricas y percentiles	Cantidad	Precio
Conteo	714.604,000	714.604,000
Media	11,663	2,900
Desviación estándar	55,851	4,148
Mín.	1,000	0,001
1%	1,000	0,290
5%	1,000	0,420
10%	1,000	0,550
25%	2,000	1,250
50%	4,000	1,950

Métricas y percentiles	Cantidad	Precio
75%	12,000	3,750
90%	24,000	6,350
95%	36,000	8,500
99%	100,000	12,750
Máx.	10.000,000	649,500

Fuente: elaboración propia

Posterior a la limpieza manual realizada, siguiendo reglas de negocio, se observa aún la presencia de datos extremos, respecto al resto de los datos, siguiendo a Chen et al. (2012), estos datos pueden ser transacciones válidas respecto al negocio, sin embargo se consideran atípicas respecto a los datos. Es por esto que, siguiendo la metodología propuesta, se realiza un análisis de detección de atípicos utilizando 2 algoritmos: *Local Outlier Factor* (Breunig et al., 2000) e *Isolation Forest* (Liu et al., 2008). Para este fin, se crea una nueva variable "Precio total", la cual es el resultado de la multiplicación de los precios unitarios "Precio" por las cantidades de cada producto "Cantidad". En la Tabla 4-11 se muestra las métricas y percentiles de las 3 variables consideradas, a saber: *Cantidad, Precio* y *Precio_total*; donde las 2 variables a utilizar en los algoritmos de detección de atípicos son *Cantidad* y *Precio_total.*

Tabla 4-11: Resumen métricas y percentiles para detección de atípicos

Métricas y percentiles	Cantidad	Precio	Precio_total
Conteo	714.604,000	714.604,000	714.604,000
Media	11,663	2,900	19,565
Desviación estándar	55,851	4,148	80,466
Mín.	1,000	0,001	0,001
1%	1,000	0,290	0,550
5%	1,000	0,420	1,250
10%	1,000	0,550	1,950
25%	2,000	1,250	4,200
50%	4,000	1,950	10,200
75%	12,000	3,750	17,700
90%	24,000	6,350	33,000
95%	36,000	8,500	59,700

Métricas y percentiles	Cantidad	Precio	Precio_total
99%	100,000	12,750	175,200
Máx.	10.000,000	649,500	38.970,000

Fuente: elaboración propia

Como se observa en la Tabla 4-11, los datos con comportamiento anómalo se presentan principalmente por encima del percentil 99%, por lo que, para la implementación de los algoritmos de detección de atípicos, el factor de contaminación como parámetro de entrada, se fija en 1%. Así mismo, la Figura 4-3 representa de dispersión de los datos entre las variables *Cantidad* y *Precio_total,* en la que se observa la presencia de valores alejados, respecto al resto de los datos.

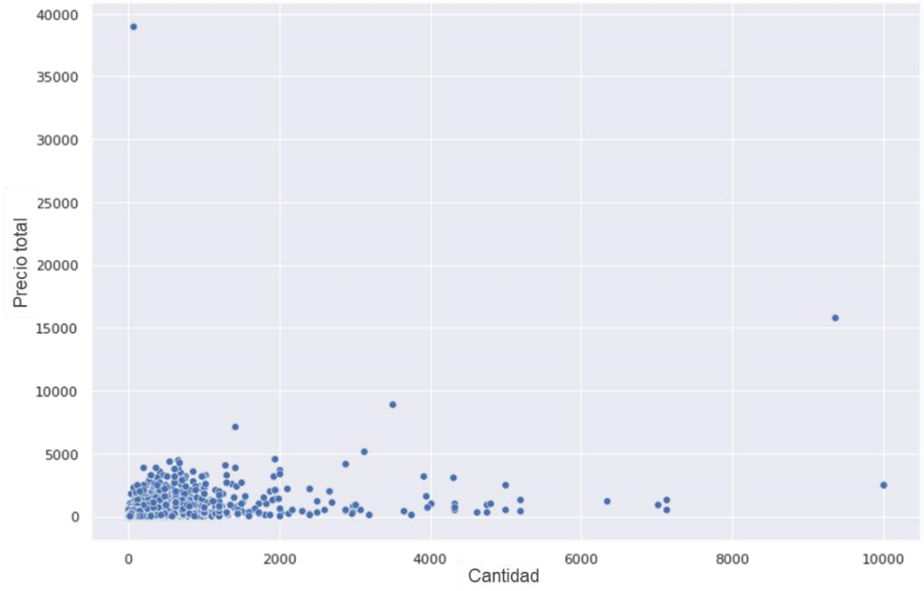

Figura 4-3: Dispersión de los datos en variables Cantidad y Precio_total

Fuente: elaboración propia

Para la implementación del algoritmo de *Isolation Forest*, se utilizó un factor de contaminación del 1%, *random_state* de 0 y *n_jobs* de -2. En el caso del algoritmo de *Local Outlier Factor*, se utilizó un factor de contaminación del 1%, *novelty* en True y *n_jobs* en -2. En ambos algoritmos el escalado de los datos se hizo mediante *StandarScaler*, y las variables utilizadas fueron Precio_total y Cantidad. En la Tabla 4-12 se muestra un resumen de los resultados obtenidos. A partir de los resultados, se observa que *Isolation Forest* logró una mejor corrección, respecto a los valores alejados que se presentaban en el conjunto de datos, mientras que en el algoritmo de *Local*

Outlier Factor aún se evidencia valores atípicos grandes. La elección final es los resultados de *Isolatión Forest*.

Tabla 4-12: Comparativa de algoritmos de detección de atípicos en Cantidad y Precio_total

Algoritmo	Variable	Datos incluidos	Datos atípicos	Media	Valor mínimo	Percentil 1%	Percentil 99%	Valor máximo
Isolation Forest	Cantidad	707565	7039	9,054	1,000	1,000	72,000	200,000
	Precio_total	707565	7039	15,786	0,001	0,550	119,000	282,150
Local Outlier Factor	Cantidad	707766	6838	62,630	1,000	1,000	432,000	10000,000
	Precio_total	707766	6838	100,989	0,100	0,620	651,075	38970,000

Fuente: elaboración propia

Por consiguiente, una vez surgido todo el proceso de exploración y preprocesamiento detallado en la presente sección, el conjunto de datos transaccionales final es conformado por 707.766 registros. En la Figura 4-4 se visualiza la dispersión de los datos entre las variables *Cantidad* y *Precio_total,* donde se observa un comportamiento más homogéneo en los datos, con una corrección de valores alejados o atípicos.

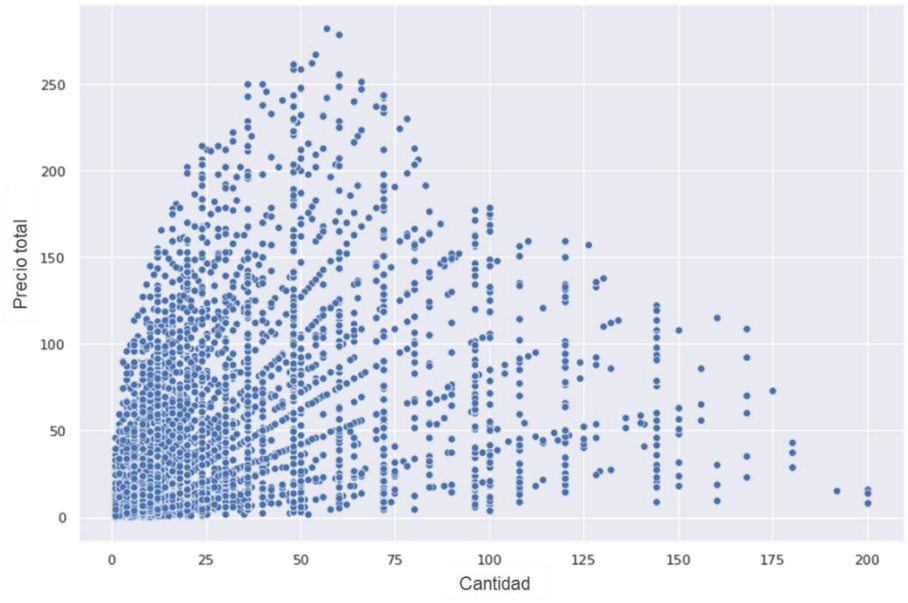

Figura 4-4: Dispersión de los datos en variables Cantidad y Precio_total posterior a detección de atípicos
Fuente: elaboración propia

4.3.3 Preparación de los datos a nivel de clientes

La exploración y prepropresamiento de los datos en la sección 4.3.1, se realizó a nivel de cada registro de transacciones o compras; sin embargo, para el proceso de modelado y dado que los objetivos de la presente tesis se enfocan en la segmentación de clientes, en esta sección se realiza una agregación inicial de los datos a nivel de cada cliente. En primer paso, se toma como insumo que a este punto el conjunto de datos cuenta con una variable adicional al conjunto de datos inicial, el *Precio_total.* Adicionalmente, la variable *Fecha,* se convierte de formato tipo fecha y hora, a solo conservar la fecha, motivado en que, en una misma transacción[2] o compra, para una misma factura, puede haber diferencia de minutos o segundos por cada producto, y en el hecho de que el análisis se centra en los días en los que ocurrió o no evento de compra. En la Tabla 4-13 se refleja una muestra de los datos de insumo para esta sección.

Tabla 4-13: Muestra del conjunto de datos inicial para preparación de datos

Código_ transacción	Código_ producto	Descripción	Cantidad	Fecha	Precio	ID_ Cliente	País	Precio _total
567879	22295	HEART FILIGREE DOVE LARGE	12	22/09/2011	1,65	16161	United Kingdom	19,80
517394	21481	FAWN BLUE HOT WATER BOTTLE	6	29/07/2010	2,95	18264	United Kingdom	17,70
60228	17012D	ORIGAMI ROSE INCENSE/CANDLE SET	10	15/07/2011	0,85	17750	United Kingdom	8,50
568152	22665	RECIPE BOX BLUE SKETCHBOOK DESIGN	1	25/09/2011	2,95	16480	United Kingdom	2,95
561867	47566	PARTY BUNTING	4	31/07/2011	4,95	16931	United Kingdom	19,80

Fuente: elaboración propia

En este punto, el objetivo es conformar un conjunto de datos a nivel de cada cliente, asociando la fecha de compra y el valor total de la compra, a este valor total de la compra

[2] En una misma factura o *Código_transacción* pueden existir varios registros, puesto que en cada compra se pueden adquirir diferentes productos, y cada producto representaría una fila diferente.

se le denominará *Valor monetario de la fecha*[3]. En ese sentido, se crea un conjunto de datos compuesto por el ID_Cliente, Fecha y Valor_monetario_fecha, el cual consta de 29.053 registros y 4.443 clientes; en la Tabla 4-14 se relaciona una muestra de los datos.

Tabla 4-14: Conjunto de datos agregado a nivel de clientes y fecha

ID_Cliente	Fecha	Valor_monetario_fecha
13127	23/09/2011	259,17
16966	13/04/2011	409,20
17243	21/02/2010	148,40
16029	13/09/2010	397,92
15570	30/11/2010	287,15

Fuente elaboración propia

Como se mencionó en la sección 2.3.1 de este capítulo, el periodo total del campo *Fecha* se encuentra entre el 1 de diciembre de 2009 y el 9 de diciembre de 2011. Ahora, para el desarrollo de esta investigación se tiene, entre otros, el objetivo de predecir el valor monetario del cliente; en este sentido, se propone aislar los últimos 6 meses de los datos, para efectos de pronóstico y validación, por consiguiente, se hace una partición de los datos en la fecha 31 de mayo de 2011. La Figura 4-5 ilustra las particiones realizadas en términos temporales. En este punto, la partición es únicamente de marcación de qué compras tuvo cada clientes antes de la fecha de corte (periodo de calibración) y qué compras después (periodo pronóstico).

[3] Esta variable es el insumo principal que se utilizará más adelante para el cálculo del Valor monetario del cliente

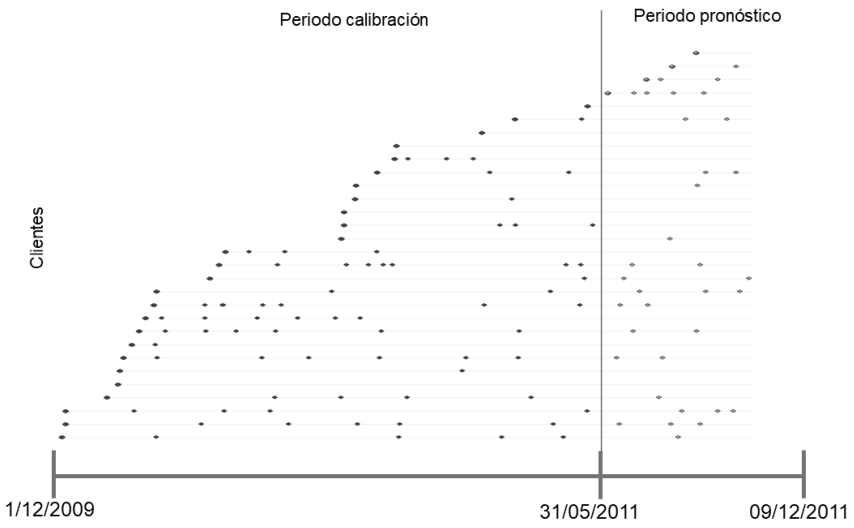

Figura 4-5: Partición temporal de los datos de compras por clientes
Fuente: elaboración propia

Es de resaltar que, en esta fase de la preparación de los datos, se tienen en cuenta solo aquellos clientes que presentan al menos 2 o más compras en el periodo de modelado, esto sustentado en que no se puede predecir comportamiento de compras de clientes que en ese periodo “no existen” para los modelos. Siguiendo Fader et al. (2005); Chen et al. (2019) y Jasek et al. (2019); los clientes de una única compra no son tenidos en cuenta en este tipo de estudios, dado que no es posible identificar un patrón de compra a partir de sus registros transaccionales e inducen ruidos o sesgos a los análisis. Aplicando estas reglas, se tiene que, de los 4.443 clientes, 2.990 cumplen, es decir el 67,29% de los clientes.

Finalmente, los clientes son separados aleatoriamente en 2 grupos, con igual cantidad de clientes cada uno. Un primer grupo con los cuales se desarrollará el Capítulo 5 de esta tesis “Selección de métodos para la predicción del valor monetario” y un segundo grupo con los cuales se desarrollará el Capítulo 6 “Método para la segmentación de clientes incorporando la predicción del valor monetario”; de esta manera, mitigamos el riesgo de un posible sesgo en los resultados, del mismo modo, esta partición contribuye a validar los resultados de la predicción del valor monetario en productivo. El esquema de la Figura 4-6 representa esta partición.

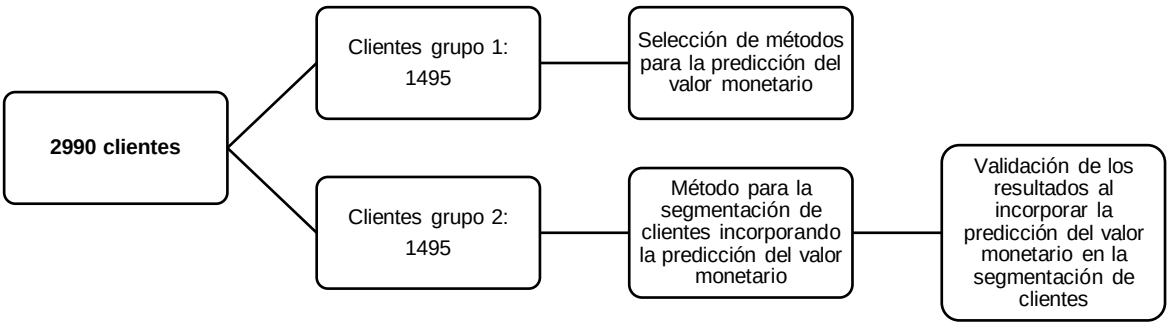

Figura 4-6: Esquema de partición de los datos por grupos
Fuente: elaboración propia

4.4 Conclusiones

Se da cumplimiento al objetivo planteado en el presente capítulo, en el que se logra convertir los datos transaccionales en datos a nivel de clientes, lo cual es el insumo principal para los capítulos siguientes de la presente tesis. En este punto, se destaca la relevancia de este paso, dado que representa pate fundamental para realizar los procesos de modelado, tanto en la predicción del valor monetario, como en la segmentación de clientes, que se realiza entre los capítulos 5 y 7; puesto que estos análisis se realizan a nivel de clientes y no de cada transacción. Así mismo, con la preparación y alistamiento de datos realizado, se espera mitigar los riesgos de generación de sesgos en los resultados, ocasionados por inadecuada calidad en los datos.

5. Selección de métodos para la predicción del valor monetario

5.1 Introducción

En este capítulo se da cumplimiento al objetivo específico 3 de la presente tesis, en el cual se deben seleccionar el método a utilizar para la predicción del valor monetario. Siguiendo a Win & Bo (2020), el valor monetario será calculado como lo muestra la Ecuación 5-1. Donde *Customer Value,* corresponde al valor monetario del cliente, *Average Order Value,* corresponde al valor promedio de cada compra del cliente y *Purchase Frequency,* representa el promedio de frecuencia de compras o cantidad de compras por cliente.

$$\textit{Customer Value} = Average\ Order\ Value \times Purchase\ Frequency \quad (5\text{-}1)$$

Los modelos paramétricos predicen ambas variables por separado y luego se multiplican, para lograr la predicción del valor monetario del cliente, mientras que los modelos de aprendizaje de máquinas predicen directamente el Valor Monetario.

En este capítulo se implementan técnicas de modelado paramétrico, así como técnicas de aprendizaje de máquinas, al final se selecciona el modelo que mejor predice el Valor Monetario.

5.2 Metodología

A partir de literatura, se propone realizar la predicción del valor monetario usando técnicas paramétricas y por otro lado utilizando técnicas de aprendizaje de máquinas, al final se selecciona el que da mejores resultados, los métodos a utilizar se ilustran en la Figura 5-1.

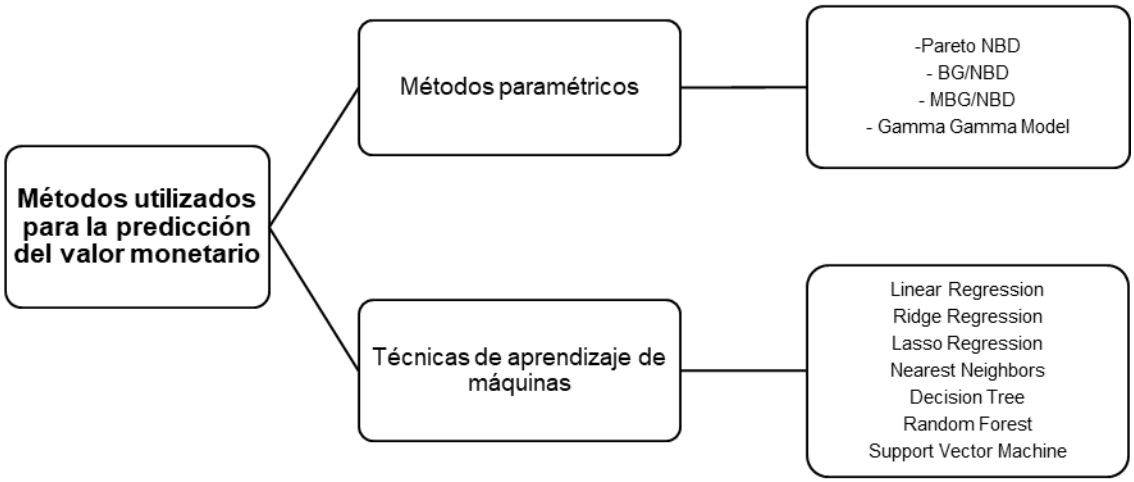

Figura 5-1: Metodología predicción valor monetario
Fuente: elaboración propia

5.3 Resultados

5.3.1 Selección e implementación de métodos paramétricos

En el caso de los métodos paramétricos, estos requieren de un conjunto de datos agregados a nivel de clientes, con la distinción de qué comportamiento de compras tienen los clientes en el periodo de ajuste o calibración y qué comportamiento tienen en el periodo de pronóstico o reserva (Platzer, 2021), es decir, entre el 1 de diciembre del 2009 y el 31 de mayo de 2011 para el periodo de ajuste y entre el 1 de junio de 2011 y el 9 de diciembre del 2011 para el periodo de pronóstico. Ahora, el conjunto de datos representado en la Tabla 4-14 del capítulo 4 es el principal insumo para la conformación del conjunto de datos requerido en esta sección. La descripción de las variables a utilizar en el conjunto de datos para esta sección se presenta en la Tabla 5-1.

Tabla 5-1: Conjunto de datos para métodos paramétricos

Variable	Tipo de variable	Descripción
ID_Cliente	Cualitativa nominal	Identificador único asignado a cada

Variable	Tipo de variable	Descripción
		cliente
Transacciones_periodo_ajuste	Cuantitativa discreta	Cantidad de transacciones o compras repetidas[4] que realizó el cliente en el periodo de ajuste o calibración
Recencia_periodo_ajuste	Cuantitativa discreta	Corresponde al tiempo entre la primera y última compra del cliente
Antigüedad_periodo_ajuste	Cuantitativa discreta	Es la antigüedad del cliente, representada en el tiempo entre la primera compra del cliente y el final del periodo de ajuste
Valor_monetario_periodo_ajuste	Cuantitativa continua	Es la suma de dinero o monto que el cliente ha gastado en el negocio en el periodo de ajuste
Frecuencia_periodo_pronóstico	Cuantitativa discreta	Es la cantidad de compras reales que realizó el cliente en el periodo de pronóstico o reserva
Valor_monetario_periodo_pronóstico	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el periodo de pronóstico o reserva

Fuente: elaboración propia

Siguiendo la metodología propuesta en este capítulo, el proceso de modelado se realiza con los clientes del grupo 1. En la Tabla 5-2 se presenta una muestra de los datos, donde las variables que tienen el término “ajuste”, corresponden a los datos del cliente en el periodo de ajuste o calibración, mientras que los datos que presentan el término “pronóstico”, corresponden a los datos reales del cliente en el periodo de pronóstico o reserva; estos últimos serán utilizados luego para calcular las métricas de error de los métodos paramétricos para la predicción de la frecuencia de compras esperada de los clientes.

[4] En la literatura, las compras o transacciones repetidas son la cantidad de periodos en los que el cliente hizo compras, por ejemplo, si el cliente hizo 2 compras, tiene un periodo o una compra repetida, si hizo 3, tiene 2. Es decir, la frecuencia en el periodo de ajuste es la cantidad de compras totales del cliente en ese periodo menos 1.

Tabla 5-2: Muestra del conjunto de datos para métodos paramétricos

ID_ Cliente	Transacciones_ periodo _ajuste	Recencia_ periodo _ajuste	Antigüedad_ periodo _ajuste	Valor_monetario _periodo _ajuste	Frecuencia _periodo _pronóstico	Valor_monetario _periodo _pronóstico
14896	7,000	529,000	541,000	253,774	0.000	0.000
14741	2,000	56,000	225,000	339,900	4.000	226,260
13209	7,000	457,000	492,000	674,720	3.000	1.176,390
16993	1,000	113,000	384,000	968,350	0.000	0,000
16175	4,000	359,000	371,000	244,045	2.000	285,305

Fuente: elaboración propia

Teniendo como entrada el conjunto de datos de la Tabla 5-2, se continúa con la implementación de los modelos paramétricos propuestos en la metodología para la predicción de la variable *"Predicción_Frecuencia"*, la cual indica la cantidad de compras que se estima realizará cada uno de los clientes en el periodo de pronóstico o de reserva. En ese sentido, los modelos a implementados son: Pareto NBD; BG/NBD y MBG/NBD, para los cuales, se ilustra en la Figura 5-2 el código utilizado para este fin.

```
penal_coef = 0.1

model_instances = [
    lifetimes.BetaGeoFitter(penalizer_coef=penal_coef),
    lifetimes.ParetoNBDFitter(penalizer_coef=penal_coef),
    lifetimes.ModifiedBetaGeoFitter(penalizer_coef=penal_coef),
]

model_names = [
    'BG/NBD',
    'Pareto/NBD',
    'MBG/NBD'
]

def model_fit(model, data, name):
    model.fit(data.frequency_cal, data.recency_cal, data.T_cal)
    print(f'Execute fit for model {name}')

    prediction = model.conditional_expected_number_of_purchases_up_to_time(
        data.duration_holdout.loc[0], data.frequency_cal, data.recency_cal, data.T_cal
    )
    print(f'Predict data for model {name}')

    mae = mean_absolute_error(data.frequency_holdout, prediction)
    rmse = np.sqrt(mean_squared_error(data.frequency_holdout, prediction))
    r2 = r2_score(data.frequency_holdout, prediction)
    print(f'Evaluate data for model {name}')

    return name, prediction.tolist(), mae, rmse, r2
```

Figura 5-2: Código para la implementación métodos paramétricos predicción frecuencia
Fuente: elaboración propia

Como resultados de la implementación de los modelos paramétricos, se encontró que el modelo BG/NBD logra las mejores métricas de error en la predicción de la "*Predicción_Frecuencia*", puesto que arroja el menor Error Absoluto Medio (MAE), la

menor Raíz del Error Cuadrático Medio (RMSE), y el mayor Coeficiente de determinación (R2). Estos resultados se presentan en la Tabla 5-3.

Tabla 5-3: Métricas de error del desempeño de los métodos paramétricos

Modelo	MAE	RMSE	R2
BG/NBD	1,324	2,233	0,702
Pareto/NBD	1,361	2,318	0,679
MBG/NBD	1,419	2,305	0,682

Fuente: elaboración propia

Ahora, para la predicción del monto promedio esperado de las transacciones o compras, se utiliza el modelo Gamma-Gamma, el cual predice el monto promedio esperado de las compras, es decir, independientemente de la cantidad de compras que realicen los clientes, qué valor monetario promedio se espera tenga cada una de ellas, a esta variable se le denominará “Monto_promedio_pronosticado”. En este sentido, se presenta en la Tabla 5-4 los coeficientes de resultado del modelo Gamma-Gamma, con los cuales se logra un mejor ajuste o predicción del monto promedio esperado en el periodo de pronóstico o reserva.

Tabla 5-4: Coeficientes de parametrización modelo Gamma-Gamma

Parámetro	coef	se(coef)	lower 95% bound	upper 95% bound
p	1,030119	0,041001	0,949757	1,11048
q	0,181031	0,005074	0,171086	0,190976
v	0,937948	0,042683	0,854289	1,021606

Fuente: elaboración propia

En este punto, se calcula el valor monetario pronosticado, a esta variable se le denominará “Predicción_valor_monetario”, y se calcula multiplicando la “*Predicción_frecuencia*” por el “Monto_promedio_esperado”, para cada cliente. En la Tabla 5-5 se muestran los resultados de las métricas y percentiles de las variables del conjunto de datos de salida, el cual incorpora la “Predicción_valor_monetario”.

Tabla 5-5: Métricas y percentiles resultados métodos paramétricos

Métrica	ID_Cliente	Transacciones_periodo_ajuste	Valor_monetario_periodo_ajuste	Frecuencia_periodo_pronóstico	Valor_monetario_periodo_pronóstico	Predicción_frecuencia	Predicción_valor_monetario
Conteo	374,000	374,000	374,000	374,000	374,000	374,000	374,000
Media	15626,444	5,652	330,666	2,182	236,477	2,316	695,398
Desviación estándar	1625,540	10,654	214,302	4,834	276,691	3,403	636,038
Mín	12346,000	1,000	11,356	0,000	0,000	0,000	12,245
25%	14229,500	1,000	181,425	0,000	0,000	0,756	324,055
50%	15671,000	3,000	286,708	1,000	173,836	1,534	480,719
75%	17046,500	6,000	416,480	3,000	365,490	2,690	836,971
Máx	18286,000	132,000	1465,380	71,000	1798,807	41,796	4809,475

Fuente: elaboración propia

Finalmente, en la Tabla 5-6, se presentan las métricas de error de los resultados de la implementación de los modelos paramétricos, los cuales serán luego comparados con los resultados generados por los métodos de aprendizaje de máquina para la predicción del valor monetario del cliente.

Tabla 5-6: Métricas de error predicción valor monetario métodos paramétricos

Modelo	MAE	RMSE	R2
BG/NBD y Gamma-Gamma	749,330	1312,435	0,419

Fuente: elaboración propia

5.3.2 Selección e implementación de métodos de aprendizaje de máquinas

Los modelos de aprendizaje de máquinas tienen la ventaja de que permiten predecir directamente el valor monetario, sin realizar previamente una predicción de la frecuencia, así mismo, permiten agregar nuevas variables al proceso de modelado. Ahora, los clientes y variable objetivo o dependiente, la cual es el valor monetario del cliente en el periodo de pronóstico, son exactamente iguales a los utilizados en la selección de métodos paramétricos. Sin embargo, partiendo de la flexibilidad que permiten los métodos de aprendizaje de máquinas, se incorporan nuevas variables independientes al conjunto de datos. En la Tabla 5-7, se describen las variables a usar en esta sección, donde una de las novedades es el uso de variables independientes que representan el valor monetario que tuvieron los clientes desde el mes 1 hasta el mes 18 del periodo de ajuste.

Tabla 5-7: Conjunto de datos selección de métodos aprendizaje de máquinas

Variable	Tipo de variable	Descripción
Id_Cliente	Cualitativa nominal	Identificador único asignado a cada cliente
Recencia_periodo_ajuste	Cuantitativa discreta	Tiempo transcurrido entre la última compra del cliente y la fecha final del periodo de ajuste
Antigüedad_periodo_ajuste	Cuantitativa discreta	Es la antigüedad del cliente, representada en el tiempo entre la primera compra del cliente y el final del periodo de ajuste
Transacciones_periodo_ajuste	Cuantitativa discreta	Cantidad de transacciones o compras repetidas[5] que realizó el cliente en el periodo de ajuste o calibración
Tiempo_activo	Cuantitativa discreta	Corresponde al tiempo entre la primera y última compra del cliente
Valor_monetario_mes1	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el mes 1 del periodo de ajuste
Valor_monetario_mes2	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el mes 2 del periodo de ajuste
Valor_monetario_mes3	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el mes 3

[5] En la literatura, las compras o transacciones repetidas son la cantidad de periodos en los que el cliente hizo compras, por ejemplo, si el cliente hizo 2 compras, tiene un periodo o una compra repetida, si hizo 3, tiene 2. Es decir, la frecuencia en el periodo de ajuste es la cantidad de compras totales del cliente en ese periodo menos 1.

Variable	Tipo de variable	Descripción
		del periodo de ajuste
.	.	.
.	.	.
.	.	.
Valor_monetario_mes17	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el mes 17 del periodo de ajuste
Valor_monetario_mes18	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el mes 18 del periodo de ajuste
Valor_monetario_periodo_pronóstico	Cuantitativa continua	Es el valor o monto real que el cliente depositó en la empresa en el periodo de pronóstico o reserva

Fuente: elaboración propia

En la Tabla 5-8 se relacionan diferentes métricas sobre el conjunto de datos, el cual está representado por los 1.495 clientes del anteriormente denominado “Grupo 1”, y 23 variables, dado que no se ilustra la variable “Id_Cliente”, al no ser cuantitativa.

Tabla 5-8: Métricas y percentiles conjunto de datos métodos de aprendizaje de máquinas

	Conteo	Media	Desviación Estándar	Mín	25%	50%	75%	Máx
Recencia_periodo_ajuste	1495	123,251	110,581	0,000	25,000	88,000	198,000	534,000
Antiguedad_periodo_ajuste	1495	409,068	132,059	13,000	330,000	450,000	531,000	546,000
Transacciones_periodo_ajuste	1495	6,474	8,660	2,000	2,000	4,000	7,000	133,000
Tiempo_activo	1495	103,155	77,284	1,000	49,000	82,000	136,500	540,000
Valor_monetario_mes1	1495	99,033	210,054	0,000	0,000	0,000	122,645	1794,700
Valor_monetario_mes2	1495	77,842	206,859	0,000	0,000	0,000	0,000	2521,600
Valor_monetario_mes3	1495	86,999	226,230	0,000	0,000	0,000	0,000	2491,780
Valor_monetario_mes4	1495	110,918	225,698	0,000	0,000	0,000	143,345	1787,900
Valor_monetario_mes5	1495	104,818	249,868	0,000	0,000	0,000	109,025	3181,970
Valor_monetario_mes6	1495	99,590	225,505	0,000	0,000	0,000	105,175	2220,000
Valor_monetario_mes7	1495	108,219	238,800	0,000	0,000	0,000	139,700	2199,390
Valor_monetario_mes8	1495	92,444	220,073	0,000	0,000	0,000	66,360	3113,610

	Conteo	Media	Desviación Estándar	Mín	25%	50%	75%	Máx
Valor_monetario_mes9	1495	98,736	246,419	0,000	0,000	0,000	19,425	3894,640
Valor_monetario_mes10	1495	138,617	297,014	0,000	0,000	0,000	184,075	3123,270
Valor_monetario_mes11	1495	161,973	275,655	0,000	0,000	0,000	287,755	1838,950
Valor_monetario_mes12	1495	161,297	269,300	0,000	0,000	0,000	265,138	2648,250
Valor_monetario_mes13	1495	135,762	320,560	0,000	0,000	0,000	117,250	3377,830
Valor_monetario_mes14	1495	69,280	186,776	0,000	0,000	0,000	0,000	2616,580
Valor_monetario_mes15	1495	72,457	195,753	0,000	0,000	0,000	0,000	2346,340
Valor_monetario_mes16	1495	87,238	221,355	0,000	0,000	0,000	0,000	2737,240
Valor_monetario_mes17	1495	83,554	209,224	0,000	0,000	0,000	0,000	2530,020
Valor_monetario_mes18	1495	104,685	224,094	0,000	0,000	0,000	131,400	2260,400
Valor_monetario_periodo_pronóstico	1495	856,086	1850,141	0,000	0,000	312,950	973,890	29469,490

Fuente: elaboración propia

Ahora, se realiza un análisis de correlaciones, para identificar si existe alguna dependencia estadística entre las variables, para ello se utilizó el coeficiene correlación, no paramétrico, de Spearman (Roy-García et al., 2019). En la Figura 5-3, se presenta la matriz de correlaciones con los resultados obtenidos, donde no se evidencia la presencia de variables altamente correlacionadas, en ninguno de los dos sentidos, a saber, directamente o inversamente correlacionadas.

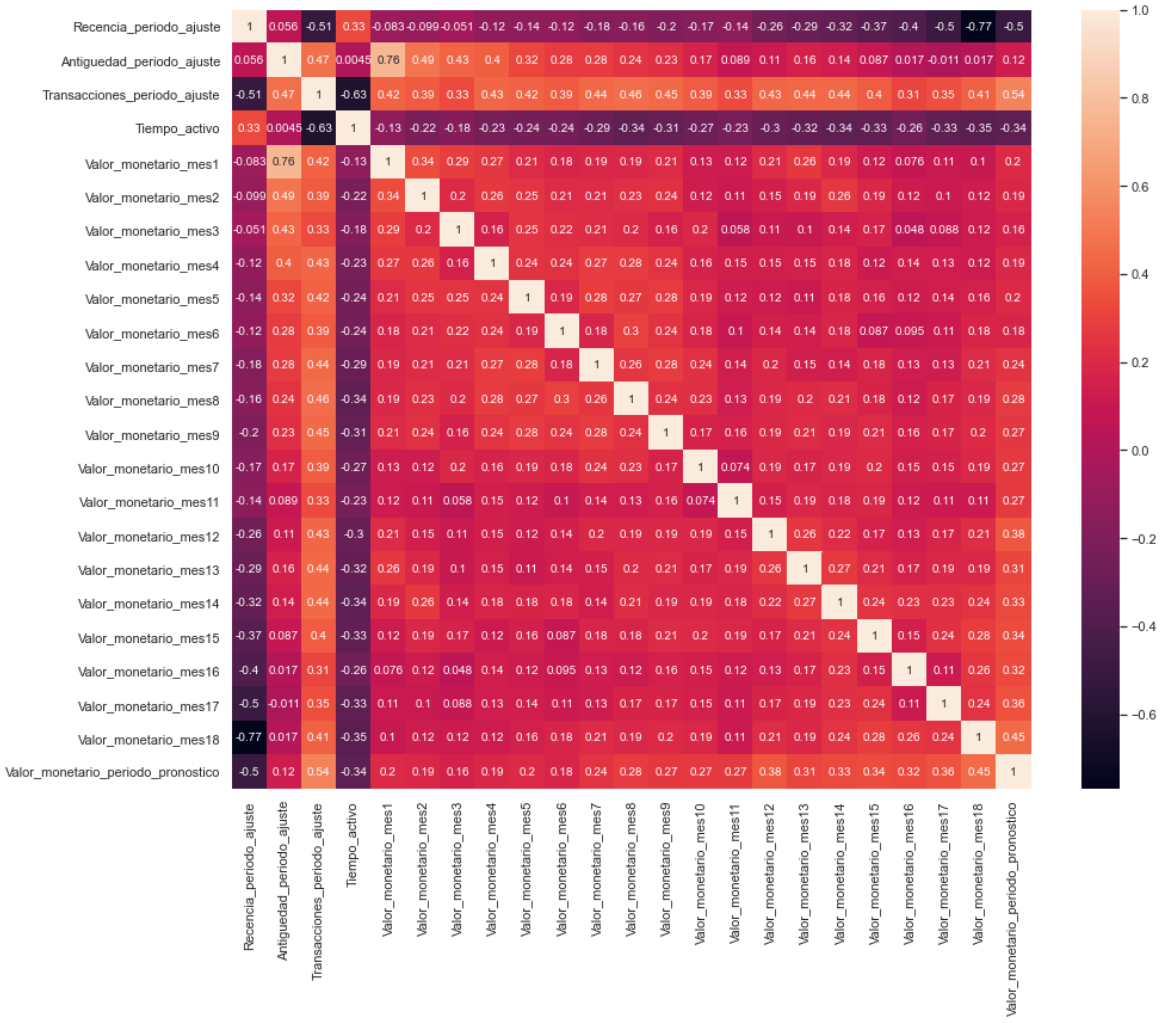

Figura 5-3: Matriz de correlaciones conjunto de datos métodos de aprendizaje de máquinas. Fuente: elaboración propia.

En esta etapa, se procede a dar inicio al proceso de modelado, no obstante, primero se realiza el escalado de los datos, para lo cual se testearon 4 escaladores, la elección del mejor, se seleccionó mediante el uso de 3 algoritmos sensibles a la presencia de variables no escaladas, en este caso, *LinearRegression*, *KneighborsRegressor y RandomForest* y finalmente el criterio de decisión fue el Error Absoluto Medio (MAE); es así como los resultados obtenidos son los siguientes: *StandardScaler* con el mejor *MAE* de 509,566 en *KneighborsRegressor, MinMaxScaler* con el mejor *MAE* de 515,102 en *KneighborsRegressor, MaxAbsScaler* con el mejor *MAE* de *514,560* en *KneighborsRegressor* y finalmente *RobustScaler* con el mejor *MAE de 527,190* en *RandomForestRegressor*. Es así, como el mejor MAE se logra en StantardScaler, por lo tanto, es el escalador seleccionado para el conjunto de datos.

Siguiendo lo con la metodología propuesta, el paso siguiente consiste en predecir el Valor monetario en el periodo de pronóstico, representado por la variable "Valor_monetario_periodo_pronóstico". Se definió dividir el conjunto de datos en 75% para modelado y 25% para testeo, utilizando un *random_state* de 42; con esto, se implementaron 7 técnicas de aprendizaje de máquinas y 3 métricas de error para la elección del mejor modelo. Así mismo, se realizó un tuneo iterativo de hiperparámetros, para lograr un mejor desempeño en cada modelo; con lo anterior, los resultados de este proceso se presentan en la Tabla 5-9.

Tabla 5-9: Resultados implementación de modelos de aprendizaje de máquinas

Modelo	Mejores parámetros	MAE	RMSE	R2
Estimator				
Linear Regression		592,371	1255,323	0,468
Ridge Regression		591,670	1253,757	0,470
Lasso Regression		590,456	1252,966	0,470
Nearest Neighbors	n_neighbors=10, weights='distance'	510,397	1093,824	0,596
Decision Tree	max_depth=10	716,776	1690,701	0,036
Random Forest	max_features=None	534,615	1371,611	0,365
Support Vector Machine	C=2.0	544,004	1355,345	0,380

Fuente: elaboración propia

Dado los resultados expresados en la Tabla 5-9, se tiene que *Nearest Neighbors* logra los mejores resultados, dado que tiene las mejores métricas de error en las 3 variables, el menor MAE con un valor de 510,397; el menor RMSE con un valor de 1093,824; y el mayor R2 con un valor de 0,59; por tanto, se selecciona este modelo.

Finalmente, en la Figura 5-4, se presentan los gráficos de densidad de la variable dependiente, tanto en el valor real del valor monetario en el periodo de pronóstico, como de la predicción del valor monetario en el periodo de pronóstico, donde, se observa la presencia de valores alejados en los datos reales, que no se predicen a tal nivel en la predicción del valor; sin embargo, en términos generales, se observa un comportamiento similar en ambos gráficos.

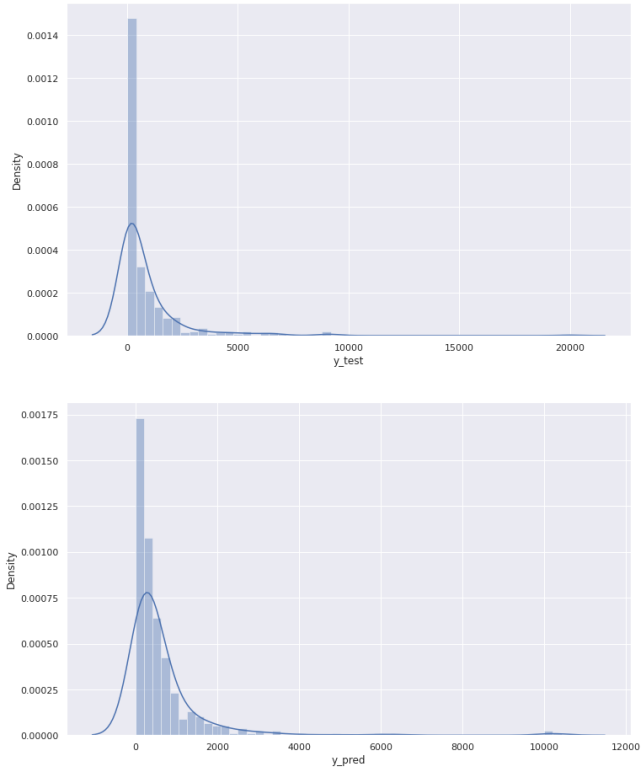

Figura 5-4: Gráfico de densidades del valor monetario real versus la predicción.
Fuente: elaboración propia

5.3.3 Comparación y selección de método para la predicción del valor monetario

En esta sección, se hace un comparativo entre los resultados de las métricas de error del mejor modelo de aprendizaje de máquinas versus el mejor modelo paramétrico o estadístico. En este punto, se hace necesario recordar que, la forma en la que funcionan los modelos paramétricos, permiten la predicción del valor monetario para todos los clientes, dado que la partición de ajuste y predicción se hace por fecha, para todos los cliente, sin embargo, los modelos de aprendizaje de máquinas, requieren dejar una proporción de los datos (clientes), para realizar el testeo de la predicción; en ese sentido, para comparar los resultados de ambos modelos, se tomará la misma muestra de clientes, tanto del modelo paramétrico seleccionado, como del modelo de aprendizaje de máquinas seleccionado. Esta proporción corresponde al 25% de los clientes, seleccionados aleatoriamente y que fueron usados en el modelado de aprendizaje de máquinas para el testeo, lo cual corresponde a 374 clientes.

En la Tabla 5-10 se muestran los resultados de las métricas de error del modelo *Nearest Neighbors* o Vecinos más cercanos y del modelo paramétrico, que tiene implícito un BG/NBD y un Gamma-Gamma model. Se evidencia que, los resultados usando el modelo de aprendizaje de máquinas, logra mejores métricas de error en todos los 3 criterios (MAE, RMSE y R2). En consecuencia, la selección final de método para la predicción del valor monetario es *Nearest Neghbors*.

Tabla 5-10: Resultados métricas de error selección método predicción valor monetario

Estimator	MAE	RMSE	R2
Nearest Neighbors	510,397	1093,824	0,596
Gamma-Gamma	749,330	1312,435	0,419

Fuente: elaboración propia

5.4 Conclusiones

Se concluye que el método de aprendizaje de máquinas Nearest Neighbors, con la estructuración de los datos propuestos, logra un mejor desempeño que los modelos paramétricos tradicionales. Aquí se identifica un aporte relevante a la literatura, puesto que aún son pocos los autores, en la literatura formal o científica, que han abordado la predicción del valor monetario del cliente, desde el enfoque de aprendizaje de máquinas, tales como Bauer & Jannach (2021) y Chen et al. (2019); principalmente porque los modelos tradicionales o paramétricos logran, en términos generales, buenos desempeños.

En ese sentido, con la implementación realizada en este capítulo, se aporta evidencia empírica a favor de la línea investigativa que propone abordar los problemas de predicción del valor monetario del cliente desde aplicando técnicas de aprendizaje de máquinas.

6. Método para la segmentación de clientes incorporando la predicción del valor monetario

6.1 Introducción

En este capítulo, se incorpora la predicción del valor monetario al proceso de segmentación de clientes. Se divide en 2: primero se hace la segmentación de clientes sin incorporar la predicción del valor monetario y luego la segmentación incorporando la predicción. Este capítulo da cumplimiento al objetivo específico 3 de la presente tesis.

Este capítulo se desarrolla con el conjunto de datos que, en el capítulo 2, fue denominado como “grupo 2”, que conforma un conjunto de datos de clientes que no fue utilizado para el proceso de modelado en la predicción del valor monetario del cliente; lo que permite minimizar sesgos en los datos.

Los algoritmos de segmentación utilizados en este capítulo son:

- KMeans
- Agglomerative Clustering
- GaussianMixture
- Birch

6.2 Metodología

La metodología propuesta, se divide en 2: primero se hace la segmentación de clientes sin incorporar la predicción del valor monetario y luego la segmentación incorporando la predicción, en la Figura 6-1, se presenta un esquema de la metodología propuesta.

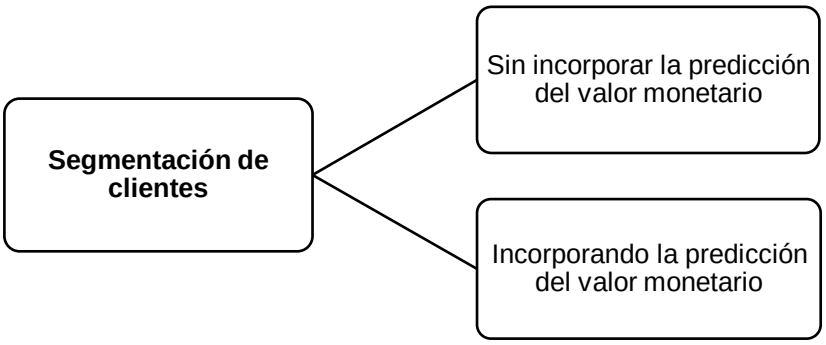

Figura 6-1: Metodología propuesta para la segmentación de clientes
Fuente: elaboración propia

6.3 Resultados

6.3.1 Segmentación sin incorporar la predicción del valor monetario

El primer punto, consiste en la realización de un proceso de segmentación de clientes sin incorporar la predicción del valor monetario que, a partir de la literatura y metodología propuesta, se realiza con las variables usadas en análisis tipo Recencia, Frecuencia, Monto (RFM) (Yoseph & Heikkila, 2019). En este sentido, se muestra en la Tabla 6-1 las métricas de resumen y percentiles de los datos, donde se verifica que el conjunto de datos corresponde a 1.495 clientes, y cada variable está asociada al comportamiento de compra de los clientes durante el periodo de ajuste, a saber, desde el 1 de diciembre de 2009 hasta el 31 de mayo de 2011.

Tabla 6-1: Métricas y percentiles conjunto de datos segmentación sin incorporar predicción del valor monetario

	Recencia	**Frecuencia**	**Valor_monetario**
Conteo	1495,000	1495,000	1495,000
Media	127,588	5,152	2098,904
Desviación estándar	114,382	7,690	5230,735
Mín	0,000	1,000	8,500
1%	0,000	1,000	43,466
5%	7,000	1,000	113,500

	Recencia	Frecuencia	Valor_monetario
10%	11,000	1,000	165,192
25%	26,000	1,000	325,040
50%	92,000	3,000	852,310
75%	202,000	6,000	2000,400
90%	284,000	11,000	4450,756
95%	355,000	16,000	6889,334
99%	455,120	34,120	21677,267
Máx	544,000	146,000	93842,050

Fuente: elaboración propia

Luego, avanzando en un análisis exploratorio, en la Figura 6-1 se presentan los gráficos de dispersión entre las variables de Recencia, Frecuencia y Valor_monetario, donde visualmente se encuentran indicios de una posible correlación entre la Frecuencia y el Valor_monetario; que al ser validada con coeficiente de correlación de Spearman, se halla que es del 0,87. Por lo que se define calcular el valor monetario promedio de cada cliente, el cual es el resultado de la división entre el Valor_monetario y la Frecuencia de compra; con esto, se presenta en la Figura 6-2 la matriz de correlaciones resultante, utilizando coeficiente de Spearman (Roy-García et al., 2019); en el que no se evidencia la presencia de variables correlacionadas.

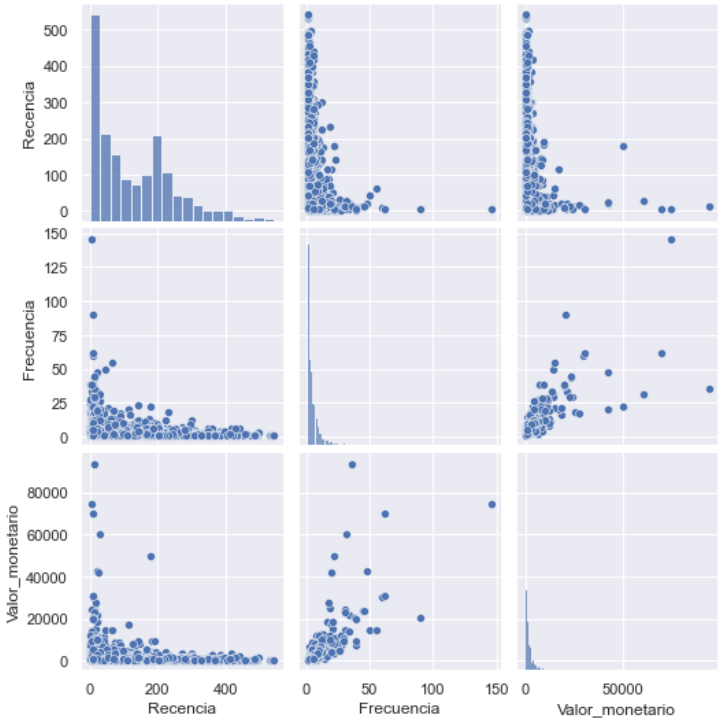

Figura 6-2: Dispersión de los datos en variables RFM
Fuente: elaboración propia

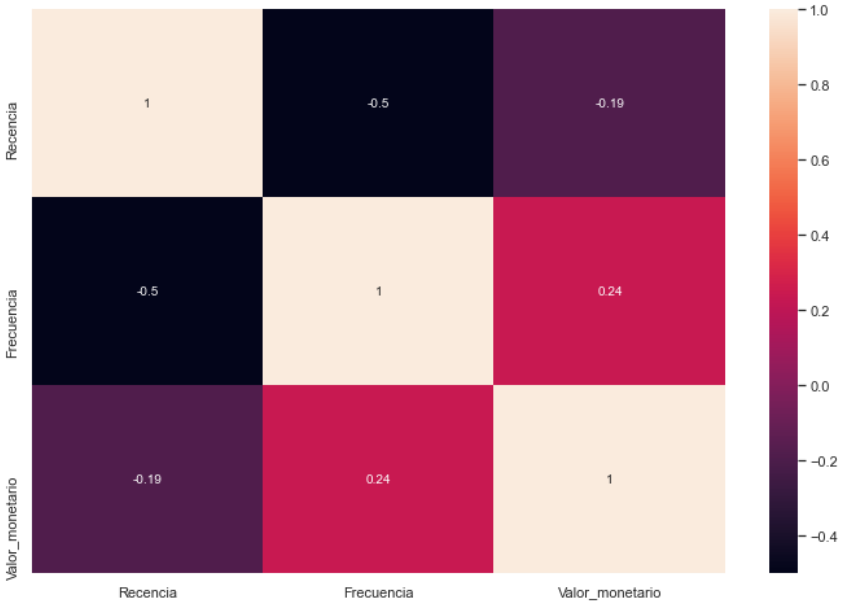

Figura 6-3: Matriz de correlaciones variables RFM
Fuente: elaboración propia

En la Figura 6-3 se visualiza la dispersión de los datos, entre las variables Valor_monetario y Frecuencia, donde se observa la presencia de valores alejados o atípicos. Ahora, dado que el objetivo principal de este capítulo es realizar una

segmentación de clientes, los clientes con valores atípicos no pueden ser eliminados, sin embargo, mantenerlos dentro del conjunto global de los datos, afectaría los resultados, dado que diferentes algoritmos a utilizar en este capítulo son sensibles a la presencia de datos atípicos; por lo que se propone la implementación de un algoritmo de detección de atípicos para dividir el conjunto de datos en datos normales *(inliers)* y datos atípicos *(outliers);* y posteriormente implementar algoritmos de segmentación de ambos grupos, para finalmente unificarlos.

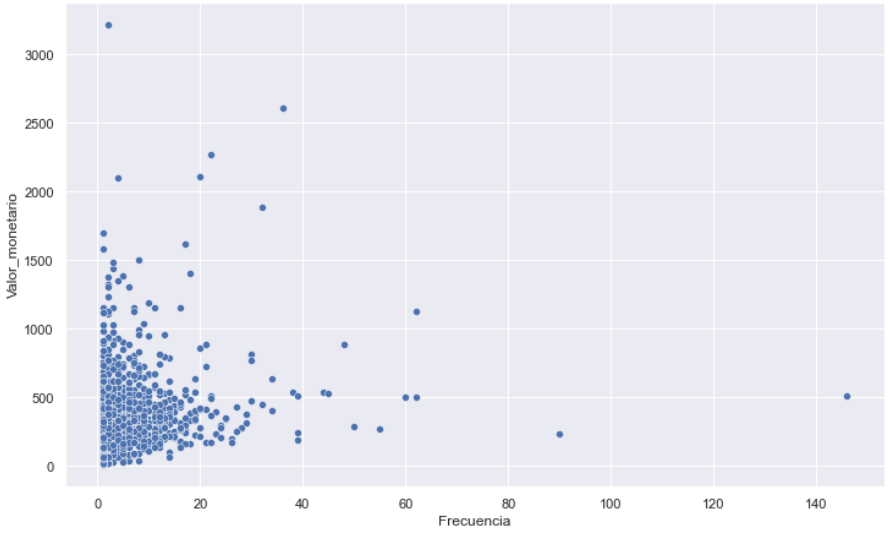

Figura 6-4: Dispersión de los datos valor monetario promedio versus frecuencia
Fuente: elaboración propia

La identificación de atípicos se realiza utilizando el algoritmo de *Isolation Forest* (Liu et al., 2008), con un factor de contaminación automático (12,9% de los datos); el escalado de los datos se realiza con *MaxAbsScaler* (Pedregosa et al., 2011). En consecuencia, se obtuvo como resultado 193 clientes con datos atípicos y 1302 con datos normales, y se representan en la Figura 6-4.

Luego, se realiza la segmentación de los datos en 2 momentos: en un primer momento se segmentan los datos del grupo marcado como atípicos y en un segundo momento se segmentan los clientes del grupo marcado como normales. Se presenta en la Tabla 6-2, los resultados de la segmentación de los clientes del conjunto de datos marcado como atípicos. Donde, a partir de las métricas seleccionadas, los criterios son: a menor *Davies_Bouldin*, mejor; a mayor *Calinski_Harabasz* mejor; y a mayor *Silhouette;* en este sentido, K-means con 2 grupos, logra los mejores resultados.

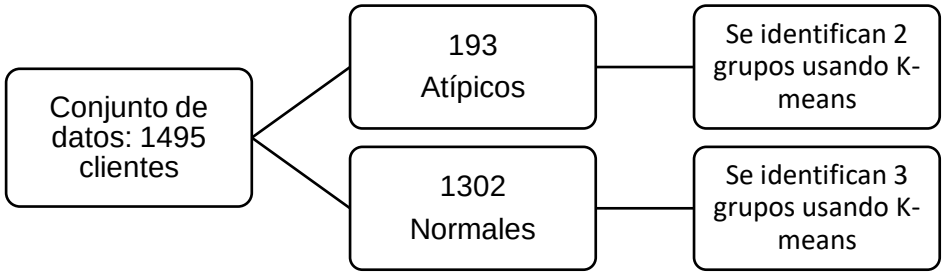

Figura 6-5: Esquema de segmentación de los datos
Fuente: elaboración propia

Tabla 6-2: Resultados segmentación de clientes en conjunto de datos atípicos

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
KMeans	0,540	402,332	0,599	[2, 2, 2]
AgglomerativeClustering	0,541	343,247	0,569	[2, 2, 2]
GaussianMixture	0,633	328,982	0,546	[2, 2, 2]
Birch	0,566	272,514	0,579	[3, 3, 3]

Fuente: elaboración propia

Seguido, se procede a implementar los métodos de agrupamiento al grupo de clientes marcado como normales, con lo que se logran los resultados mostrados en la Tabla 6-3, en la que se observa que no hay un único método ganador en las 3 métricas, Birch arroja mejores resultados en *Davies_Bouldin y Silhouette,* mientras que K-means ahorra mejores resultados en *Calinski_Harabasz.* En este punto, se decide la elección de K-means, utilizando como criterio adicional la ya elección de K-means en el grupo de atípicos, posibilitando así una mayor homogeneidad en el método seleccionado.

Tabla 6-3: Resultados segmentación de clientes en conjunto de datos normales

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
KMeans	1,1280	822,7183	0,3499	[3, 3, 3]
AgglomerativeClustering	1,1943	696,2101	0,3180	[3, 3, 3]
GaussianMixture	1,6851	421,4880	0,2009	[3, 3, 3]
Birch	1,0243	755,3470	0,3578	[3, 3, 3]

Fuente: elaboración propia

Ahora, se procede a agregar los grupos conformados en un mismo resultado, y se presenta en la Figura 6-5, el gráfico de dispersión de los datos, teniendo en cuenta las variables Recencia, Frecuencia y Valor_monetario.

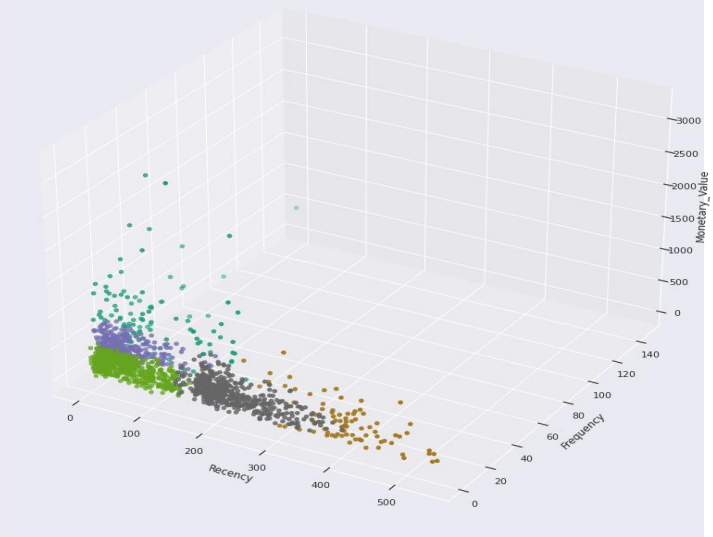

Figura 6-6: Dispersión de datos segmentos unificados sin predicción de valor monetario
Fuente: elaboración propia

Se presenta en la Figura 6-6 los gráficos de barra que representan el tamaño de cada uno de los grupos conformados. Donde se evidencia que el grupo 5 es el de mayor tamaño con 502 clientes, seguido del grupo 3 con 492, en tercera posición el grupo 2 con 308, luego el grupo 1 con 114 y el grupo 4 el más pequeño con 79 clientes.

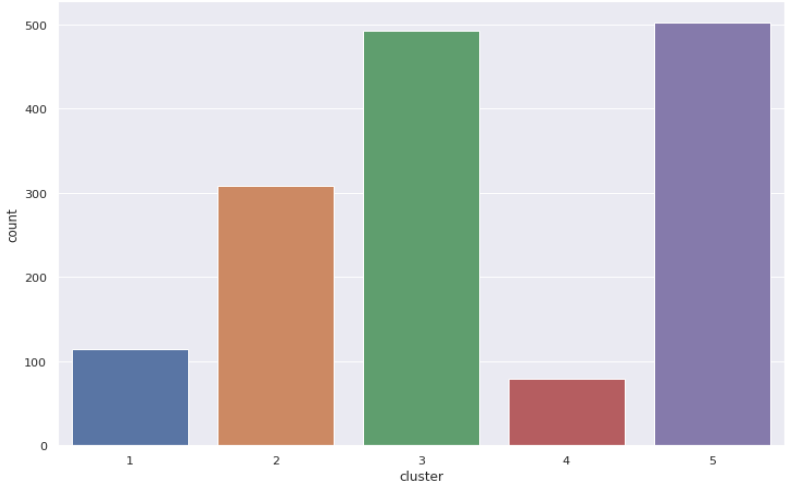

Figura 6-7: Gráfico de barras total de clientes por grupos o clúster sin predicción de valor monetario

Fuente: elaboración propia

Finalmente, se presenta en la Tabla 6-4, el resumen del resultado final de todos los 5 clústers conformados, según las 3 variables tenidas en cuenta en este análisis. Aplicando la regla de que, a menor Recencia, mejor, y que a mayor Frecuencia y Valor_monetario, mejor; se podría concluir que el Clúster 1 es el mejor grupo.

Tabla 6-4: Tabla resumen final segmentación sin incorporar predicción valor monetario

		Recencia			Frecuencia			Valor_monetario		
Clúster	**Conteo**	**Mín**	**Media**	**Máx**	**Mín**	**Media**	**Máx**	**Mín**	**Media**	**Máx**
1	114	0	65,219	240	1	19,088	146	30,802	809,879	3219,905
2	308	0	41,880	193	1	8,562	20	134,974	451,200	846,576
3	492	0	53,622	145	1	3,185	10	8,500	239,228	550,520
4	79	232	402,380	544	1	2,342	18	10,080	308,524	1149,160
5	502	137	223,586	399	1	2,265	9	10,900	269,275	811,100

Fuente: elaboración propia

6.3.2 Segmentación de clientes incorporando la predicción del valor monetario

El primer punto, consiste en la realización de un proceso de segmentación de clientes incorporando la variable de predicción del valor monetario, y se mantienen las variables de Recencia, Frecuencia y Valor_monetario, donde se precisa que este último valor monetario es el histórico que han presentado los clientes hasta el 31 de mayo de 2011, mientras que la predicción del valor monetario es el pronóstico del valor que aportaría el cliente entre el 1 de junio de 2011 y el 9 de diciembre de 2011.

En esta sección, se implementa en productivo la salida del modelo de *Nearest Neighbors*, que fue seleccionado en el Capítulo 5, como mejor modelo para la predicción del valor monetario del cliente, en ese sentido, el modelo se pone en productivo con los 1.495 clientes del llamado “Grupo 2”, y que el modelo no conoció en su etapa de ajuste. En consecuencia, se muestra en la Tabla 6-5 las métricas de resumen y percentiles de los datos, donde se verifica que el conjunto de datos corresponde a 1.495 clientes, los cuales son los mismos clientes con los que abordó la sección 6.3.1 y cada variable está asociada al comportamiento de compra de los clientes durante el periodo de ajuste, a saber, desde el 1 de diciembre de 2009 hasta el 31 de mayo de 2011.

Tabla 6-5: Métricas y percentiles de conjunto de datos segmentación incorporando predicción de valor monetario

Métrica y percentiles	Recencia	Frecuencia	Valor monetario	Predicción valor monetario
Conteo	1495,000	1495,000	1495,000	1495,000
Media	127,588	5,152	2098,904	622,134
Desviación estándar	114,382	7,690	5230,735	967,822
Mín	0,000	1,000	8,500	0,000
1%	0,000	1,000	43,466	0,000
5%	7,000	1,000	113,500	9,838
10%	11,000	1,000	165,192	50,127
25%	26,000	1,000	325,040	151,014
50%	92,000	3,000	852,310	325,644
75%	202,000	6,000	2000,400	701,219
90%	284,000	11,000	4450,756	1443,418
95%	355,000	16,000	6889,334	2141,456
99%	455,120	34,120	21677,267	5000,817
Máx	544,000	146,000	93842,050	12709,860

Fuente: elaboración propia

Se presenta en la Figura 6-8 la matriz de correlaciones, utilizando coeficiente de Spearman, en este punto, se recuerda que el valor monetario del periodo de ajuste presentaba una correlación con la variable Frecuencia, por lo que el valor monetario del periodo de ajuste que se relaciona en esta sección es el promedio del valor monetario. En la Figura 6-8 no se evidencia la presencia de correlaciones entre las variables.

En la Figura 6-9, se presenta la dispersión de los datos, en las variables de Valor_monetario y Frecuencia, donde se identifica nuevamente la presencia de datos atípicos o lejanos. Ahora, dado que el objetivo principal de este capítulo es realizar una segmentación de clientes, los clientes con valores atípicos no pueden ser eliminados, sin embargo, mantenerlos dentro del conjunto global de los datos, afectaría los resultados, dado que diferentes algoritmos a utilizar en este capítulo son sensibles a la presencia de datos atípicos; por lo que se propone la implementación de un algoritmo de detección de atípicos para dividir el conjunto de datos en datos normales *(inliers)* y datos atípicos *(outliers);* y posteriormente implementar algoritmos de segmentación de ambos grupos, para finalmente unificarlos.

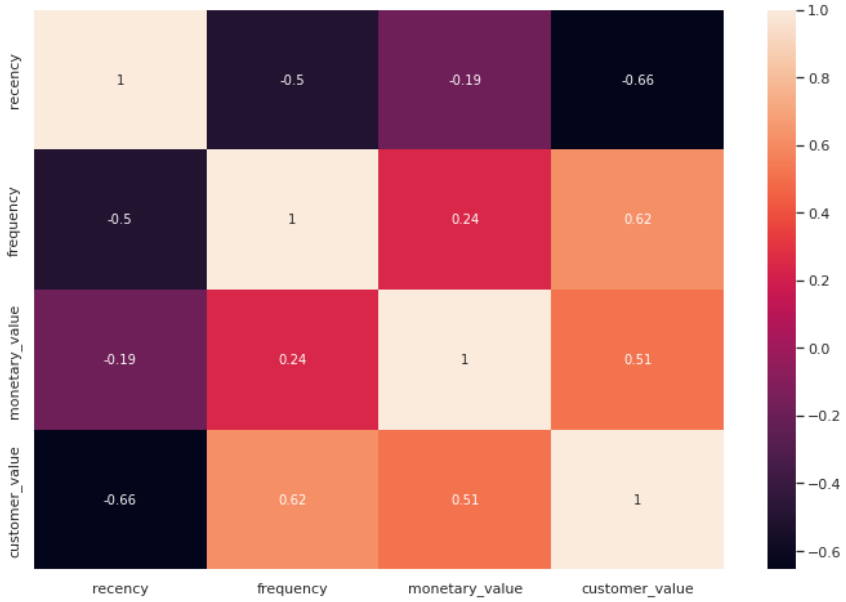

Figura 6-8: Matriz de correlaciones conjunto de datos con predicción de valor monetario
Fuente: elaboración propia

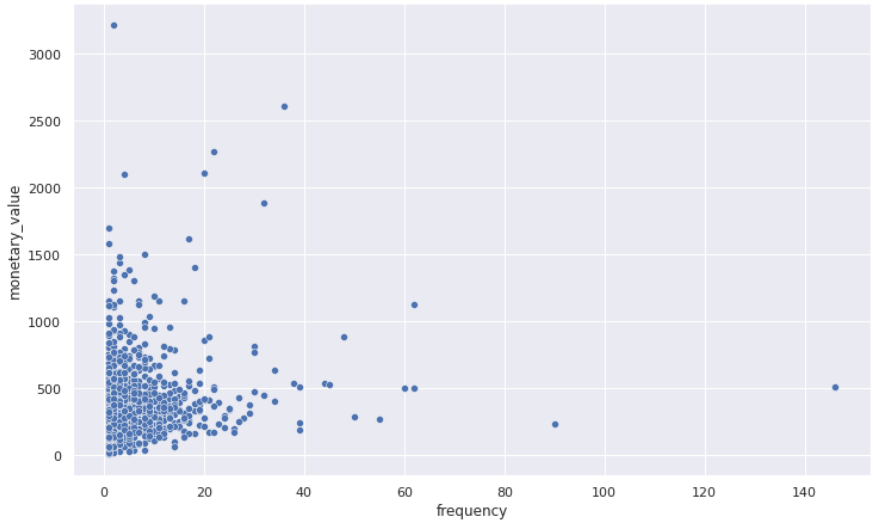

Figura 6-9: Dispersión de los datos incorporando la variable predicción de valor monetario
Fuente: elaboración propia

La identificación de atípicos se realiza utilizando el algoritmo de *Isolation Forest* (Liu et al., 2008), con un factor de contaminación automático (11,8% de los datos); el escalado de los datos se realiza con *MaxAbsScaler* (Pedregosa et al., 2011). En consecuencia, se obtuvo como resultado 177 clientes con datos atípicos y 1318 con datos normales, y se representan en la Figura 6-10.

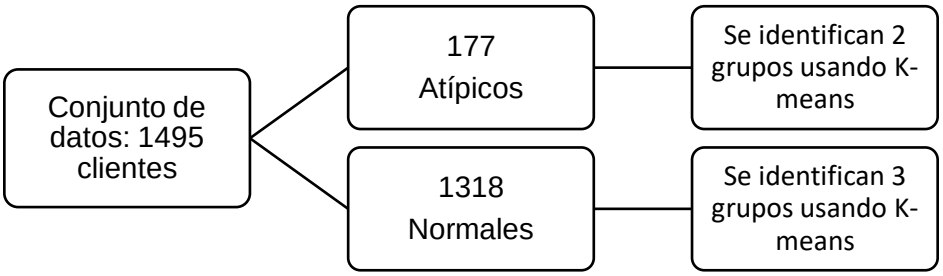

Figura 6-10: Esquema de distribución del conjunto de datos en atípicos y normales
Fuente: elaboración propia

Luego, se realiza la segmentación de los datos en 2 momentos: en un primer momento se segmentan los datos del grupo marcado como atípicos y en un segundo momento se segmentan los clientes del grupo marcado como normales. Se presenta en la Tabla 6-6, los resultados de la segmentación de los clientes del conjunto de datos marcado como atípicos. Donde, a partir de las métricas seleccionadas, los criterios son: a menor *Davies_Bouldin*, mejor; a mayor *Calinski_Harabasz* mejor; y a mayor *Silhouette;* en este sentido, K-means con 2 grupos, logra los mejores resultados.

Tabla 6-6: Resultados métodos de agrupamiento para datos atípicos incorporando predicción del valor monetario

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
KMeans	0,615048	253,271558	0,55161	[2, 2, 2]
AgglomerativeClustering	0,563484	238,430163	0,551627	[2, 2, 2]
GaussianMixture	0,793891	187,130401	0,475656	[2, 2, 2]
Birch	0,542984	173,179162	0,515283	[3, 3, 3]

Fuente: elaboración propia

Seguido, se procede a implementar los métodos de agrupamiento al grupo de clientes marcado como normales, con lo que se logran los resultados mostrados en la Tabla 6-7, en la que se observa que K-means con 3 clústeres, logra los mejores resultados en todas las métricas.

Tabla 6-7: Resultados métodos de agrupamiento para datos normales incorporando predicción del valor monetario

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
KMeans	1,100	804,203	0,337	[3, 3, 3]

Método de agrupamiento	Davies_Bouldin	Calinski_Harabasz	Silhouette	n_grupos
AgglomerativeClustering	1,229	643,002	0,304	[3, 3, 3]
GaussianMixture	1,562	469,024	0,241	[3, 3, 3]
Birch	1,044	667,991	0,335	[3, 3, 3]

Fuente: elaboración propia

En la Figura 6-10 se muestran los resultados agregados o unificados de los clústeres generados, donde se identifican los 5 grupos conformados. Y finalmente en la Tabla 6-8 muestra el resumen de los grupos conformados, donde se identifica que el clúster 1 es el mejor, según los criterios de menor Recencia, mayor Frecuencia, mayor Valor_monetario y mayor Predicción_valor_monetario.

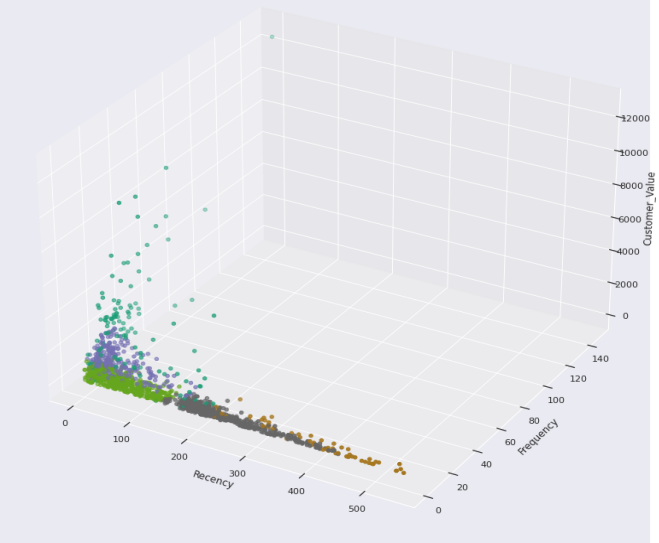

Figura 6-11: Dispersión de los datos por clústers incorporando predicción del valor monetario
Fuente: elaboración propia

Tabla 6-8: Resultado final de segmentación incorporando predicción de valor monetario

		Recencia			Frecuencia			Valor monetario			Predicción valor monetario	
Clúster	**Conteo**	**Mín**	**Media**	**Máx**	**Mín**	**Media**	**Máx**	**Mín**	**Media**	**Máx**	**Mín**	**Media**	**Máx**
1	121	0	47,421	225	1	18,950	146	162,143	796,199	3219,905	44,976	2785,809	12709,860
2	247	0	40,939	176	1	8,672	22	135,338	453,025	805,573	286,999	1156,554	2463,184
3	543	0	54,871	146	1	3,569	14	8,500	250,811	604,000	0,000	370,518	1249,368
4	56	228	390,982	544	1	2,643	18	13,500	444,446	1581,520	0,000	176,510	745,335
5	528	138	233,341	436	1	2,237	12	10,080	263,691	814,200	0,000	182,315	778,602

Fuente: elaboración propia

6.4 Conclusiones

Se logra el cumplimiento del objetivo, donde se logra hacer la implementación de la segmentación de clientes sin incorporar la predicción del valor monetario y la segmentación de clientes incorporando la predicción del valor monetario, con el fin de identificar qué método logra mejores resultados en el proceso de segmentación de clientes. Los grupos conformados en cada clúster de este capítulo son el insumo principal para la validación de los resultados y del método propuesto en el capítulo 7.

Se encuentra que, el algoritmo de K-means logra buenos resultados o desempeños, en casi todos los casos, en comparación con los demás algoritmos propuestos, tanto en la predicción sin incorporar la predicción del valor monetario, como en la predicción incorporando la predicción del valor monetario. Se destaca en este capítulo la definición de que en el proceso de limpieza no se pueden eliminar clientes que tengan un comportamiento aparentemente atípico, por el contrario, se propone el uso de algoritmos de detección de atípicos para identificar estos clientes y posteriormente segmentarlos; más no eliminarlos.

Finalmente, con el desarrollo de este capítulo, se contribuye en uno de los vacíos identificados en la revisión de la literatura, en el sentido que muestra de qué manera se puede realizar el proceso de segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable.

7. Validación de los resultados del método de segmentación propuesto

7.1 Introducción

En este capítulo, se logra el cumplimiento del último objetivo específico de la presente tesis, donde se propone un experimento hipotético para realizar la valoración de los resultados en un escenario real, es decir, si un tomador de decisión realiza su segmentación de clientes mediante el método propuesto en la presente tesis, qué tanto logra mejores resultados respecto a realizar un método de segmentación tradicional. En ese sentido, se comparan los resultados en términos económicos de segmentar incorporando la predicción del valor monetario, versus segmentar sin incorporar esta predicción.

A partir de lo desarrollado en los capítulos anteriores, el diagrama de flujo del método propuesto se presenta en la Figura 7-1.

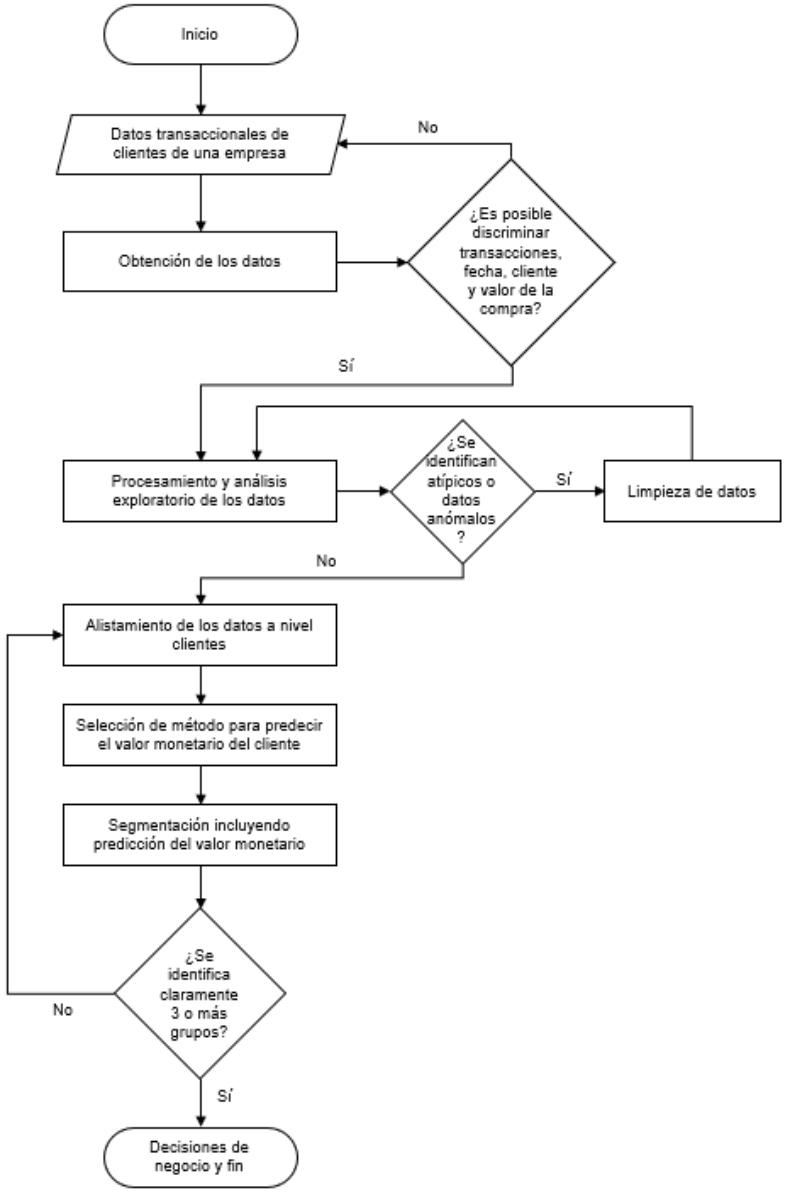

Figura 7-1: Representación de método propuesto para la segmentación de clientes incorporando la predicción del valor monetario
Fuente: elaboración propia

7.2 Metodología

Para la validación de los resultados se diseña un experimento hipotético real, en el que se plantea que un tomador de decisión debe seleccionar el mejor grupo de clientes para una estrategia de fidelización de 6 meses. La decisión la debe tomar a partir de las segmentaciones realizadas en el Capítulo 6, a saber: i) un método de segmentación que

no incorpora la predicción del valor monetario; y ii) un método de segmentación que incorpora la predicción del valor monetario como una variable de segmentación.

En ambas debe seleccionar el mejor grupo de clientes, teniendo en cuenta que ambos métodos de segmentación se hicieron con datos a corte 31 de mayo del 2011. Luego se valora qué información real de compra obtuvieron esos clientes desde el 1 de junio de 2011 al 9 de diciembre del 2011. El grupo de clientes que tenga una mayor contribución económica al negocio en esos 6 meses habrá sido la mejor elección.

Para lo anterior, el análisis se centra en el clúster 1 de clientes en ambos métodos de segmentación, en el sentido que cumple los criterios de menor recencia, mayor frecuencia, mayor valor monetario y mayor predicción del valor monetario. Finalmente se analizan los datos de valor monetario total de estos clústeres en el periodo de validación, así como los valores monetario promedio por cliente en cada clúster.

7.3 Resultados

En primera instancia, se toma como insumo los resultados obtenidos en la segmentación de clientes sin incorporar la predicción del valor monetario, como se muestra en la Tabla 7-1; y la segmentación de clientes incorporando la predicción del valor monetario, como se presenta en la Tabla 7-2. En ambas tablas, se identifica que el clúster 1, corresponde al grupo de clientes *top* o de clientes más valiosos*;* siguiendo la regla de menor recencia, mayor frecuencia, mayor valor monetario y mayor predicción del valor monetario[6].

Tabla 7-1: Resumen segmentación sin incorporar predicción del valor monetario

		Recencia	**Frecuencia**	**Valor monetario histórico**
Clúster	**Conteo**	**Media**	**Media**	**Media**
1	114	65,219	19,088	809,879
2	308	41,880	8,562	451,200
3	492	53,622	3,185	239,228
4	79	402,380	2,342	308,524

[6] En las Tablas 7-1 y 7-2, se presenta el resumen de los valores promedio de cada variable tenida en cuenta en el proceso de segmentación. No se presentan los valores mínimos y máximos de cada clúster, dado que fueron detallados anteriormente en el Capítulo 6.

		Recencia	Frecuencia	Valor monetario histórico
5	502	223,586	2,265	269,275

Fuente: elaboración propia

Tabla 7-2: Resumen de segmentación incorporando la predicción del valor monetario

		Recencia	Frecuencia	Valor monetario histórico	Predicción valor monetario
Clúster	**Conteo**	**Media**	**Media**	**Media**	**Media**
1	121	47,421	18,950	796,199	2785,809
2	247	40,939	8,672	453,025	1156,554
3	543	54,871	3,569	250,811	370,518
4	56	390,982	2,643	444,446	176,510
5	528	233,341	2,237	263,691	182,315

Fuente: elaboración propia

Luego, siguiendo con la metodología propuesta, se realizan 2 validaciones. La primera, tomando el clúster 1 de cada método para segmentar (incorporando y sin incorporar la predicción del valor monetario), y se analiza el valor total que ese grupo le aporta a la empresa entre el 1 de junio del 2011 y el 9 de diciembre del 2011; periodo que no fue tenido en cuenta en la fase de segmentación, y cuyos datos fueron reservados para esta fase de validación. Es así, como se observa en la Figura 7-2, que el clúster 1, de la segmentación incorporando la predicción del valor monetario, genera en el periodo de validación un total de valor monetario real de 12% por encima del clúster 1 sin incorporar la predicción del valor monetario; lo cual equivale a 56.723 libras esterlinas de ingresos adicionales.

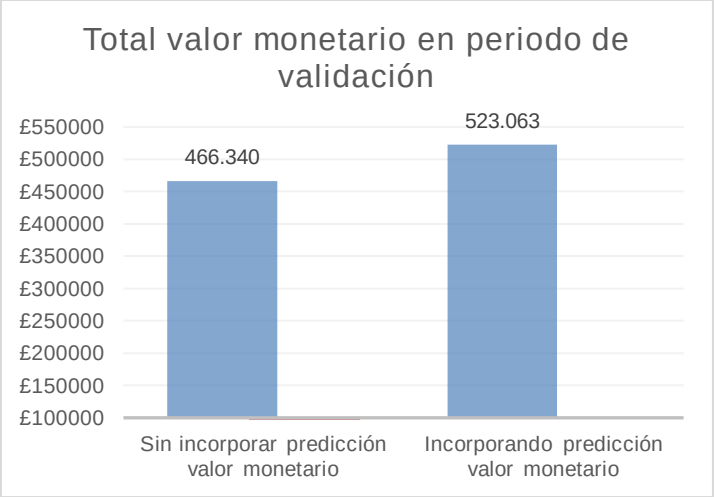

Figura 7-2: Validación de total valor monetario en periodo de validación
Fuente: elaboración propia

Ahora, respecto al valor monetario promedio por cliente en el periodo de validación, el cual se calcula dividiendo el valor monetario total sobre la cantidad de clientes en el clúster, también se observa que el valor promedio del grupo que incorpora la predicción del valor monetario es 6% superior a la del grupo que no incorpora la predicción del valor monetario; lo cual equivale a 232 libras esterlinas de contribución económica adicional por cada cliente, tal y como se presenta en la Figura 7-3.

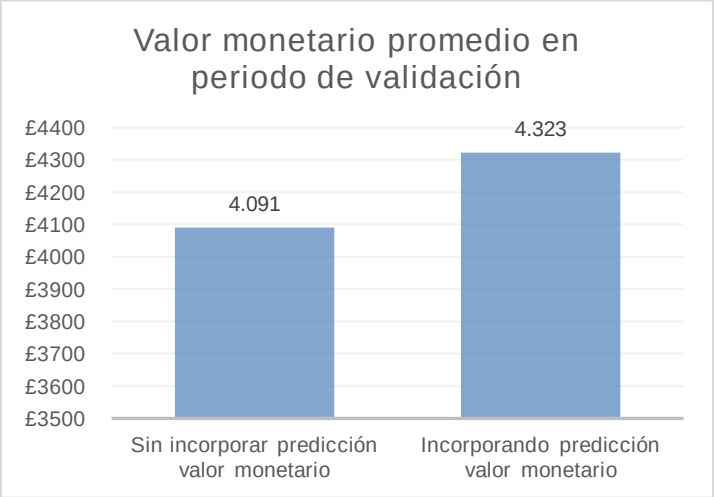

Figura 7-3: Validación de total valor monetario en periodo de validación
Fuente: elaboración propia

7.4 Conclusiones

El método propuesto de segmentación de clientes de la presente tesis demuestra que, con los datos utilizados, se valida la hipótesis principal del presente estudio, relacionada a que al realizar segmentación de clientes incorporando la variable de predicción del valor monetario del cliente, se contribuye a una mejor optimización de presupuesto con relación al mercadeo dirigido a sus clientes más valiosos. En este capítulo, se logró validar en un caso empírico que, si el tomador de decisión selecciona el grupo de clientes más valiosos, en los cuáles invertiría en campañas de fidelización de clientes, usando el método propuesto, que incorpora la predicción del valor monetario en el proceso de segmentación, se logra identificar un grupo que aporta un 12% de ingresos adicionales a la empresa, respecto a la selección de este grupo sin incorporar la predicción del valor monetario. Así mismo, se encuentra que, en cuanto a valor monetario promedio, cada cliente en este grupo representa ingresos adicionales por un 6% superior a la definición de un grupo sin incorporar esta predicción.

8. Conclusiones y recomendaciones

8.1 Conclusiones

Con los resultados de la presente tesis de investigación se logran aportes pertinentes al campo de conocimiento relacionado con la segmentación de clientes y la predicción del valor monetario de los clientes, a continuación, se plantean las conclusiones de la investigación realizada:

- Con el análisis bibliométrico y la revisión de la literatura realizada se identifica un campo del conocimiento en crecimiento, con oportunidades del uso del aprendizaje de máquinas para la predicción del valor monetario de los clientes, y así mismo su uso en la segmentación de clientes. En la revisión de la literatura, se identificó una línea de investigación en la cual autores proponen abordar la predicción del valor monetario del cliente mediante técnicas de aprendizaje de máquinas, sin embargo, no se encontró de manera detallada cómo comparar estos resultados con los métodos paramétricos o estadísticos.

  Se observa una baja participación de países latinoamericanos en los principales países del campo, donde en el caso de Colombia, solo se cuenta con 3 publicaciones en revistas de alto impacto indexadas a Scopus. Adicionalmente, se identifica que las publicaciones son de interés tanto en el campo de las revistas de negocios y ciencias sociales, como en revistas relacionadas a las ciencias de la computación. Finalmente, se concluye que es un campo del conocimiento que presenta ciclos de aproximadamente 5 años, en los cuales se renuevan los intereses alrededor del campo.

- Para la predicción del valor monetario del cliente se concluye que el método de aprendizaje de máquinas Nearest Neighbors, con la estructuración de los datos propuestos, logra un mejor desempeño que los modelos paramétricos

tradicionales. Aquí se identifica un aporte relevante a la literatura, puesto que aún son pocos los autores, en la literatura formal o científica, que han abordado la predicción del valor monetario del cliente, desde el enfoque de aprendizaje de máquinas, tales como Bauer & Jannach (2021) y Chen et al. (2019); principalmente porque los modelos tradicionales o paramétricos logran, en términos generales, buenos desempeños.

En ese sentido, con la implementación realizada en este capítulo, se aporta evidencia empírica a favor de la línea investigativa que propone abordar los problemas de predicción del valor monetario del cliente desde aplicando técnicas de aprendizaje de máquinas.

- Para el proceso de segmentación de los clientes, se realiza la segmentación de clientes sin incorporar la predicción del valor monetario y la segmentación de clientes incorporando la predicción del valor monetario, con el fin de identificar qué método logra mejores resultados en el proceso de segmentación de clientes. Los grupos conformados en cada clúster de ambas segmentaciones son el insumo principal para la validación de los resultados y del método propuesto.

  Se encuentra que, el algoritmo de K-means logra buenos resultados o desempeños, en casi todos los casos, en comparación con los demás algoritmos propuestos, tanto en la predicción sin incorporar la predicción del valor monetario, como en la predicción incorporando la predicción del valor monetario. Se destaca en este capítulo la definición de que en el proceso de limpieza no se pueden eliminar clientes que tengan un comportamiento aparentemente atípico, por el contrario, se propone el uso de algoritmos de detección de atípicos para identificar estos clientes y posteriormente segmentarlos; más no eliminarlos. De esta manera, se contribuye en uno de los vacíos identificados en la revisión de la literatura, en el sentido que muestra de qué manera se puede realizar el proceso de segmentación de clientes incorporando la predicción del valor monetario del cliente como una variable.

- Finalmente, el método propuesto de segmentación de clientes de la presente tesis demuestra que, con los datos utilizados, se valida la hipótesis principal del presente estudio, relacionada a que al realizar segmentación de clientes incorporando la variable de predicción del valor monetario del cliente, se contribuye a una mejor optimización de presupuesto con relación al mercadeo dirigido a sus clientes más valiosos. En este capítulo, se logró validar en un caso empírico que, si el tomador de decisión selecciona el grupo de clientes más valiosos, en los cuáles invertiría en campañas de fidelización de clientes, usando el método propuesto, que incorpora la predicción del valor monetario en el proceso de segmentación, se logra identificar un grupo que aporta un 12% de ingresos adicionales a la empresa, respecto a la selección de este grupo sin incorporar la predicción del valor monetario. Así mismo, se encuentra que, en cuanto a valor monetario promedio, cada cliente en este grupo representa ingresos adicionales por un 6% superior a la definición de un grupo sin incorporar esta predicción.

8.2 Recomendaciones

La presente tesis se concentró en proponer un método para la segmentación de clientes, incorporando la predicción del valor monetario como una variable de investigación. En la que se logró la implementación en un caso de datos reales y empírico, este método puede ser adaptado a futuros estudios, en los que como agenda de futuras investigaciones se sugiere:

- Proponer métodos para la predicción del valor monetario del cliente en otro tipo de negocios, diferentes al sector retail o minorista. Así mismo, considerar otros horizontes temporales, es decir, el uso de datos más recientes o antiguos, así como la disponibilidad de datos con mayor longitud, teniendo en cuenta que en el presente estudio se trabajó con 2 años de historia.

- Proponer otros algoritmos o técnicas de clustering para realizar la segmentación de clientes. Se puede considerar agrupar previamente a los clientes por cohortes, acorde con la fecha en la cual realizan la primera compra en la empresa.

- Realizar procesos de segmentación de clientes incorporando la predicción del valor monetario del cliente, incorporando otras variables no transaccionales durante el modelado, por ejemplo, la posibilidad de incluir variables sociodemográficas.

- En la presente tesis se incluyó la predicción de la segmentación de clientes como una variable en la predicción del valor monetario del cliente. En trabajos futuros se puede contemplar y validar el uso de la predicción del valor monetario del cliente como una variable para la segmentación de clientes.

- Finalmente, para trabajos futuros se puede contemplar un modelamiento previo de la deserción esperada de los clientes, a partir del comportamiento histórico que han tenido otros clientes con características sociodemográficas o transaccionales similares a los clientes del conjunto de análisis. Luego esta variable puede ser usada en el proceso de predicción del valor monetario del cliente, así como en el proceso de segmentación de clientes.

Referencias bibliográficas

Alipour, S. P. (2016). Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis. *Kybernetes*, *45*(7), 1129–1157. https://doi.org/10.1108/K-07-2015-0180

Ardanuy, J. (2012). Breve introducción a la bibliometría. In *Universitat de Barcelona* (pp. 1–25). https://doi.org/10.1038/nmat3485

Aria, M., & Cuccurullo, C. (2017). bibliometrix : An R-tool for comprehensive science mapping analysis. *Journal of Informetrics*, *11*(4), 959–975. https://doi.org/10.1016/j.joi.2017.08.007

Bauer, J., & Jannach, D. (2021). Improved Customer Lifetime Value Prediction with Sequence-To-Sequence Learning and Feature-Based Models. *ACM Transactions on Knowledge Discovery from Data*, *1*(1). https://doi.org/10.1145/3441444

Breunig, M. M., Kriegel, H.-P., Ng, R. T., & Sander, J. (2000). LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.*, *29*(2), 93–104. https://doi.org/10.1145/335191.335388

Cadavid, L., Awad, G., & Franco, C. (2012). Análisis bibliométrico del campo modelado de difusión de innovaciones. *Estudios Gerenciales*, *28*(EE), 213–236. http://www.scielo.org.co/pdf/eg/v28nspe/v28nspea12.pdf

Camps, D. (2008). Limitaciones de los indicadores bibliométricos en la evaluación de la actividad científica biomédica. In *Colombia Médica* (Vol. 39, pp. 74–79). scieloco.

Cancer Research UK. (2022). *Cancer Research UK Together we will beat cancer*. https://www.cancerresearchuk.org/

Channa, H. S. (2019). Customer lifetime value: An ensemble model approach. *Advances in Intelligent Systems and Computing*, *808*, 353–363. https://doi.org/10.1007/978-981-13-1402-5_27

Chatterjee, S., Rana, N. P., Tamilmani, K., & Sharma, A. (2021). The effect of AI-based CRM on organization performance and competitive advantage: An empirical analysis in the B2B context. *Industrial Marketing Management*, *97*, 205–219. https://doi.org/10.1016/j.indmarman.2021.07.013

Chen, D. (2019). *Online Retail II Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Online+Retail+II

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

Chen, P. P., Guitart, A., Del Río, A. F., & Periáñez, A. (2019). Customer Lifetime Value in Video Games Using Deep Learning and Parametric Models. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, 2134–2140. https://doi.org/10.1109/BigData.2018.8622151

Christy, A. J., Umamakeswari, A., Priyatharsini, L., & Neyaa, A. (2018). RFM ranking – An effective approach to customer segmentation. *Journal of King Saud University - Computer and Information Sciences*, 1–7. https://doi.org/10.1016/j.jksuci.2018.09.004

Dhamija, P., & Bag, S. (2020). Role of artificial intelligence in operations environment: a review and bibliometric analysis. *TQM Journal*, *32*(4), 869–896. https://doi.org/10.1108/TQM-10-2019-0243

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: An alternative to the Pareto/NBD model. *Marketing Science*, *24*(2), 275–284. https://doi.org/10.1287/mksc.1040.0098

Forliano, C., De Bernardi, P., & Yahiaoui, D. (2021). Entrepreneurial universities: A bibliometric analysis within the business and management domains. *Technological Forecasting and Social Change*, *165*, 1–16. https://doi.org/10.1016/j.techfore.2020.120522

Guadarrama, T., & Rosales, E. M. (2015). Marketing relacional: valor, satisfacción, lealtad y retención del cliente. Análisis y reflexión teórica. *Ciencia y Sociedad*, *40*(2), 307–340.

Gutiérrez-Salcedo, M., Martínez, M. Á., Moral-Munoz, J. A., Herrera-Viedma, E., & Cobo, M. J. (2018). Some bibliometric procedures for analyzing and evaluating research fields. *Applied Intelligence*, *48*(5), 1275–1287. https://doi.org/10.1007/s10489-017-1105-y

Hall, C. M. (2011). Publish and perish? Bibliometric analysis, journal ranking and the assessment of research quality in tourism. *Tourism Management*, *32*(1), 16–27. https://doi.org/10.1016/J.TOURMAN.2010.07.001

Heldt, R., Silveira, C. S., & Luce, F. B. (2021). Predicting customer value per product: From RFM to RFM/P. *Journal of Business Research*, *127*(March 2019), 444–453. https://doi.org/10.1016/j.jbusres.2019.05.001

Herrera-Franco, G., Montalván-Burbano, N., Carrión-Mero, P., Apolo-Masache, B., & Jaya-Montalvo, M. (2020). Research trends in geotourism: A bibliometric analysis using the scopus database. *Geosciences (Switzerland)*, *10*(10), 1–29. https://doi.org/10.3390/geosciences10100379

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*,

1–5. https://doi.org/10.1073/pnas.0507655102

Hurwitz, J., & Kirsch, D. (2018). *Machine Learning for dummies* (IBM Limite).

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2018). Modeling and application of customer lifetime value in online retail. *Informatics*, *5*(1), 1–22. https://doi.org/10.3390/informatics5010002

Jasek, P., Vrana, L., Sperkova, L., Smutny, Z., & Kobulsky, M. (2019). Comparative analysis of selected probabilistic customer lifetime value models in online shopping. *Journal of Business Economics and Management*, *20*(3), 398–423. https://doi.org/10.3846/jbem.2019.9597

Kansal, T., Bahuguna, S., Singh, V., & Choudhury, T. (2018). Customer Segmentation using K-means Clustering. *2018 International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS)*, 135–139. https://doi.org/10.1109/CTEMS.2018.8769171

Linnenluecke, M. K., Marrone, M., & Singh, A. K. (2020). Conducting systematic literature reviews and bibliometric analyses. *Australian Journal of Management*, *45*(2), 175–194. https://doi.org/10.1177/0312896219877678

Liu, F. T., Ting, K. M., & Zhou, Z.-H. (2008). Isolation Forest. *2008 Eighth IEEE International Conference on Data Mining*, 413–422. https://doi.org/10.1109/ICDM.2008.17

Nita, A. (2019). Empowering impact assessments knowledge and international research collaboration - A bibliometric analysis of Environmental Impact Assessment Review journal. *Environmental Impact Assessment Review*, *78*(March), 106283. https://doi.org/10.1016/j.eiar.2019.106283

Oblander, E. S., Gupta, S., Mela, C. F., Winer, R. S., & Lehmann, D. R. (2020). The past, present, and future of customer management. *Marketing Letters*, *31*(2–3), 125–136. https://doi.org/10.1007/s11002-020-09525-9

Pareto, V. (1896). *Cours d'économie politique*. Université de Lausanne.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Peralta, M. J., Frías, M., & Gregorio, O. (2015). Criterios, clasificaciones y tendencias de los indicadores bibliométricos en la evaluación de la ciencia. *Revista Cubana de Información En Ciencias de La Salud*, *26*(3), 290–309. http://scielo.sld.cu

Platzer, M. (2021). *Customer Base Analysis with BTYDplus* (pp. 1–33). https://cran.r-project.org/web/packages/BTYDplus/vignettes/BTYDplus-HowTo.pdf

Rathi, T., & Ravi, V. (2017). Customer Lifetime Value Measurement using Machine Learning Techniques. In *Artificial Intelligence: Concepts, Methodologies, Tools, and Applications* (pp. 3013–3022). https://doi.org/10.4018/978-1-5225-1759-7.ch124

Roy-García, I., Rivas-Ruiz, R., Pérez-Rodríguez, M., & Palacios-Cruz, L. (2019). Correlation: Not all correlation entails causality. *Revista Alergia Mexico*, *66*(3), 354–360. https://doi.org/10.29262/ram.v66i3.651

Rueda, G., Gerdsri, P., & Kocaoglu, D. F. (2007). Bibliometrics and Social Network Analysis of the Nanotechnology Field. *PICMET '07 - 2007 Portland International Conference on Management of Engineering & Technology*, 2905–2911. https://doi.org/10.1109/PICMET.2007.4349633

Sifa, R., Runge, J., Bauckhage, C., & Klapper, D. (2018). Customer Lifetime Value Prediction in Non-Contractual Freemium Settings: Chasing High-Value Users Using Deep Neural Networks and SMOTE. *Hawaii International Conference on System Sciences*, 923–932.

Simeone, O. (2018). A Brief Introduction to Machine Learning for Engineers. *Foundations and Trends® in Signal Processing*, *12*(3–4), 200–431. https://doi.org/10.1561/2000000102

Sodhi, P., Awasthi, N., & Sharma, V. (2019). Introduction to Machine Learning and Its Basic Application in Python. *SSRN Electronic Journal*, 1354–1375. https://doi.org/10.2139/ssrn.3323796

Srivastava, R. (2017). Identification of customer clusters using RFM model: a case of diverse purchaser classification. *International Journal of Information, Business and Management*, *9*(4), 201–208.

Tsai, T. Y., Lin, C. T., & Prasad, M. (2019). An Intelligent Customer Churn Prediction and Response Framework. *Proceedings of IEEE 14th International Conference on Intelligent Systems and Knowledge Engineering, ISKE 2019*, 928–935. https://doi.org/10.1109/ISKE47853.2019.9170380

Valérie, D., & Pierre, A. G. (2010). Bibliometric idicators: Quality masurements of sientific publication. *Radiology*, *255*(2), 342–351. https://doi.org/10.1148/radiol.09090626

Velasco, B., Eiros Bouza, J. M., Pinilla, J. M., & San Román, J. A. (2012). La utilización de los indicadores bibliométricos para evaluar la actividad investigadora. *Aula Abierta*, *40*(2), 75–84. http://dialnet.unirioja.es/servlet/articulo?codigo=3920967&info=resumen&idioma=ENG

Villa, E., Ruiz, L., Valencia, A., & Picón, E. (2018). Electronic commerce: factors involved in its adoption from a bibliometric analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, *13*(1), 39–70. https://doi.org/10.4067/S0718-18762018000100104

Win, T. T., & Bo, K. S. (2020). Predicting Customer Class using Customer Lifetime Value with Random Forest Algorithm. *Proceedings of the 4th International Conference on Advanced Information Technologies, ICAIT 2020*, 236–241. https://doi.org/10.1109/ICAIT51105.2020.9261792

Yoseph, F., & Heikkila, M. (2019). Segmenting retail customers with an enhanced RFM and a hybrid regression/clustering method. *Proceedings - International Conference*

on Machine Learning and Data Engineering, ICMLDE 2018, 77–82. https://doi.org/10.1109/iCMLDE.2018.00029