



UNIVERSIDAD  
**NACIONAL**  
DE COLOMBIA

# **Método para la predicción de la intención de compra de los usuarios en línea utilizando técnicas de aprendizaje de máquina**

**Luis Felipe Ortiz-Clavijo**

Universidad Nacional de Colombia

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

Medellín, Colombia

2024

# **Método para la predicción de la intención de compra de los usuarios en línea utilizando técnicas de aprendizaje de máquina**

**Luis Felipe Ortiz-Clavijo**

Tesis de investigación presentada como requisito parcial para optar al título de:  
**Magister en Ingeniería - Analítica**

Director:

Sergio Armando Gutiérrez Betancur, Ph.D. (Universidad de Antioquia, Medellín)

Codirector:

John Willian Branch Bedoya, Ph.D. (Universidad Nacional de Colombia, Medellín)

Línea de Investigación:

Redes Neuronales Artificiales, Computación Evolutiva, y Reconocimiento de Patrones

Grupo de Investigación:

GIDIA – Grupo I+D en Inteligencia Artificial

Universidad Nacional de Colombia

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

Medellín, Colombia

2024

*A mis padres,  
Martha Lucia y Luis Eduardo.*

## **Agradecimientos**

A mis directores de tesis, los profesores Sergio Armando Gutiérrez y John Willian Branch, por su invaluable apoyo y acompañamiento durante todo el proceso.

Al Grupo de Investigación y Desarrollo en Inteligencia Artificial (GIDIA) y su seminario permanente de investigación. Gracias por los espacios de divulgación, transferencia de conocimiento y retroalimentación.

## Resumen

### **Método para la predicción de la intención de compra de los usuarios en línea utilizando técnicas de aprendizaje de máquina**

El desafío de predecir la intención de compra de los usuarios en línea constituye un aspecto crítico en el dinámico mundo del comercio electrónico. En este contexto, comprender y anticipar el comportamiento del consumidor se ha vuelto una prioridad para las empresas con presencia digital. Esta investigación aborda dicho desafío desarrollando un método predictivo para la intención de compra de usuarios digitales, empleando técnicas de aprendizaje de máquina. A través de un conjunto de datos representativo, que incluye información detallada sobre las actividades de los usuarios en un sitio de comercio electrónico, se ajustan técnicas de aprendizaje de máquina, utilizando un enfoque de ensamble de Bosques Aleatorios y XGBoost. Los resultados obtenidos demuestran que el método propuesto alcanza una precisión general del 89.69%, demostrando su habilidad para identificar correctamente cuándo los usuarios tienen la intención de realizar una compra en línea y cuándo no. La investigación aporta al campo del comercio electrónico un enfoque predictivo que se centra en la precisión y la generalización, adaptándose a variaciones en el comportamiento de compra de los usuarios digitales. Este enfoque de generalización implica que el modelo no está restringido a un conjunto de datos específico o a condiciones de mercado particulares, sino que puede ajustarse y mantener su precisión ante los cambios dinámicos que presenta el comercio electrónico.

**Palabras clave:** Aprendizaje de máquina, predicción de compra en línea, comercio electrónico, bosques aleatorios, XGBoost, análisis predictivo, comportamiento del consumidor.

## Abstract

### **Method for Predicting Online Shopping Intentions Using Machine Learning Techniques**

The challenge of predicting online users' purchase intentions constitutes a critical aspect in the dynamic world of e-commerce. In this context, understanding and anticipating consumer behavior has become a priority for companies with a digital presence. This research addresses this challenge by developing a predictive method for predicting purchase intentions of digital users, employing machine learning techniques. Through a representative dataset, which includes detailed information about users' activities on an e-commerce site, machine learning techniques are adjusted, using an ensemble approach of Random Forests and XGBoost. The results obtained demonstrate an overall accuracy of 89.69%, proving its ability to correctly identify when users intend to make an online purchase and when they do not. The research contributes to the field of e-commerce a predictive approach that focuses on accuracy and generalization, adapting to variations in the purchasing behavior of digital users. This generalization approach implies that the model is not restricted to a specific dataset or market conditions but can be adjusted and maintain its accuracy amidst the dynamic changes presented in e-commerce.

**Keywords:** Machine learning, online purchase prediction, e-commerce, random forests, XGBoost, predictive analysis, consumer behavior.

# Contenido

	<b>Pág.</b>
<b>Resumen</b>	<b>IX</b>
<b>Lista de figuras</b>	<b>13</b>
<b>Lista de tablas</b>	<b>14</b>
<b>1. Introducción</b>	<b>15</b>
1.1 Motivación	15
1.2 Descripción del problema	17
1.3 Objetivos	18
1.3.1 Objetivo general	18
1.3.2 Objetivos específicos	18
1.4 Contribución	18
1.5 Estructura del documento	19
<b>2. Marco teórico de referencia</b>	<b>20</b>
2.1 Comercio electrónico	20
2.2 Técnicas de aprendizaje de máquina	21
2.3 Intención de compra	22
2.3.1 Predicción de la intención de compra	22
<b>3. Revisión sistemática de la literatura RSL</b>	<b>24</b>
3.1 Criterios de selección y fuentes de información	24
3.2 Resultados revisión de la literatura	26
3.2.1 Descripción de los estudios seleccionados	29
3.2.2 Análisis y discusión	32
<b>4. Predicción de la intención de compra de usuarios en línea</b>	<b>34</b>
4.1 Conjunto de datos para la predicción de la intención de compra	34
4.2 Análisis exploratorio de datos	39
4.2.1 Variables numéricas	39
4.2.2 Variables categóricas	46
4.2.3 Intención de compra vs ejecución de la compra	50
4.3 Desarrollo del método predictivo	51
4.3.1 Técnicas reportadas en la literatura	51
4.3.2 Métricas de evaluación	52
4.3.3 Implementación	53

<b>5. Resultados y discusión</b>	<b>55</b>
5.1 Evaluación de técnicas	55
5.2 Método para la predicción de la intención de compra de los usuarios en línea	59
5.2.1 Fundamentación del método predictivo	59
5.2.2 Técnica de Bosques Aleatorios	60
5.2.3 Técnica XGBoost	63
5.2.4 Definición del método predictivo: ensamble de técnicas	66
5.3 Trabajos reportados en la literatura	71
<b>6. Conclusiones</b>	<b>73</b>
<b>7. Trabajo futuro</b>	<b>75</b>
<b>Bibliografía</b>	<b>76</b>

## Lista de figuras

	<b>Pág.</b>
<b>Figura 4-1.</b> Matriz de correlación de variables numéricas .....	39
<b>Figura 4-2.</b> Diagrama de caja: contribución de página al ingreso.....	41
<b>Figura 4-3.</b> Diagrama de caja: sesiones con tasa de rebote .....	42
<b>Figura 4-4.</b> Diagrama de caja: sesiones sin compra .....	43
<b>Figura 4-5.</b> Distribución de las variables numéricas .....	44
<b>Figura 4-6.</b> Distribución de variables categóricas: Meses .....	46
<b>Figura 4-7.</b> Distribución de variables categóricas: tipo de visitante .....	47
<b>Figura 4-8.</b> Distribución de variables categóricas: fines de semana.....	47
<b>Figura 4-9.</b> Distribución de variables categóricas: tipo de visitante vs revenue .....	48
<b>Figura 4-10.</b> Distribución de variables categóricas: fines de semana vs Revenue .....	49
<b>Figura 5-1.</b> Comparativa de la capacidad discriminativa de las técnicas.....	57
<b>Figura 5-2.</b> Características principales para Bosques Aleatorios .....	60
<b>Figura 5-3.</b> Matriz de confusión – Técnica Bosques Aleatorios.....	61
<b>Figura 5-4.</b> Resultados métricas de evaluación para Bosques Aleatorios .....	62
<b>Figura 5-5.</b> Características principales para XGBoost.....	63
<b>Figura 5-6.</b> Matriz de confusión – Técnica XGBoost .....	64
<b>Figura 5-7.</b> Resultados métricas de evaluación para XGBoost .....	65
<b>Figura 5-8.</b> Resultados de métricas de evaluación del método propuesto .....	68
<b>Figura 5-9.</b> Diagrama de flujo del método propuesto .....	69

## Lista de tablas

	Pág.
<b>Tabla 3-1:</b> Parámetros de la revisión de literatura .....	24
<b>Tabla 3-2:</b> Artículos seleccionados para la revisión .....	26
<b>Tabla 3-3:</b> Técnicas de referencia reportadas en la literatura para la predicción.....	33
<b>Tabla 4-1:</b> Diccionario de variables del conjunto de datos .....	37
<b>Tabla 5-1:</b> Comparativa de rendimiento de técnicas .....	55
<b>Tabla 5-2:</b> Trabajos similares reportados en la literatura .....	71

# 1.Introducción

## 1.1 Motivación

El comercio electrónico ha experimentado un crecimiento exponencial en las últimas décadas, impulsado por avances en las tecnologías de la información y las comunicaciones, así como por el acceso generalizado a Internet y el auge de las redes sociales (Cheba et al., 2021; Rosário y Raimundo, 2021; Viu-Roig y Alvarez-Palau, 2020). Este crecimiento ha llevado a la expansión de los mercados en línea a nivel mundial, cambiando la forma en que las personas compran y venden productos y servicios (Alsaad y Taamneh, 2019). En Colombia, el crecimiento del comercio electrónico ha sido impulsado por el aumento en la cantidad de usuarios de Internet y el creciente número de usuarios de teléfonos móviles (Quintero, 2021; Suárez, 2020). Este contexto nacional subraya una tendencia global hacia la digitalización del consumo, que a su vez genera grandes cantidades de datos sobre el comportamiento de los consumidores. Estos datos indican una área de oportunidad para investigar aspectos críticos relacionados con el comportamiento del consumidor, como la intención de compra en línea, la cual es fundamental para las empresas que buscan mejorar sus estrategias de mercadeo y ventas en entornos digitales (Ricci, 2022).

La intención de compra ha atraído un interés académico, reflejado en investigaciones que exploran temas como la predicción (Chaudhuri et al., 2021; Esmeli et al., 2021; Micol Policarpo et al., 2021; Peña-García et al., 2020), la segmentación de mercado (Huang y Benyoucef, 2015; Huseynov y Özkan Yıldırım, 2019), el análisis de los elementos influyentes en el comportamiento del consumidor (Blackwell et al., 2006; Kotler y Keller, 2016; Solomon, 2017), la formulación de estrategias de mercadeo (Svobodová y

Rajchlová, 2020), y el análisis de los factores que influyen el comportamiento del consumidor digital (Jiménez et al., 2019; Miao et al., 2021).

El análisis de la intención de compra se ha consolidado como una herramienta con potencial para las empresas que buscan no solo entender a sus clientes sino también anticiparse a sus necesidades y deseos. Este análisis profundiza en factores como percepciones, actitudes, motivaciones y decisiones, permitiendo a las empresas ajustar su oferta y personalizar la experiencia de compra (Islam et al., 2023). Además, en la literatura se han reportado diversos desafíos asociados a la intención de compra, tales como la evolución acelerada de las intenciones, la presencia de ruido en los datos de navegación de usuarios de comercios digitales, y la dificultad para distinguir la intención de compra de otros factores como el precio, la disponibilidad del producto y su conveniencia (Kazancoglu y Aydin, 2018; Z. Wang et al., 2021; Zozalbo y Astuti, 2022).

La ciencia de datos, correspondiente a un campo interdisciplinario que integra matemáticas, estadística e informática, resulta fundamental para descubrir tendencias valiosas (He y Yin, 2021), mostrando un potencial significativo para analizar dichas tendencias y comportamientos, con ejemplos de enfoques sugeridos por diversos autores para prever las necesidades del mercado y fortalecer las estrategias comerciales (Bacile, 2020; Bangyal et al., 2022).

En este contexto, las técnicas de aprendizaje automático se destacan por su habilidad para generalizar el conocimiento obtenido de datos existentes hacia nuevos datos o situaciones no vistas previamente, proporcionando flexibilidad y precisión cruciales en el comercio electrónico (Chen et al., 2022).

## 1.2 Descripción del problema

La intención de compra es reconocida por su complejidad y naturaleza multifacética, influenciada por una amplia variedad de factores tanto internos como externos al consumidor. Los factores internos, que abarcan las características personales y psicológicas del individuo, tales como la motivación, las actitudes y las preferencias personales, se combinan con elementos externos relacionados con el entorno de mercado, como el precio y la disponibilidad del producto (Gupta et al., 2021). Además, el contexto específico en el que se encuentra el consumidor digital, caracterizado por variables situacionales como el momento del día, la estacionalidad y la ubicación geográfica, ejerce una influencia significativa en la intención de compra, añadiendo mayor complejidad al fenómeno (Sokolova y Kefi, 2020).

Esta interacción entre factores internos, externos y contextuales no solo subraya la complejidad inherente al proceso de compra, sino que también presenta desafíos significativos en el desarrollo de modelos predictivos generalizables y precisos. La capacidad de un modelo para generalizar a través de diversas situaciones de compra, adaptándose a diferentes tipos de consumidores y entornos de mercado, es crucial para su aplicabilidad práctica en el ámbito dinámico del comercio electrónico (Bawack et al., 2022; Mannering et al., 2020).

Los principales desafíos identificados en la literatura relacionada con la predicción de la intención de compra incluyen la necesidad de datos precisos y representativos que reflejen la complejidad del comportamiento del consumidor, la capacidad para manejar interacciones complejas entre una amplia gama de variables influyentes, y la adaptabilidad frente a cambios en la intención de compra a lo largo del tiempo (Chaudhuri et al., 2021; Dong y Jiang, 2019; Kabir et al., 2019; Mokryn et al., 2019).

Frente a estos desafíos, la aplicación de técnicas de aprendizaje de máquina constituye un enfoque prometedor. Estas técnicas permiten no solo mejorar la precisión de las predicciones sino también facilitar una mejor comprensión de los factores determinantes de la intención de compra, gracias a la capacidad de éstas para identificar patrones complejos y no lineales en grandes conjuntos de datos (Chaudhuri et al., 2021).

La elección del aprendizaje de máquina para abordar la predicción de la intención de compra en el comercio electrónico se justifica por su potencial para superar las limitaciones de los métodos analíticos tradicionales, ofreciendo un modelo predictivo que es tanto generalizable como adaptable a las dinámicas fluctuantes del mercado y las características variadas de los consumidores (Dong y Jiang, 2019; Kabir et al., 2019).

## **1.3 Objetivos**

### **1.3.1 Objetivo general**

Proponer un método para predicción de la intención de compra de los usuarios en línea utilizando técnicas de aprendizaje de máquina.

### **1.3.2 Objetivos específicos**

- Seleccionar un conjunto de datos que incluya información de compradores en línea para ser utilizado en el desarrollo del método.
- Diseñar un método de predicción que identifique la intención de compra de los usuarios en línea empleando técnicas de aprendizaje de máquinas.
- Validar el desempeño del método propuesto, con respecto a otros reportados en la literatura.

## **1.4 Contribución**

Esta investigación aporta al campo del comercio electrónico mediante la propuesta de un método predictivo que integra el análisis del comportamiento de los usuarios en línea con técnicas de aprendizaje de máquina. Se destaca especialmente la aplicación de Bosques Aleatorios y XGBoost en un enfoque de ensamble como estrategia para mejorar la precisión en la predicción.

## **1.5 Estructura del documento**

La tesis se estructura de la siguiente manera: El Capítulo 2 establece un marco teórico que aborda la problemática de la intención de compra en el comercio electrónico y su predicción a través del aprendizaje de máquina, explorando diferentes métodos y enfoques dentro de este campo. En el Capítulo 3, se presenta una revisión sistemática de la literatura, destacando estudios clave que han influido en el desarrollo de la investigación y las tendencias actuales en la predicción de la intención de compra en línea. El Capítulo 4 detalla la metodología y el desarrollo del método predictivo, incluyendo la selección del conjunto de datos, las técnicas exploradas, las métricas de evaluación y la implementación del método predictivo. Los resultados se presentan en el Capítulo 5, donde se discute el método propuesto y se compara con otros reportados en la literatura, concluyendo con las reflexiones derivadas de su implementación. El Capítulo 6 recoge las conclusiones generales de la tesis, resumiendo los hallazgos y reflexiones principales. Por último, el Capítulo 7 se delinean las propuestas de trabajo futuro en relación con la predicción de la intención de compra en el ámbito del comercio electrónico.

## 2. Marco teórico de referencia

En esta sección, se exploran los aspectos fundamentales relacionados con la temática de estudio. Se examinan conceptos clave tales como el comercio electrónico, el aprendizaje de máquina, la predicción de la intención de compra y el comportamiento del consumidor en línea.

### 2.1 Comercio electrónico

Desde su aparición en la década de los años 90, el comercio electrónico ha experimentado un crecimiento exponencial y transformaciones significativas (Rico et al., 2008). El comercio electrónico se define como la compra y venta de bienes y servicios a través de medios electrónicos, especialmente Internet (Ahn et al., 2004; Laudon y Traver, 2013). Este comercio se clasifica en varias categorías: B2B (Business-to-Business), B2C (Business-to-Consumer) y C2C (Consumer-to-Consumer), que incluyen transacciones entre empresas, entre empresas y consumidores, o entre consumidores. Además, se encuentran las categorías B2A (Business-to-Administration) y C2A (Consumer-to-Administration), que involucran transacciones con entidades de administración pública (Babenko et al., 2019; Turban et al., 2018).

Con el avance de la tecnología, especialmente en lo que respecta al uso de teléfonos inteligentes y redes sociales, el comercio electrónico se ha vuelto una parte integral de la vida cotidiana. Los consumidores, ahora capaces de realizar compras en cualquier lugar y momento, y demandan cada vez más personalización y respuestas inmediatas por parte de las empresas. Así, el ámbito del comercio electrónico se ha expandido para abarcar no solo la venta de bienes físicos sino también la oferta de servicios y contenidos digitales, diversificando las experiencias de compra y venta en línea.

Algunas empresas, en respuesta a esta evolución, adoptan modelos analíticos para personalizar las experiencias de sus usuarios. Estos modelos se basan en el análisis del comportamiento y las preferencias de los consumidores, utilizando para ello herramientas de inteligencia de negocios y técnicas de aprendizaje de máquina (Du et al., 2019; Lu et al., 2021; Micol et al., 2021).

## **2.2 Técnicas de aprendizaje de máquina**

El aprendizaje de máquina (Machine Learning - ML por su sigla en inglés) es un subcampo de las ciencias de la computación que se centra en el uso de datos y algoritmos para crear modelos que mejoran gradualmente su precisión al imitar la forma en que los humanos aprenden. Este campo se enfoca en la creación de algoritmos que mejoran su desempeño en una tarea específica a medida que se les proporciona más datos (Zhou y Liu, 2021). Los algoritmos de ML pueden clasificarse en supervisados, no supervisados, o semi-supervisados, según la disponibilidad de datos etiquetados para el entrenamiento (Sindhu y Suriya, 2020).

La aplicación del ML trasciende diversos dominios, incluyendo la minería de datos, el procesamiento de imágenes y el análisis predictivo. En el ámbito del comercio electrónico, el aprendizaje de máquina se revela como una herramienta poderosa para analizar y predecir la intención de compra, permitiendo a las empresas ofrecer experiencias de usuario personalizadas y afinar sus estrategias de mercadeo (Esmeli et al., 2021; Xiao et al., 2019). Entre las aplicaciones destacadas en este sector se encuentran:

- **Gestión de marca:** Los algoritmos de regresión y las máquinas de soporte vectorial mejoran el marketing online y el posicionamiento de marca, facilitando la comunicación en redes sociales (Pamuksuz et al., 2021; Q. Wang et al., 2020)
- **Seguridad y eficiencia:** El ML contribuye a la seguridad empresarial y fomenta sistemas de negocio más eficientes, al optimizar la velocidad y la seguridad de las transacciones financieras, así como al mejorar la interacción y satisfacción del usuario en plataformas digitales (Rath, 2020).

- Sistemas de recomendación: Recomendaciones de productos, estimación de tallas, precios dinámicos y detección de reseñas falsas, enriqueciendo la experiencia de compra (Islek y Oguducu, 2022; Tahir et al., 2021)
- Predicción: Métodos de predicción de intención de compra basados en modelos de ensamble (Dhali et al., 2020). Predicción de género de los clientes a partir de datos de comportamiento y contexto (Khan et al., 2020).
- Aplicaciones en la gestión empresarial: Mejoras en la cadena de suministro, experiencia del cliente y eficiencia operativa a través del ML (Pallathadka et al., 2023).

Dentro de las diversas aplicaciones, el enfoque de ensamble emerge como una extensión del aprendizaje de máquina, buscando fortalecer aún más la precisión y la robustez de los modelos predictivos. Este método consiste en combinar las predicciones de múltiples modelos de ML para formar una predicción final más precisa y confiable (Dhali et al., 2020).

## **2.3 Intención de compra**

La intención de compra se define como el propósito o la disposición de un usuario de adquirir un producto o servicio. Es una medida de la inclinación del usuario hacia la compra, que puede no siempre traducirse en una compra real. La intención de compra es un indicador crucial en el comercio electrónico, ya que permite predecir la probabilidad de que un consumidor realice una transacción en el futuro (Rasheed et al., 2014).

Es importante distinguir entre la intención de compra y la ejecución de compra. Mientras que la intención de compra refleja el propósito o la disposición de un usuario de adquirir un producto o servicio, la ejecución de compra se refiere a la acción concreta de realizar una transacción comercial efectiva, es decir, cuando un usuario completa el proceso de compra y efectivamente adquiere un producto o servicio.

### **2.3.1 Predicción de la intención de compra**

La predicción de la intención de compra es un elemento clave en el ámbito del comercio, tanto en el entorno físico como en el digital. Este proceso se enfoca en la capacidad de las empresas para anticipar la probabilidad de que un consumidor realice una compra en el futuro. Es un procedimiento complejo que implica la utilización de técnicas analíticas

para examinar múltiples factores que pueden influir en las decisiones de compra de un consumidor (Trivedi y Sama, 2020).

Entre estos factores, se incluyen el comportamiento de compra anterior del consumidor, sus preferencias y gustos personales, la influencia de las opiniones de otros consumidores y elementos externos como las condiciones económicas generales o las tendencias del mercado (Kotler y Keller, 2016; Solomon, 2017). Estos factores son cruciales para refinar estrategias que mejoren la eficacia en la retención y adquisición de clientes (Chaudhuri et al., 2021; Svobodová y Rajchlová, 2020).

El comportamiento del consumidor en línea comprende cómo los individuos buscan, evalúan, seleccionan y compran productos y servicios a través de Internet, según Constantinides (2004). Este concepto va más allá de la simple acción de comprar en línea, ya que involucra una serie de factores psicológicos y sociales que moldean el proceso de compra.

Entre estos factores, se destaca la motivación que impulsa a los consumidores a buscar y elegir determinados productos o servicios, según Padmavathy et al. (2019). Asimismo, la percepción juega un papel crucial, ya que determina cómo los consumidores interpretan la información y las ofertas disponibles (Marceda et al., 2020).

Además, la actitud hacia ciertas marcas, productos o servicios, que puede ser influenciada por diversos elementos (Wang et al., 2021). Dicha actitud incluye también el proceso de toma de decisiones de compra, un recorrido que el consumidor realiza desde la identificación de una necesidad hasta la concreción de la compra (Esmeli et al., 2021).

De ahí que el estudio del comportamiento del consumidor en línea requiera una comprensión profunda tanto de los consumidores como de los factores que influyen en sus decisiones y acciones en el entorno digital (Zozalbo y Astuti, 2022).

### 3.Revisión sistemática de la literatura RSL

En esta sección se desarrolla una revisión sistemática de la literatura, centrada en estudios e investigaciones recientes que exploran la predicción de la intención de compra empleando técnicas de aprendizaje de máquina. Se incluyen los criterios de selección y fuentes de información, los resultados de la revisión presentando la descripción de los estudios seleccionados, y el análisis y discusión de los aspectos relevantes.

#### 3.1 Criterios de selección y fuentes de información

Los parámetros de esta revisión se detallan en la Tabla 3-1.

**Tabla 3-1:** Parámetros de la revisión de literatura

<b>Objetivo</b>	Identificar las principales técnicas de predicción reportadas en la literatura, con potencial uso para el tratamiento de la intención de compra en el comercio electrónico.	
<b>Marco temporal</b>	Publicaciones entre 2019 - 2024	
<b>Palabras clave</b>	Online purchase intention	
	Predict	
	Machine learning	
	Algorithms	
	Consumer behavior e-commerce	
<b>Fuentes de información</b>	Artículos de revista y trabajos presentados en conferencias científicas.	<b>Bases de datos</b>
		IEEE
		ACM Elsevier

La búsqueda se llevó a cabo en los metabuscaadores Web of Science y Scopus, a partir de la siguiente ecuación de búsqueda:

(((((ALL=(online purchase intention)) OR ALL=(purchase intention)) AND ALL=(prediction)) AND ALL=(machine learning)) AND ALL=(algorithms)) AND ALL=(consumer behavior)) OR ALL=(consumer behaviour)) AND ALL=(e-commerce)) OR ALL=(electronic commerce) and IEEE or Assoc Computing Machinery or Elsevier (Publishers) and Article or Proceeding Paper (Document Types) and 2024 or 2023 or 2022 or 2021 or 2020 or 2019 (Publication Years) and Article or Proceeding Paper (Document Types).

La búsqueda documental se focalizó en la búsqueda en las bases de datos Elsevier, IEEE y ACM.

### **Criterios de inclusión y exclusión:**

Los estudios fueron seleccionados en varias etapas, incluyendo la eliminación de duplicados, la revisión de títulos y resúmenes, y la lectura completa de los artículos seleccionados.

Se evaluó la calidad de los estudios utilizando criterios como:

- Relevancia del contenido
- Metodología (uso de técnicas de aprendizaje de máquina)
- Validez de los resultados

**Inclusión:** Investigaciones centradas en técnicas de aprendizaje de máquina aplicadas a la predicción de la intención de compra en línea y estudios con resultados empíricos y cuantitativos.

**Exclusión:** Estudios no enfocados en la predicción de la intención de compra en línea, investigaciones basadas en técnicas distintas al aprendizaje de máquina y estudios sin resultados empíricos o cuantitativos.

Los atributos considerados para cada estudio incluyeron título, palabras clave, autores, año y fuente de publicación.

**Análisis de los estudios seleccionados.** Se extrajeron datos relevantes de los estudios seleccionados, categorizándolos en técnicas de aprendizaje de máquina utilizadas, precisión de los modelos, y variables y características consideradas.

De esta manera, se seleccionaron 19 artículos para esta investigación, basándose en su relevancia y aporte a la temática de estudio.

A continuación, en la Tabla 3-2 se detallan las características de los artículos seleccionados, proporcionando una visión general de los enfoques y hallazgos más relevantes.

### 3.2 Resultados revisión de la literatura

**Tabla 3-2.** Artículos seleccionados para la revisión

<b>Título</b>	<b>Palabras clave</b>	<b>Autores / año</b>	<b>Revista</b>
Impact of Brand Marketing Strategies Based on Consumer Purchase Intention Mining	Brand marketing strategies, machine learning, purchase intention.	(P. Wang, 2024)	Computer-Aided Design and Applications
Predicting online customer purchase: The integration of customer characteristics and browsing patterns	Customer purchase behavior, customer decision journey, RFM graph metrics, predictive analysis, clustering analysis.	(Kim et al., 2024)	Decision Support Systems
Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods	Missing values, imbalanced classes, machine learning, consumer preference.	(Liu et al., 2024)	Expert Systems with Applications
A gradient boosting classifier for purchase intention prediction of online shoppers	Gradient boosting classifier, imbalanced dataset, feature selection, online shopper's purchase intention, real time prediction.	(Abdullah-All-Tanvir et al., 2023)	Heliyon

<b>Título</b>	<b>Palabras clave</b>	<b>Autores / año</b>	<b>Revista</b>
Analysis and Prediction of Purchase Intention of Online Customers with Deep Learning	Purchase intension, machine learning, deep learning, online shopping, customer retention.	(Bansal y Vyas, 2023)	Proceedings of Data Analytics and Management. Lecture Notes in Networks and Systems
Prediction of Buying Intention: Factors Affecting Online Shopping	Online Shopping behaviour, machine learning, purchase intention, prediction, e-commerce.	(Islam et al., 2023)	Proceedings of International Conference on Next-Generation Computing, IoT and Machine Learning
Data Mining Model for Predicting Customer Purchase Behavior in e-Commerce Context	Apriori PT algorithm, C4.5, CS-MC4, datamining, decision tree, e-commerce, k-means.	(Alghanam et al., 2022)	International Journal of Advanced Computer Science and Applications
Predicting the Intention of Online Shoppers' Purchasing	Computational modeling, web pages, predictive models, real-time systems, robustness, behavioral sciences, electronic commerce.	(Sang y Wu, 2022)	2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)
An intuitionistic fuzzy data-driven product ranking model using sentiment analysis and multi-criteria decision-making	Online reviews, product ranking, sentiment analysis, multi-criteria decision-making, intuitionistic fuzzy sets, IF-IDOCRIW.	(Heidary et al., 2021)	Technological Forecasting and Social Change
User Value Identification Based on Improved RFM Model and k-Means++ Algorithm for Complex Data Analysis	User value identification, RFM model, K-means++ algorithm, complex data analysis, e-commerce.	(Wu et al., 2021)	Wireless Communications and Mobile Computing

<b>Título</b>	<b>Palabras clave</b>	<b>Autores / año</b>	<b>Revista</b>
M-GAN-XGBOOST model for sales prediction and precision marketing strategy making of each product in online stores	Prediction Model, marketing strategy, marketing strategies, digital marketing, content type, sales data, adaptive prediction, proposed model, sales prediction, online stores.	(Wang y Yang, 2021)	Data Technologies and Applications
The Application of Machine Learning in Online Purchasing Intention Prediction	User value identification, neural network, e-commerce, user behavior mining, user classification, marketing cost.	(Shi, 2021)	ICBDC '21: Proceedings of the 6th International Conference on Big Data and Computing
Recommending Products by Fusing Online Product Scores and Objective Information Based on Prospect Theory	Automobiles, psychology, internet, probabilistic logic, linguistics, business, economics.	(Song et al., 2020)	IEEE Access
Deep Learning Based Prediction Model for the Next Purchase	Time series analysis, deep learning, prediction, e-commerce.	(Utku y Akcayol, 2020)	Advances in Electrical and Computer Engineering
Improving The Effectiveness of Classification Using the Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction	Classification, online shoppers, purchasing intention, data level approach, feature selection.	(Kurniawan et al., 2020)	Journal of Physics: Conference Series
A Stacking Ensemble of Multi-Layer Perceptrons to Predict Online Shoppers' Purchasing Intention	Classification algorithms, stacking, neurons, support vector machines, random forests, predictive models, prediction algorithms.	(Mootha et al., 2020)	2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)
Real-Time Prediction of	Real-time online	(Baati y Mohsil,	Artificial

Título	Palabras clave	Autores / año	Revista
Online Shoppers' Purchasing Intention Using Random Forest	shopper behavior, marketing offers in online stores, random forest.	2020)	Intelligence Applications and Innovations
Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks	Online shopper behavior, shopping cart abandonment, clickstream data, deep learning.	(C. O. Sakar et al., 2019)	Neural Computing and Applications
Predicting overall customer satisfaction: big data evidence from hotel online textual reviews	Customer satisfaction, hotel reviews big data, textual analysis, natural language processing.	Zhao et al. (2019)	International Journal of Hospitality Management

\*Nota: Las palabras clave se transcriben en inglés, respetando su forma original en los textos citados.

### 3.2.1 Descripción de los estudios seleccionados

En el ámbito del comercio electrónico se identifica como un problema significativo la baja tasa de conversión de compras. Es decir, el porcentaje de visitantes del sitio que realizan una compra en comparación con el total de visitantes. Abordando esta problemática, investigadores como Kim et al. (2024) exploraron esta cuestión enfocándose en dos factores clave: las características históricas del cliente y su comportamiento en la navegación en Internet. Los autores, utilizaron la técnica de mercadeo denominada análisis RFM (Recencia, Frecuencia y Monetario), estos autores identificaron variables críticas relacionadas con las características de los clientes. Al desarrollar un análisis predictivo del comportamiento de compra, encontraron que la combinación de características del cliente y patrones de navegación mejora significativamente la precisión en la predicción de compras.

Otro enfoque interesante es el de Wang (2024), que propone un método de predicción mediante la minería de datos de usuarios. Este estudio compara el desempeño alcanzado con el algoritmo ID3, utilizando como criterio el valor F1, el cual mide la precisión y la sensibilidad del modelo para equilibrar el reconocimiento de instancias positivas y negativas.

Liu et al. (2024) proponen un método adaptativo que selecciona la combinación óptima de estrategias para mejorar la calidad de los datos y el rendimiento del modelo, incluyendo la eliminación de características irrelevantes, el relleno de valores faltantes y el equilibrio de clases en la clasificación. Este enfoque permite ajustar el modelo para manejar de manera más efectiva las variaciones y complejidades de los datos. Los resultados destacan que el método de Bosques Aleatorios, en particular, sobresale en general, mostrando una capacidad superior para manejar el desequilibrio de clases en los datos, lo cual es un desafío común en la clasificación real.

Abdullah-All-Tanvir et al. (2023) desarrollan un método predictivo enfocado en sitios web de comercio electrónico, con el objetivo de identificar a potenciales compradores para ofrecerles incentivos. Este enfoque se fundamenta en la aplicación de técnicas de aprendizaje de máquina, entre las cuales se destacó el uso del clasificador XGBoost. En este caso, XGBoost se emplea para distinguir entre potenciales compradores y visitantes con baja probabilidad de compra, basándose en patrones identificados. La eficacia de este clasificador no solo se mide por su precisión reportada en un 90,65%, sino también por su capacidad para reducir falsos positivos y falsos negativos, mejorando así la identificación de aquellos usuarios más propensos a realizar una compra.

Por su parte, Bansal y Vyas (2023) implementaron técnicas de aprendizaje profundo para analizar un amplio conjunto de datos de un sitio web de comercio electrónico, enfocándose en elementos como el historial de navegación y la demografía del cliente. Mediante el uso de modelo de redes neuronales convolucionales y recurrentes, lograron una precisión del 85% en la identificación de usuarios con potencial de compra.

Sang y Wu (2022) se centraron en analizar los factores que determinan la intención de compra de los consumidores en línea, considerando aspectos como las características demográficas de los usuarios, la información específica de los productos y los patrones de comportamiento en Internet. Emplearon la regresión logística como método predictivo para evaluar el impacto de estos factores en las decisiones de compra, alcanzando una precisión del 85% en sus predicciones. Los resultados subrayan la relevancia de entender en profundidad las características demográficas de los consumidores, los detalles de los productos y las conductas de los usuarios en línea.

Alghanam et al. (2022) propusieron un método basado en la minería de datos, evidenciando que las técnicas de clasificación mediante árboles de decisión pueden predecir con precisión el comportamiento de compra. En una línea similar, Wu et al. (2021) optimizaron el análisis RFM al incorporar la fecha de la última compra como un nuevo indicador, demostrando que esta mejora contribuye efectivamente a la predicción de la intención de compra en línea. Shi (2021) exploró el uso del aprendizaje de máquina, destacando la máquina de soporte vectorial (SVM) por su alta precisión en la predicción de la intención de compra. Del mismo modo, Wang y Yang (2021) y Heidary et al. (2021) identificaron la técnica XGBoost y la combinación de análisis de sentimientos con toma de decisiones de criterios múltiples (MCDM), respectivamente, como métodos altamente efectivos para mejorar la precisión de las predicciones.

En Song et al. (2020) propusieron integrar análisis de sentimientos y evaluaciones cuantitativas basadas en reseñas y calificaciones para afinar la predicción, alcanzando un 80% de precisión. Kurniawan et al. (2020) enfocaron sus esfuerzos en perfeccionar la efectividad de la clasificación mediante técnicas avanzadas de selección de características, logrando una precisión del 93.7%. Mootha et al. (2020) desarrollaron un método de predicción utilizando un conjunto de perceptrones multicapa (MLP), que superó la precisión de los MLP individuales. En Sakar et al. (2019) emplearon redes neuronales recurrentes, demostrando ser superior en términos de exactitud y precisión sobre otros métodos reportados. Finalmente, Baati y Mohsil (2020) aplicaron un clasificador de bosque aleatorio para analizar patrones de navegación y compra, logrando un 90% de precisión en la predicción de la intención de compra.

### 3.2.2 Análisis y discusión

En los años recientes, la investigación enfocada en la predicción de la intención de compra en el comercio electrónico ha ganado relevancia, destacándose por los aportes de investigaciones como las de Kim et al. (2024), Liu et al. (2024) y Wang (2024), las cuales han contribuido a profundizar en el entendimiento del problema de la predicción.

Por su parte, Liu et al. (2024) se centran en abordar los desafíos técnicos asociados a la calidad de los datos, la ausencia de datos y el desequilibrio de clases en los conjuntos de datos de comercio electrónico. Proponiendo un enfoque adaptativo para la selección de técnicas de imputación de datos y de balance de clases, lo cual es fundamental para mejorar la fiabilidad y precisión de la predicción en este ámbito.

En Kim et al. (2024) presentan un enfoque novedoso al integrar análisis detallados sobre las características demográficas del cliente y sus patrones de navegación web. Este estudio revela cómo la combinación de características puede enriquecer significativamente la precisión en la predicción, logrando un mejor entendimiento de las dinámicas que impulsan las decisiones de los consumidores.

Diversos estudios han validado la efectividad de técnicas como el aprendizaje de máquina, mostrando potencial para el desarrollo de métodos predictivos. Estos enfoques se apoyan en datos variados, incluyendo el comportamiento de navegación, la información demográfica y el historial de compras de los usuarios (Alghanam et al., 2022; Sakar et al., 2019; Kurniawan et al., 2020). No obstante, la tarea de predecir la intención de compra sigue siendo un reto significativo debido a la influencia de factores tanto internos como externos al consumidor, así como la variabilidad del contexto en el que se realizan las compras (Gupta et al., 2021; Qi et al., 2020; Sokolova y Kefi, 2020).

Estos retos subrayan la importancia de que un método predictivo no solo gestione las interacciones entre una amplia gama de variables, sino que también se adapte a los cambios dinámicos a lo largo del tiempo. De este modo, la adaptabilidad y la capacidad para generalizar son cruciales para superar las barreras en la precisión y aplicabilidad de estos modelos. Según Tran (2021), Chaudhuri et al. (2021), Chu et al. (2019), Dong y Jiang (2019), Kabir et al. (2019) y Mokryn et al. (2019), enfrentar barreras como la volatilidad de los datos, la complejidad de las relaciones entre variables y la evolución constante de los patrones de comportamiento del consumidor, requiere la

implementación de diversas técnicas. Entre estas técnicas se incluyen desde árboles de decisión y bosques aleatorios hasta métodos de aprendizaje profundo, como las redes neuronales.

A continuación, en la Tabla 3-3, se presenta un resumen de los métodos identificados en la literatura, incluyendo las técnicas aplicadas y las métricas de evaluación usadas.

**Tabla 3-3.** Técnicas de referencia reportadas en la literatura para la predicción

Referencia	Técnica usada	Métricas de evaluación
Liu et al., 2024	Proceso adaptativo para imputación y tratamiento de desbalance	Valor F1, Precisión
Alghanam et al., 2022	Árboles de decisión	Precisión, Sensibilidad, Especificidad
Abdullah-All-Tanvir et al., 2023	XGBoost	Exactitud, TVP, Valor F1
Bansal y Vyas, 2023	Aprendizaje profundo	Tasa de acierto Área bajo la curva ROC (AUC-ROC)
Sang y Wu, 2022	Regresión logística	Exactitud, Precisión, Sensibilidad
Shi, 2021	SVM	Exactitud, TVP, Valor F1
Kurniawan et al., 2020	Bosque aleatorio	Exactitud, AUC-ROC, TVP
Mootha et al., 2020	MLP	Exactitud, TVP, TVN, Valor F1
Baati y Mohsil, 2020	Bosque aleatorio	Precisión, Sensibilidad, Valor F1
Sakar et al., 2019	Red neuronal recurrente	Exactitud, TVP, TVN, Valor F1

La revisión de estudios recientes indica una tendencia hacia el empleo de técnicas de aprendizaje de maquina tales como Árboles de decisión, SVM, Bosques Aleatorios, MLP, XGBoost, entre otros. Adicionalmente, se identifica el uso del conjunto de datos "Intención de Compradores en Línea" (Saka y Kastro, 2018), el cual se encuentra disponible en el UCI Machine Learning Repository (Islam et al., 2023).

## 4. Predicción de la intención de compra de usuarios en línea

Este capítulo aborda la elección del conjunto de datos y el diseño experimental, incluyendo la descripción de técnicas de aprendizaje de máquina seleccionadas con base en estudios previos reportados en la literatura. Se explica la elección de métricas de evaluación y se detalla la implementación en el entorno de ejecución.

### 4.1 Conjunto de datos para la predicción de la intención de compra

Se identificaron tres conjuntos de datos relacionados con la intención de compra en línea, los cuales fueron evaluados para determinar su idoneidad en el contexto de esta investigación. Los conjuntos analizados fueron:

#### **Conjunto de datos 1: Intención de compradores en línea<sup>1</sup>**

Este conjunto de datos, disponible en el UCI Machine Learning Repository, incluye 12,330 sesiones de usuarios únicos con diversas características relacionadas con el comportamiento del usuario en un sitio de comercio electrónico. Entre las características más destacadas se encuentran:

- Administrative: Número de páginas visitadas en la sección administrativa.
- Informational: Número de páginas visitadas en la sección informativa.
- ProductRelated: Número de páginas relacionadas con productos.

---

<sup>1</sup> <https://archive.ics.uci.edu/dataset/468/online+shoppers+purchasing+intention+dataset>

- BounceRates: Porcentaje de visitantes que abandonan el sitio después de ver solo una página.
- ExitRates: Porcentaje de vistas de página que terminan en esa página específica.
- PageValues: Valor promedio de una página web visitada antes de completar una transacción.
- Revenue: Indica si una sesión resultó en una transacción de compra.

La amplitud y profundidad de las características permiten identificar patrones significativos en el comportamiento del usuario, lo que hace que este conjunto de datos sea robusto para predecir la intención de compra.

### **Conjunto de datos 2: Customer propensity to purchase data<sup>2</sup>**

Este conjunto de datos incluye información sobre las interacciones de los clientes en un sitio de comercio electrónico. Las características principales son:

- page\_view: Número de visitas a la página.
- basket\_add\_list: Número de veces que el usuario añadió productos a la cesta.
- promo\_banner\_click: Número de clics en banners promocionales.
- device: Tipo de dispositivo utilizado (móvil, computadora, tablet).
- returning\_user: Si el usuario es recurrente.
- ordered: Si el usuario realizó una compra

Aunque útil para analizar el comportamiento del usuario, este conjunto carece de ciertas características detalladas como las tasas de rebote y salida, que son cruciales para entender la intención de compra.

---

<sup>2</sup> <https://www.kaggle.com/datasets/benpowis/customer-propensity-to-purchase-data>

### **Conjunto de datos 3: E-commerce platform analysis and prediction<sup>3</sup>**

Este conjunto de datos abarca una variedad de características relacionadas con el comportamiento de los usuarios en una plataforma de comercio electrónico. Entre las características principales se encuentran:

- Gender: Género del usuario.
- Age: Edad del usuario.
- City: Ciudad desde donde realiza la compra.
- Payment Method: Método de pago utilizado.
- Order Status: Estado de la orden (completada, pendiente, etc.).
- Product Details: Detalles del producto (categoría, precio, etc.).

Aunque este conjunto de datos ofrece información demográfica y sobre el estado de las órdenes, carece de detalles específicos del comportamiento del usuario en el sitio web, como la duración de la interacción con productos y las tasas de rebote y salida.

#### **Selección del conjunto de datos**

El conjunto de datos "Intención de Compradores en Línea", disponible en el repositorio UCI Machine Learning Repository (Islam et al., 2023), se destaca como la opción más adecuada para esta investigación debido a su enfoque integral en el comportamiento del usuario en el sitio web. Las características detalladas, como las tasas de rebote, tasas de salida y valores de página, permiten un análisis preciso de los factores que influyen en la intención de compra. En comparación, los otros dos conjuntos de datos carecen de algunas características críticas y son menos detallados en aspectos específicos del comportamiento del usuario, lo que los hace menos adecuados para el objetivo de esta investigación.

Adicionalmente, el conjunto de datos Intención de Compradores en Línea (Sakar y Kastro, 2018), se destaca por su aplicación en diversas investigaciones previas (Agustyaningrum et al., 2021; Frazier et al., 2022; Sakar et al., 2019).

---

<sup>3</sup> <https://www.kaggle.com/datasets/smokingkrills/ecommerce-platform-analysis-and-prediction>

Este conjunto comprende 12,330 sesiones de usuarios únicos, de las cuales el 15,5% resultaron en compras efectivas. Se destaca que "Revenue" (ingresos) actúa como una variable objetivo, determinando si hubo o no una transacción de compra por parte del cliente.

A continuación, en la Tabla 4-1, se relacionan las variables del conjunto de datos.

**Tabla 4-1.** Diccionario de variables del conjunto de datos

<b>Nombre variable</b>	<b>Descripción</b>	<b>Tipo de variable</b>
Administrative	Número de páginas visitadas por el usuario en la sección administrativa del sitio web	Entera
Administrative Duration	Tiempo que el usuario pasó en páginas administrativas	Entera
Informative	Número de páginas visitadas por el usuario en la sección informativa	Entera
Information Duration	Tiempo que el usuario pasó en páginas informativas	Entera
Product Related	Número de páginas visitadas por el usuario relacionadas con productos	Entera
Product Related Duration	Tiempo que el usuario pasó en páginas relacionadas con productos	Continua
Bounce Rate	Porcentaje de visitantes que entran al sitio web por esa página y luego abandonan el sitio sin realizar más acciones	Continua
Exit Rate	Porcentaje de vistas de página en el sitio web que terminan en esa página específica	Continua
Page Value	Valor promedio de una página web que un usuario visitó antes de completar una transacción de comercio electrónico	Entera
Special day	Indica si el día en que se dio la sesión se considera un día especial	Entera
Month	Mes en que se registró la sesión	Catagórica
OS	Tipo de sistema operativo utilizado por el usuario	Entera
Browser	Navegador utilizado por el	Entera

<b>Nombre variable</b>	<b>Descripción</b>	<b>Tipo de variable</b>
	usuario para acceder al sitio web	
Region	Región geográfica del usuario	Entera
Traffic Type	Clasificación del tipo de tráfico del usuario	Entera
Visitor Type	Indica si el usuario es nuevo o recurrente	Categórica
Weekend	Indica si la sesión ocurrió durante el fin de semana	Binaria
Revenue	Indica si la sesión resultó en una compra	Binaria

## 4.2 Análisis exploratorio de datos

### 4.2.1 Variables numéricas

El análisis exploratorio inicial de los datos permitió comprender las diversas características (Figura 4-1) que influyen en la intención de compra de los usuarios en línea.

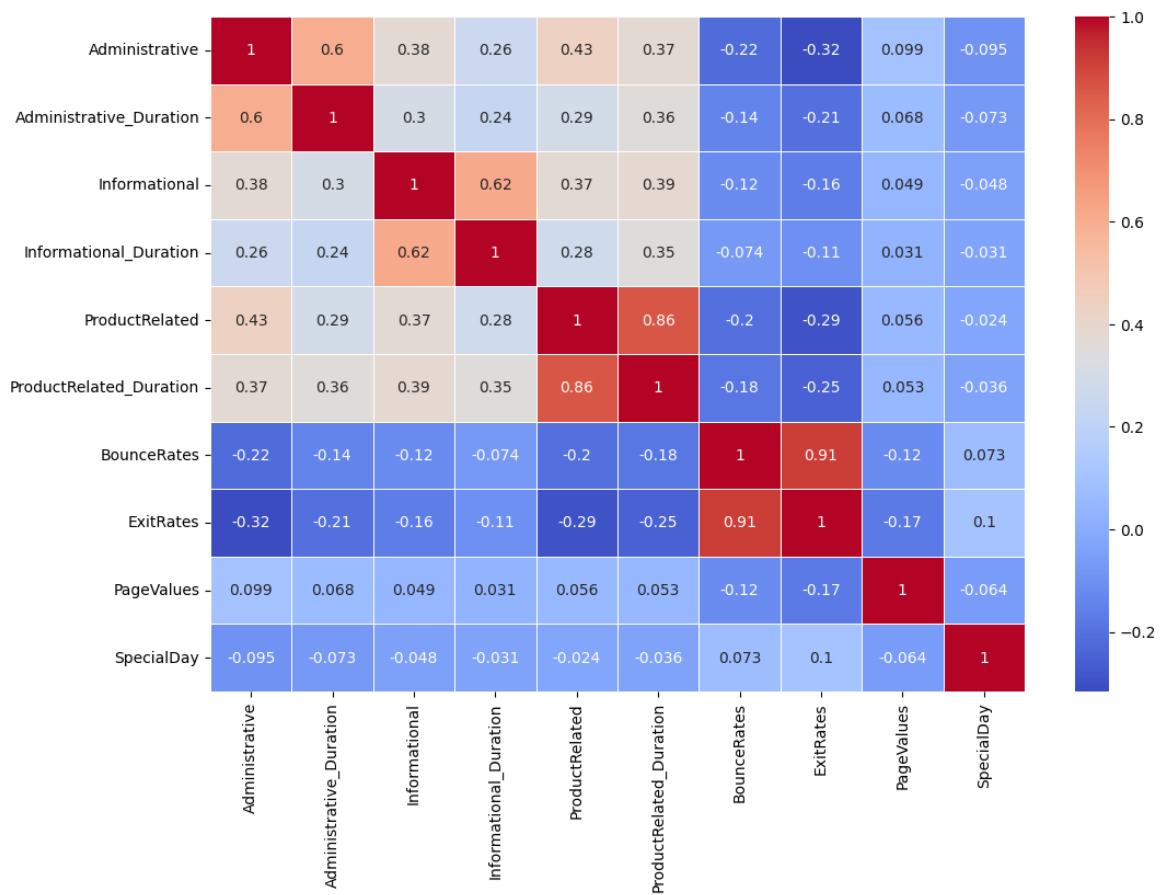
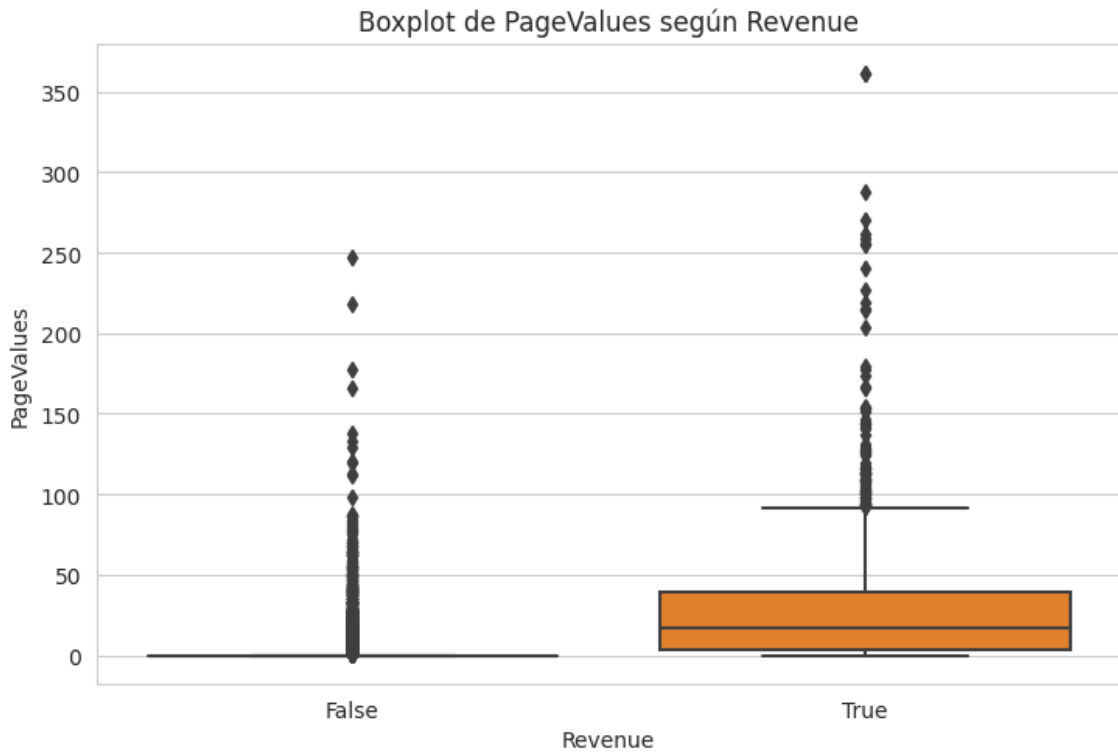


Figura 4-1. Matriz de correlación de variables numéricas

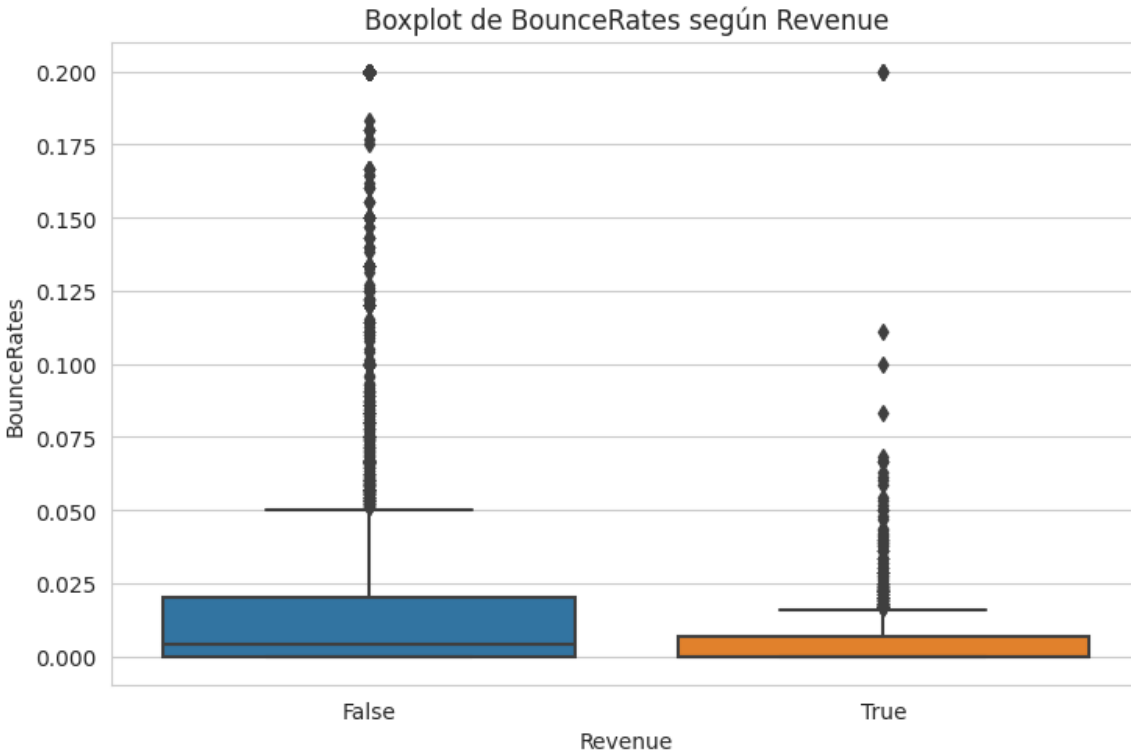
Los hallazgos más destacados en la matriz de correlación son los siguientes:

- Administrative y Administrative\_Duration: Tienen una correlación positiva fuerte (0.6). Esto indica que a medida que aumenta el número de páginas administrativas visitadas, también aumenta el tiempo que los usuarios pasan en estas páginas.
- Informational e Informational\_Duration: Tienen una correlación positiva moderada (0.62). Similarmente, más páginas informativas visitadas se asocian con más tiempo en esas páginas.
- ProductRelated y ProductRelated\_Duration: Muestran una correlación positiva muy fuerte (0.86). Esto sugiere que hay una relación directa entre el número de páginas relacionadas con productos visitadas y el tiempo pasado en estas páginas.
- BounceRates y ExitRates: Tienen una correlación muy fuerte (0.91). Esto indica que las páginas con altas tasas de rebote también tienden a tener altas tasas de salida.
- Administrative y ProductRelated: Correlación positiva moderada (0.43). Los usuarios que visitan más páginas administrativas también tienden a visitar más páginas de productos.
- Administrative y ExitRates: Correlación negativa moderada (-0.32). Las sesiones con más páginas administrativas visitadas tienden a tener menores tasas de salida.
- ExitRates y ProductRelated: Correlación negativa moderada (-0.29). Más páginas relacionadas con productos visitadas tienden a asociarse con menores tasas de salida.



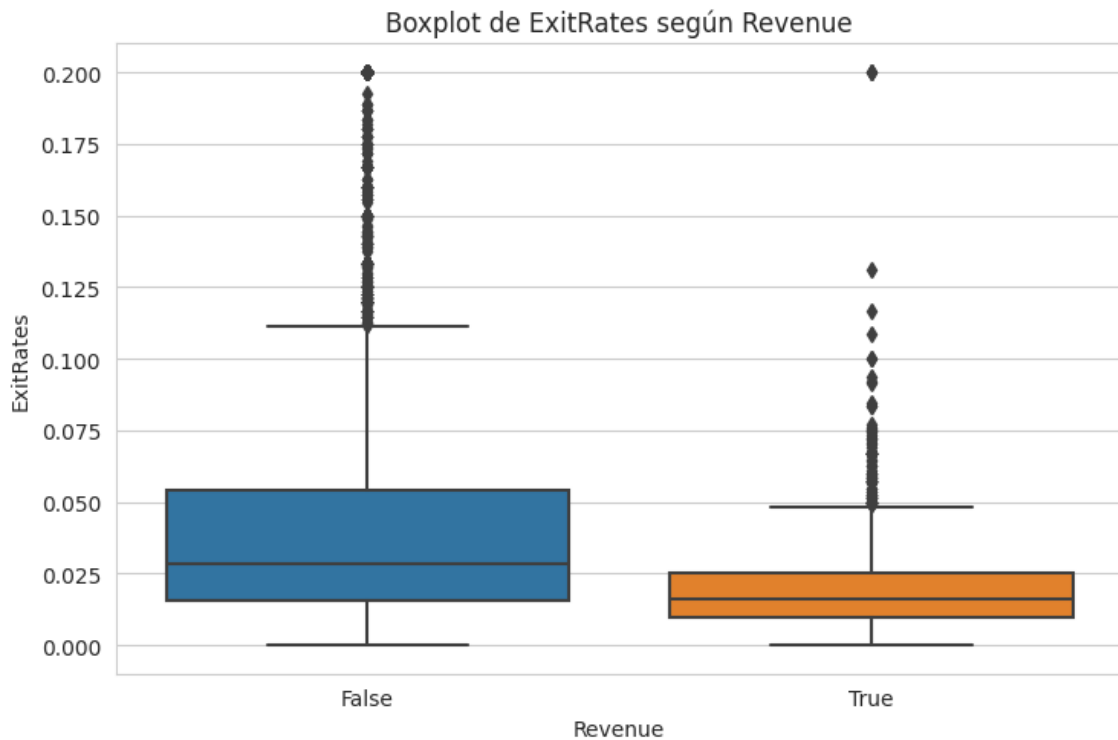
**Figura 4-2.** Diagrama de caja: contribución de página al ingreso

Los valores de página ("PageValues") representan una métrica que cuantifica la relevancia o importancia de una página web dentro de un sitio, basándose en factores como el tiempo de permanencia, la interacción del usuario y la frecuencia de visitas. Las conversiones de compra ("Revenue"), por otro lado, indican las ocasiones en que una visita al sitio web culmina en una transacción comercial efectiva. El análisis exploratorio realizado para el conjunto de datos reveló que las sesiones que terminan en una compra presentan, en promedio, valores de página más altos en comparación con aquellas que no resultan en una venta.



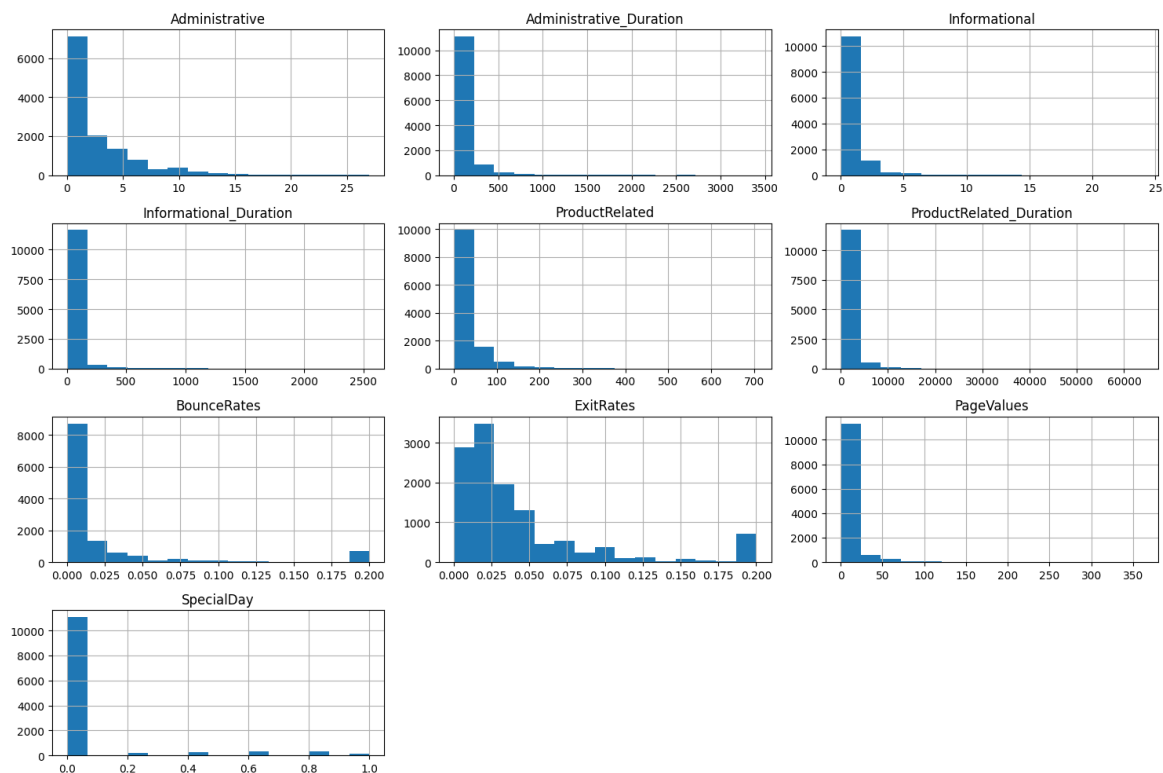
**Figura 4-3.** Diagrama de caja: sesiones con tasa de rebote

La tasa de rebote (BounceRates) se define como el porcentaje de visitas en las que el usuario abandona el sitio desde la página de entrada sin interactuar con ella o navegar a otras páginas. La tasa de salida, por otro lado, mide la frecuencia con la que los usuarios abandonan el sitio desde una página específica después de haber navegado por otras páginas. Esto implica que cuando los usuarios interactúan más con el contenido del sitio y navegan a través de diversas páginas, es más probable que completen una transacción de compra.



**Figura 4-4.** Diagrama de caja: sesiones sin compra

Al analizar detalladamente las ExitRates (Tasas de Salida), se observó una tendencia similar a la de las BounceRates (Tasas de Rebote). Las sesiones que no culminaron en una compra mostraron, en su mayoría, tasas de salida más elevadas. Esto sugiere que un abandono prematuro de la página web puede ser un indicador temprano de una menor probabilidad de conversión a compra.



**Figura 4-5.** Distribución de las variables numéricas

**Administrative:** La mayoría de los usuarios visitan entre 0 y 5 páginas administrativas, con un decrecimiento rápido en la frecuencia a medida que aumenta el número de páginas visitadas.

**Administrative\_Duration:** Similar al número de páginas, la mayoría de los usuarios pasan poco tiempo (menos de 500 segundos) en páginas administrativas, con muy pocos usuarios pasando más tiempo.

**Informational:** La distribución es muy similar a la de las páginas administrativas, con la mayoría de los usuarios visitando entre 0 y 5 páginas informativas.

Informational\_Duration: Nuevamente, la mayoría de los usuarios pasan poco tiempo (menos de 500 segundos) en páginas informativas.

ProductRelated: La mayoría de los usuarios visitan entre 0 y 100 páginas relacionadas con productos, aunque hay algunos usuarios que visitan hasta 700 páginas, lo que muestra una distribución más dispersa en comparación con las páginas administrativas e informativas.

ProductRelated\_Duration: La mayoría de los usuarios pasan menos de 10000 segundos en páginas relacionadas con productos, pero hay algunos que pasan mucho más tiempo, incluso hasta 60000 segundos, indicando una gran variabilidad.

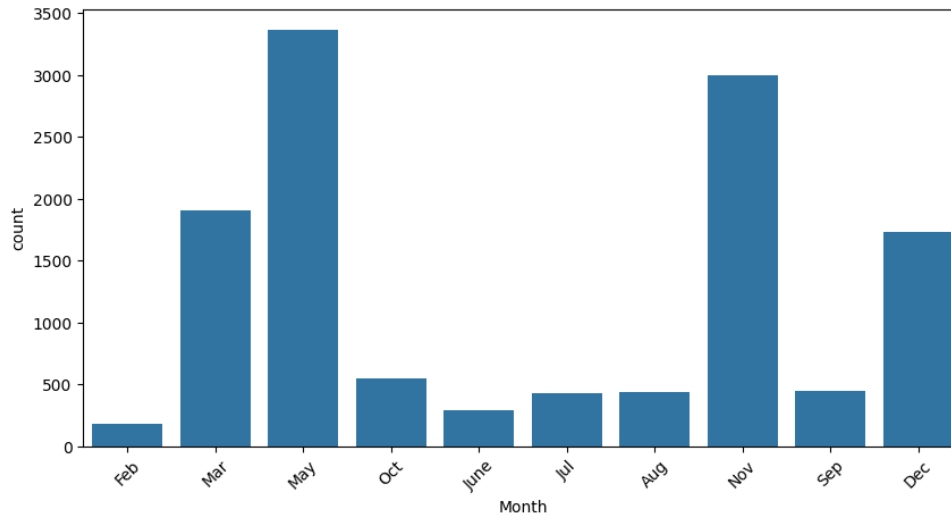
BounceRates: La mayoría de los usuarios tienen una tasa de rebote muy baja (menos de 0.025), con algunas páginas teniendo tasas de rebote más altas, pero en menor cantidad.

ExitRates: La distribución es más uniforme comparada con la tasa de rebote, pero la mayoría de las páginas tienen tasas de salida inferiores al 0.05.

PageValues: La mayoría de las páginas tienen un valor de página muy bajo, con la mayoría de los valores concentrados en menos de 50.

SpecialDay: La mayoría de los valores son 0, lo que indica que la mayoría de las visitas no ocurren cerca de días especiales. Hay algunos valores mayores a 0, representando visitas que ocurren más cerca de días especiales.

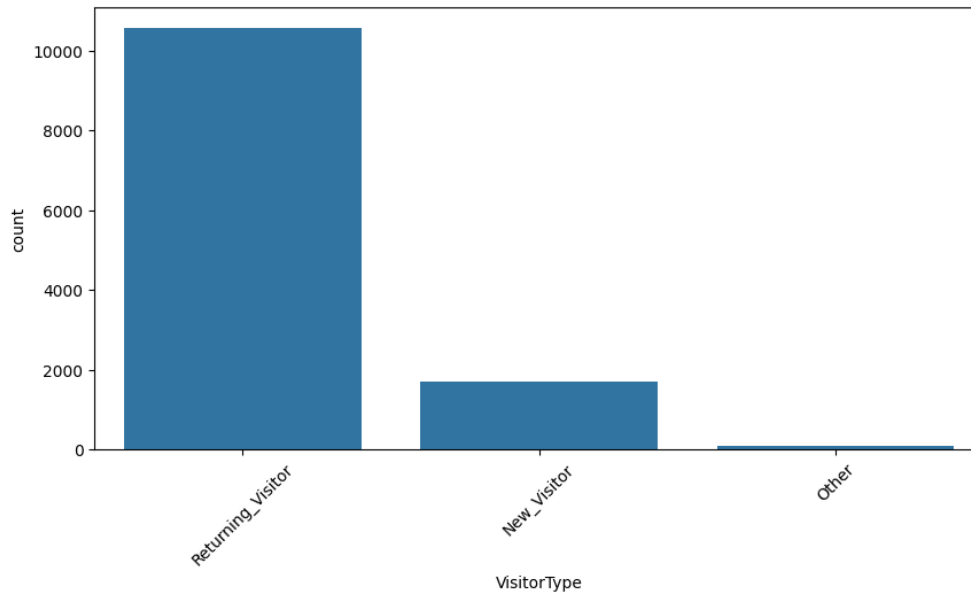
### 4.2.2 Variables categóricas



**Figura 4-6.** Distribución de variables categóricas: Meses

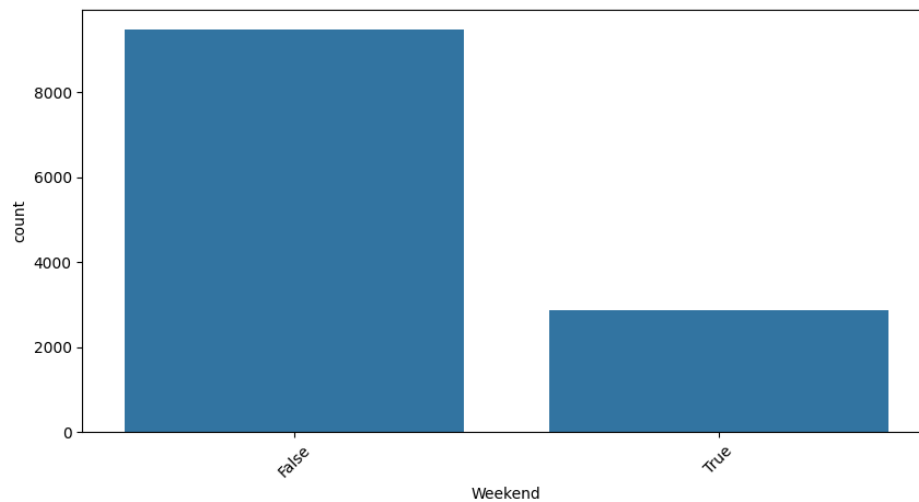
Se observa que los meses de mayo y noviembre tienen la mayor cantidad de sesiones.

La relación con Revenue muestra que noviembre y diciembre tienen una alta proporción de conversiones (ventas), indicando que las visitas en estos meses son más propensas a resultar en una compra.



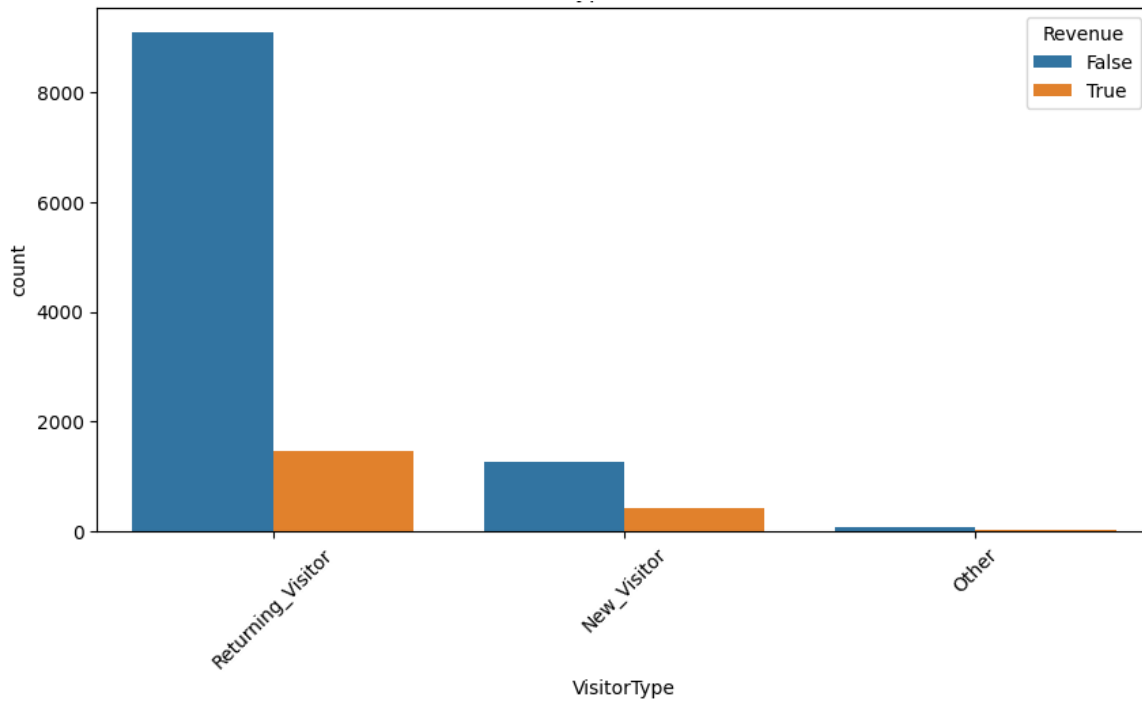
**Figura 4-7.** Distribución de variables categóricas: tipo de visitante

La mayoría de los visitantes son recurrentes. Los visitantes recurrentes tienen una mayor proporción de conversiones en comparación con los nuevos visitantes.



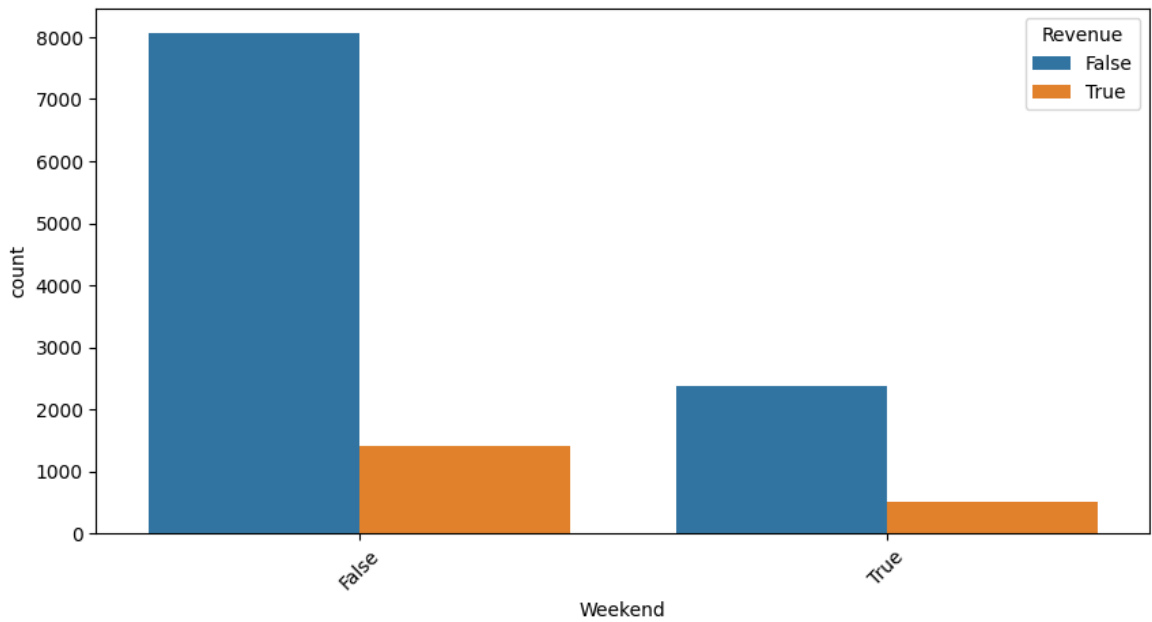
**Figura 4-8.** Distribución de variables categóricas: fines de semana

La mayoría de las sesiones ocurren entre semana. Sin embargo, las sesiones durante el fin de semana (True) tienen una proporción ligeramente mayor de conversiones.



**Figura 4-9.** Distribución de variables categóricas: tipo de visitante vs revenue

La mayoría de los visitantes son recurrentes (Returning\_Visitor). Los visitantes recurrentes tienen una mayor proporción de conversiones en comparación con los nuevos visitantes (New\_Visitor), lo que indica la importancia de atraer y mantener visitantes recurrentes.



**Figura 4-10.** Distribución de variables categóricas: fines de semana vs Revenue

La mayoría de las sesiones ocurren entre semana (False). Sin embargo, las sesiones durante el fin de semana (True) tienen una proporción ligeramente mayor de conversiones en comparación con las sesiones entre semana.

### 4.2.3 Intención de compra vs ejecución de la compra

La intención de compra se define como el propósito o la disposición de un usuario de adquirir un producto o servicio. Es una medida de la inclinación del usuario hacia la compra, que puede no siempre traducirse en una compra real. Por otro lado, la ejecución de compra (Revenue) se refiere a la acción concreta de realizar una transacción comercial efectiva, es decir, cuando un usuario completa el proceso de compra y efectivamente adquiere un producto o servicio.

En esta investigación, debido a la ausencia de una variable explícita de intención de compra en el conjunto de datos, se infiere esta intención a través de patrones de comportamiento observados en las interacciones de los usuarios con el sitio web. A continuación, se detalla cómo se realiza esta inferencia:

- **Interacción prolongada con productos específicos:** Usuarios que pasan más tiempo en páginas relacionadas con productos (ProductRelated\_Duration) y que visitan múltiples páginas de productos (ProductRelated) son considerados como usuarios con una alta probabilidad de tener intención de compra.
- **Frecuencia de visitas a páginas relacionadas con productos:** La repetición de visitas a páginas de productos indica un interés continuo, lo cual puede asociarse con una intención de compra. Número de páginas visitadas (ProductRelated).
- **Comportamientos de navegación:** Variables como BounceRates y ExitRates también proporcionan información sobre la intención de compra. Una baja tasa de rebote y una alta tasa de interacción con varias páginas antes de salir pueden indicar una mayor intención de compra.
- **Días especiales:** La variable SpecialDay indica la cercanía de la visita a días especiales (como el Día de San Valentín), donde es más probable que los

usuarios finalicen sus compras. La intención de compra puede ser inferida por el incremento de visitas y el tiempo pasado en el sitio en estos días especiales.

### 4.3 Desarrollo del método predictivo

Se reconoció la complejidad inherente a la predicción de la intención de compra en línea, caracterizada por la necesidad de analizar patrones de comportamiento del usuario que, aunque no siempre evidentes, son reveladores de sus tendencias y preferencias.

Para comprender y anticipar efectivamente las preferencias del consumidor digital, se realizó una exploración de diversas técnicas de aprendizaje máquina. El objetivo fue identificar la técnica que no solo presentara la mayor precisión en sus predicciones, sino que también demostrara una capacidad para generalizar. Este enfoque garantiza que los resultados obtenidos no se limiten a un conjunto de datos o escenario específico, sino que sean aplicables y útiles en una amplia gama de situaciones en el marco del comercio electrónico.

#### 4.3.1 Técnicas reportadas en la literatura

A continuación, se presenta el análisis de técnicas de aprendizaje de máquina reportadas en la literatura para la predicción de la intención de compra en línea, enfocándose en sus características, fortalezas y limitaciones.

**Regresión logística:** Comúnmente empleada para clasificación binaria, destaca por su simplicidad y la interpretación directa. Es efectiva cuando las características tienen relaciones lineales con la variable objetivo. Sin embargo, puede enfrentar limitaciones en situaciones donde las relaciones entre variables son más complejas, como en interacciones no lineales (Alpaydin, 2020).

**Árboles de decisión:** La técnica no lineal segmenta el espacio de características en distintas regiones. Siendo útil para interpretar y visualizar los factores influyentes en

diversas aplicaciones, pero son susceptibles al sobreajuste en presencia de muchas características (James et al., 2013).

**Bosques aleatorios:** La técnica agrupa múltiples árboles de decisión para mejorar la precisión y controlar el sobreajuste. destacándose frente a variaciones en los datos y ofrece un buen balance entre rendimiento y capacidad interpretativa (Breiman, 2001).

**SVM:** Las SVM buscan el mejor hiperplano, una superficie que separa las clases en el espacio de características, siendo efectivas en espacios de alta dimensión y cuando las clases no están claramente definidas. El rendimiento de las SVM es altamente dependiente de la elección correcta del kernel, una función matemática que transforma los datos para facilitar la separación de clases (Cortes y Vapnik, 1995)

**XGBoost:** Es una técnica de ensamble que construye secuencialmente árboles de decisión, cada uno corrigiendo los errores de los árboles previamente construidos. Ofrece una alta eficacia en una variedad de problemas de clasificación. Sin embargo, su complejidad puede implicar tiempos de entrenamiento más prolongados y requiere una configuración cuidadosa para evitar el sobreajuste (Friedman, 2001).

### 4.3.2 Métricas de evaluación

La selección de métricas para evaluar el rendimiento de las técnicas se basó en el análisis de estudios previos (ver Tabla 3-3) en el capítulo 3. Entre estas métricas, se destacan la precisión, el F1-Score y el AUC-ROC.

Estas métricas proporcionan una evaluación detallada de la eficacia de los métodos predictivos, midiendo tanto la precisión de las predicciones como su habilidad para diferenciar acertadamente entre casos positivos y negativos. A continuación, se detalla el propósito y funcionamiento de cada métrica:

**Accuracy (Precisión):** Indica la proporción de predicciones correctas entre el total de predicciones.

**Recall (Exhaustividad):** Mide la proporción de casos positivos identificados correctamente.

**F1-Score:** Representa la media armónica entre precisión y recall, particularmente esta métrica es valiosa en contextos con clases desbalanceadas.

**AUC-ROC:** Evalúa la capacidad de discriminación entre clases.

Estas métricas facilitan una evaluación multidimensional del rendimiento, permitiendo una comparación y selección informada de la técnica más adecuada para predecir la intención de compra en línea.

### 4.3.3 Implementación

La fase de implementación de las técnicas de predicción de intención de compra en línea se realizó en un entorno controlado y estructurado. Se eligió Python como el lenguaje de programación dada su reconocida eficacia y el extenso soporte que ofrece para el procesamiento de datos y el desarrollo de modelos de aprendizaje automático. Esta decisión se fundamenta en la amplia adopción de Python en la comunidad científica por su flexibilidad y biblioteca de herramientas disponibles (Naik et al., 2022). Para la ejecución y pruebas de los modelos, se utilizó Google Colab, una plataforma que provee un entorno de desarrollo integrado en la nube, y facilita el acceso a recursos computacionales de alto rendimiento.

#### **Entrenamiento y evaluación de técnicas:**

- Inicialmente, se aplicó un preprocesamiento detallado a las variables del conjunto de datos. Las variables categóricas se transformaron mediante la técnica de codificación one-hot, que consiste en representar cada categoría con un vector único en el que solo un elemento es 1 (indicando la presencia de la categoría) y el resto son 0 (indicando la ausencia). Esta técnica, descrita por Harris y Harris, (2015), es efectiva para convertir atributos categóricos en formatos que los modelos de aprendizaje de máquina pueden procesar eficientemente.
- Se exploraron cinco técnicas dada su relevancia y desempeño reportado en la literatura: Regresión Logística, Árboles de Decisión, Bosques Aleatorios, Máquinas de Soporte Vectorial (SVM) y XGBoost.

- Validación cruzada: Durante el ajuste de hiperparámetros, se utilizó validación cruzada para asegurar que la técnica no solo se ajustara bien al conjunto de entrenamiento, sino que también generalizara a nuevos datos.
- Evaluación de técnicas: Cada técnica fue evaluada en un conjunto de prueba independiente. Se calculó la precisión y se generaron curvas ROC. Esto permitió comparar la capacidad discriminativa entre las diferentes técnicas.

Para garantizar que las condiciones de evaluación sean similares a las de una aplicación práctica, el conjunto de prueba fue segregado del proceso de entrenamiento, simulando un escenario real donde se gestionan datos no tenidos en cuenta previamente.

El entorno de ejecución y los scripts de código utilizados para la implementación están disponibles para consulta y uso en el repositorio público:

[GitHub - FelipeOrtiz-Clavijo/prediccionintencioncompra.](https://github.com/FelipeOrtiz-Clavijo/prediccionintencioncompra)

## 5. Resultados y discusión

En este capítulo, se presentan y analizan los resultados obtenidos de la aplicación de diversas técnicas de aprendizaje de máquina para predecir la intención de compra de usuarios en línea. Se incluye una comparativa entre las diferentes técnicas, enfatizando las métricas de rendimiento como accuracy, recall, F1-Score y la precisión, así como una interpretación en términos de aplicabilidad práctica y eficiencia predictiva en el contexto del comercio electrónico.

### 5.1 Evaluación de técnicas

En esta sección, se detalla el desempeño de las técnicas de aprendizaje de máquina exploradas. La Tabla 5-1 presenta una comparativa del rendimiento de las técnicas para predecir la intención de compra de usuarios en línea.

**Tabla 5-1.** Comparativa de rendimiento de técnicas

<b>Técnica</b>	<b>Precision</b>	<b>Recall</b>	<b>F1-score</b>	<b>Accuracy</b>
Regresión logística	76%	35%	48%	87.3%
Árbol de decisión	57%	59%	58%	85.8%
Bosques aleatorios	75%	54%	63%	89.3%
SVM	75%	46%	57%	88.4%
Gradient Boosting	71%	58%	64%	89.1%

La regresión logística muestra una alta precisión, pero un recall relativamente bajo, lo que indica que, aunque la técnica es buena identificando sesiones de compra

verdaderas, tiende a no detectar muchas de ellas. El árbol de decisión presenta un equilibrio más cercano entre precisión y recall, aunque ambos valores son moderados.

Los Bosques Aleatorios y el Gradient Boosting demuestran ser las técnicas más equilibradas y efectivas, y la mejor precisión general, lo que sugiere un mejor rendimiento global. La SVM tiene una precisión comparable a los Bosques Aleatorios, pero un recall y un F1-score más bajos, lo que implica una menor eficacia en identificar todas las sesiones de compra positivas.

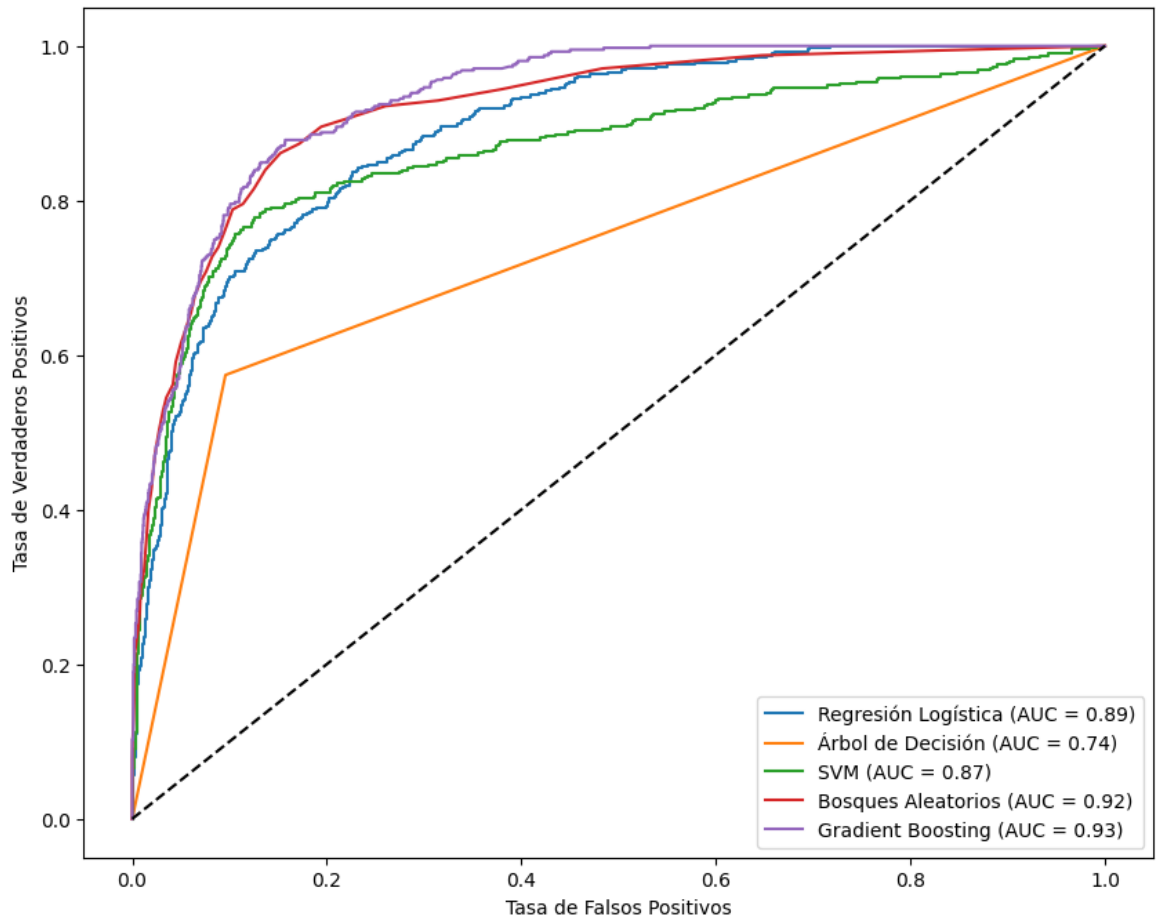
Es relevante mencionar que, más allá de las métricas individuales, la elección de la técnica adecuada debe considerar el contexto específico de aplicación, incluyendo la naturaleza del conjunto de datos, las exigencias de interpretación del método y las limitaciones de tiempo y recursos.

Las técnicas de Bosques Aleatorios y Gradient Boosting tienen un rendimiento superior en comparación con otras técnicas. Ambas muestran un buen equilibrio entre precisión y recall, reflejado en sus F1-scores más altos. La selección entre estos dos dependerá de factores como la importancia relativa de la precisión frente al recall, el tiempo de entrenamiento y la necesidad de interpretabilidad. Bosques Aleatorios puede ser preferible por su mayor interpretabilidad, mientras que Gradient Boosting puede ser más adecuado si se prioriza maximizar el rendimiento.

Por otro lado, la precisión sigue siendo un indicador valioso, especialmente en escenarios donde los falsos positivos conllevan costos significativos o impactos negativos en la experiencia del usuario.

En la Figura 5-1 se muestran las curvas ROC, donde cada una ilustra la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos, ajustada a distintos umbrales de decisión (Bouza-Herrera, 2021). Este enfoque es particularmente útil para valorar la capacidad discriminativa de las técnicas.

### Capacidad discriminativa



**Figura 5-1.** Comparativa de la capacidad discriminativa de las técnicas

La línea diagonal punteada representa el desempeño de un clasificador aleatorio. Cualquier técnica con una curva por encima de esta línea (área bajo la curva AUC) tiene capacidad discriminativa.

A continuación, se presenta un análisis general de los resultados obtenidos de la curva ROC:

- Gradient Boosting (AUC = 0.93): Muestra una excelente capacidad para clasificar correctamente con una alta tasa de verdaderos positivos y una baja tasa de falsos positivos.
- Bosques Aleatorios (AUC = 0.92): Aunque ligeramente inferior a Gradient Boosting, ofrece ventajas significativas, especialmente en contextos con un gran número de variables de entrada.
- Regresión Logística (AUC = 0.89): Aunque no alcanza los niveles de Gradient Boosting o Bosques Aleatorios, sigue siendo una opción competente.
- SVM (AUC = 0.87): Con un rendimiento similar a la Regresión Logística, SVM muestra un buen equilibrio entre la identificación de verdaderos positivos y la minimización de falsos positivos.
- Árbol de Decisión (AUC = 0.74): Presenta un rendimiento más bajo, lo que se refleja en su menor capacidad para discriminar entre clases.

## **5.2 Método para la predicción de la intención de compra de los usuarios en línea**

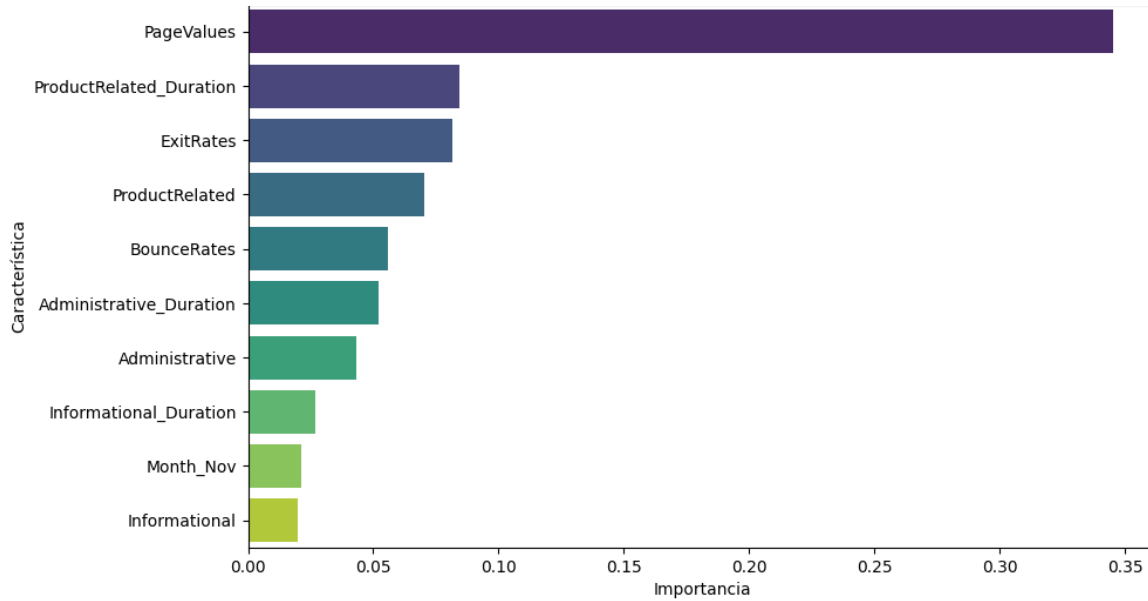
### **5.2.1 Fundamentación del método predictivo**

Esta investigación desarrolla un método para predecir la intención de compra en línea, fundamentándose en el análisis de diversas técnicas y seleccionando específicamente los Bosques Aleatorios y XGBoost por su rendimiento similar en la métrica AUC, con valores de 0.93 y 0.92, respectivamente. Este enfoque busca aprovechar las fortalezas de ambas técnicas para mejorar la precisión en la predicción de comportamientos de compra en el entorno digital.

XGBoost se distingue por su mayor precisión, demostrada en su capacidad para identificar de forma acertada las instancias positivas y negativas, un aspecto crucial para la predicción fiable de la intención de compra en línea, como señala el estudio de Yang et al. (2022). A pesar de que el AUC de los Bosques Aleatorios es ligeramente menor, esta técnica sobresale en escenarios con una gran cantidad de variables de entrada, típicos del comercio electrónico.

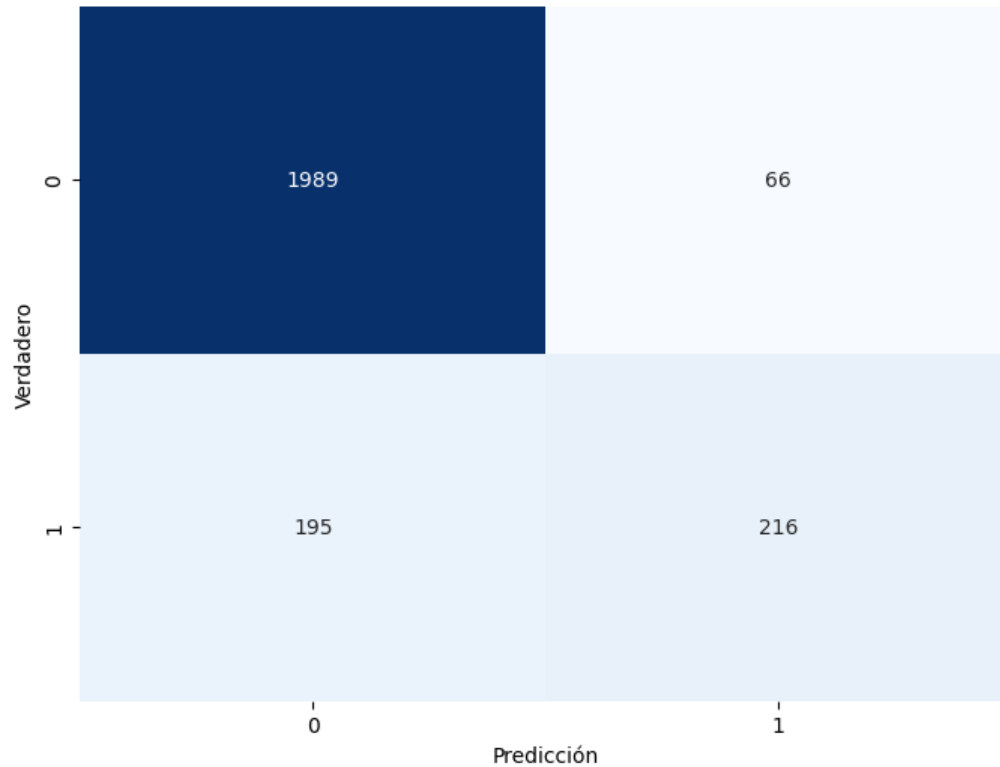
La habilidad de los Bosques Aleatorios para realizar la selección de características de manera eficiente, incluso frente a un volumen elevado de variables, contribuye significativamente a la mejora de la precisión y el desempeño en la clasificación, según lo reportado por Lilhore et al. (2021).

## 5.2.2 Técnica de Bosques Aleatorios



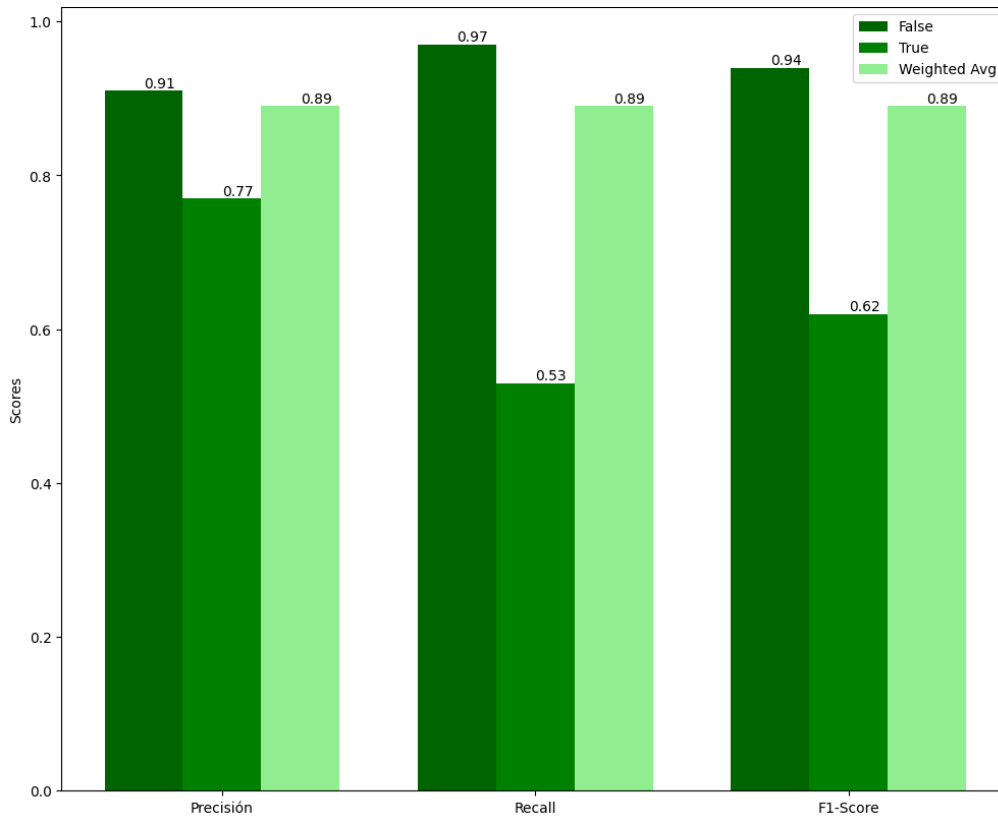
**Figura 5-2.** Características principales para Bosques Aleatorios

Se han identificado diez características clave, siendo los valores de página (PageValues) la más sobresaliente, seguida por la duración de la interacción del usuario con productos relacionados (ProductRelated\_Duration) y las tasas de salida (ExitRates). Este hallazgo subraya la importancia de las páginas visitadas, el tiempo invertido en productos relevantes y las tasas de salida para determinar la probabilidad de que un usuario efectúe una compra. La significancia de estas características resalta cómo la interacción del usuario con el sitio web y la calidad de su experiencia son factores determinantes para predecir la intención de compra.



**Figura 5-3.** Matriz de confusión – Técnica Bosques Aleatorios

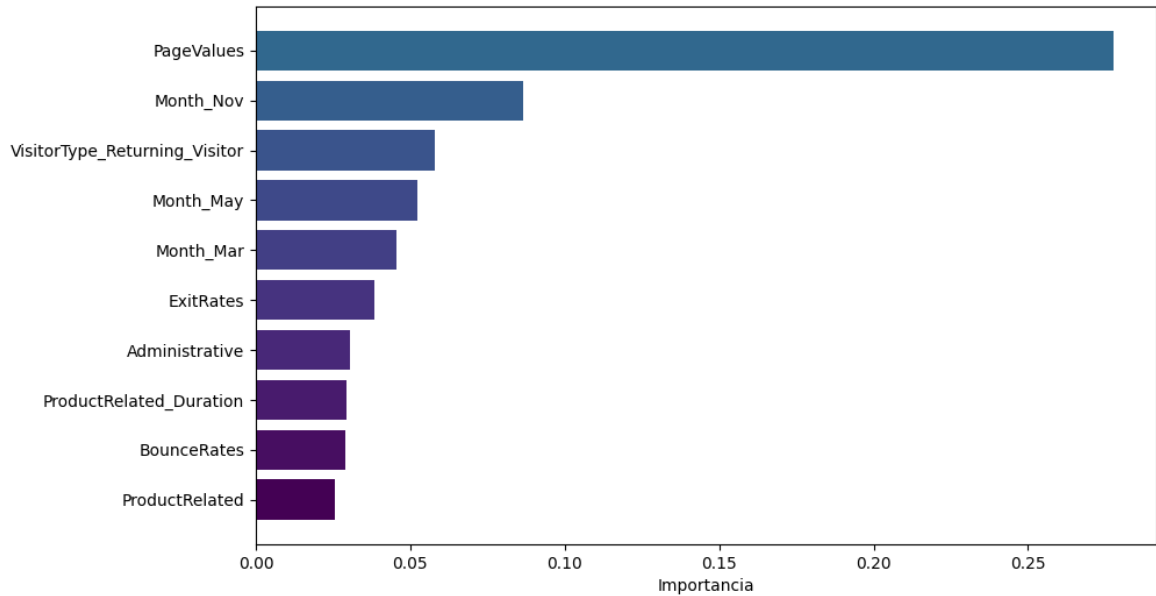
La matriz de confusión presentada en la figura anterior ilustra la capacidad de la técnica para clasificar correctamente las instancias de prueba en dos categorías: usuarios que realizaron una compra (1) y usuarios que no la realizaron (0). Los valores en la diagonal principal (verdaderos positivos y verdaderos negativos) indican el número de predicciones correctas para cada clase, mientras que los valores fuera de esta diagonal representan los errores (falsos positivos y falsos negativos).



**Figura 5-4.** Resultados métricas de evaluación para Bosques Aleatorios

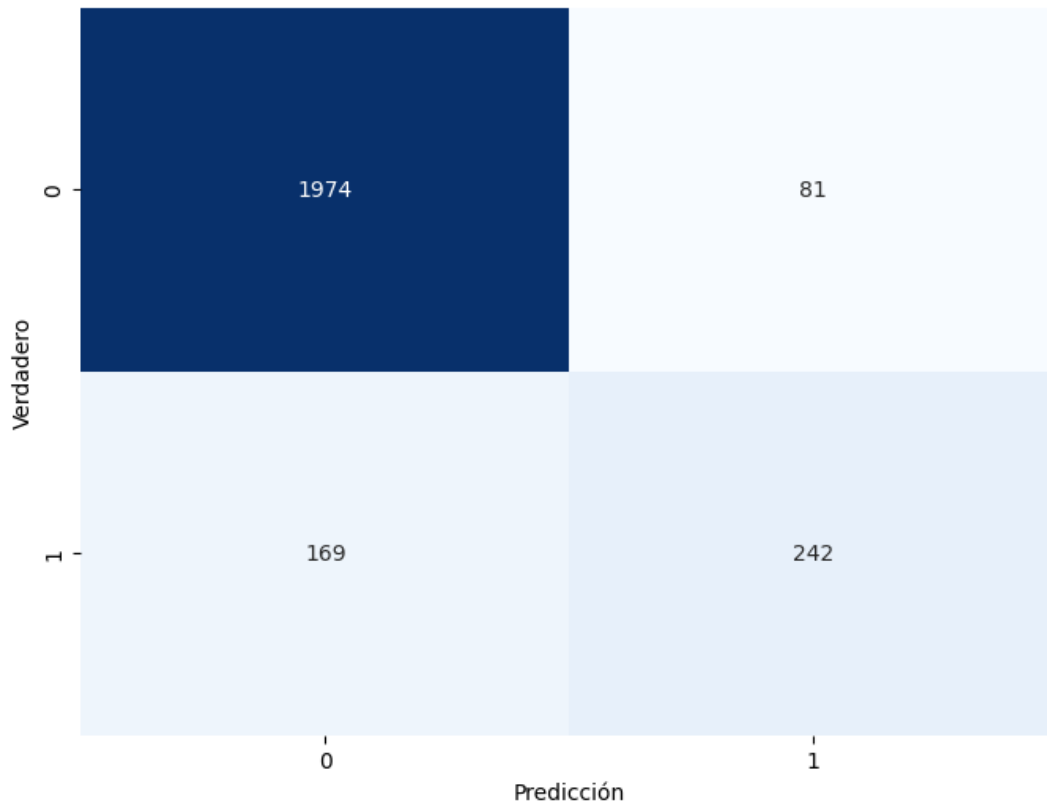
En la Figura 5-4 se compara la precisión, el recall y el F1-score de dos clases, "False" y "True". La clase "False" muestra una alta efectividad con una precisión del 91% y un recall del 97%, reflejado en un F1-score del 94%. Por otro lado, la clase "True" presenta una precisión del 77% y un recall más bajo del 53%, con un F1-score del 62%, indicando que el modelo es menos eficaz para detectar esta clase. El promedio ponderado de las métricas supera el 89%, lo que sugiere un rendimiento general bueno del modelo. La técnica alcanzó una precisión global de 89.42%.

### 5.2.3 Técnica XGBoost



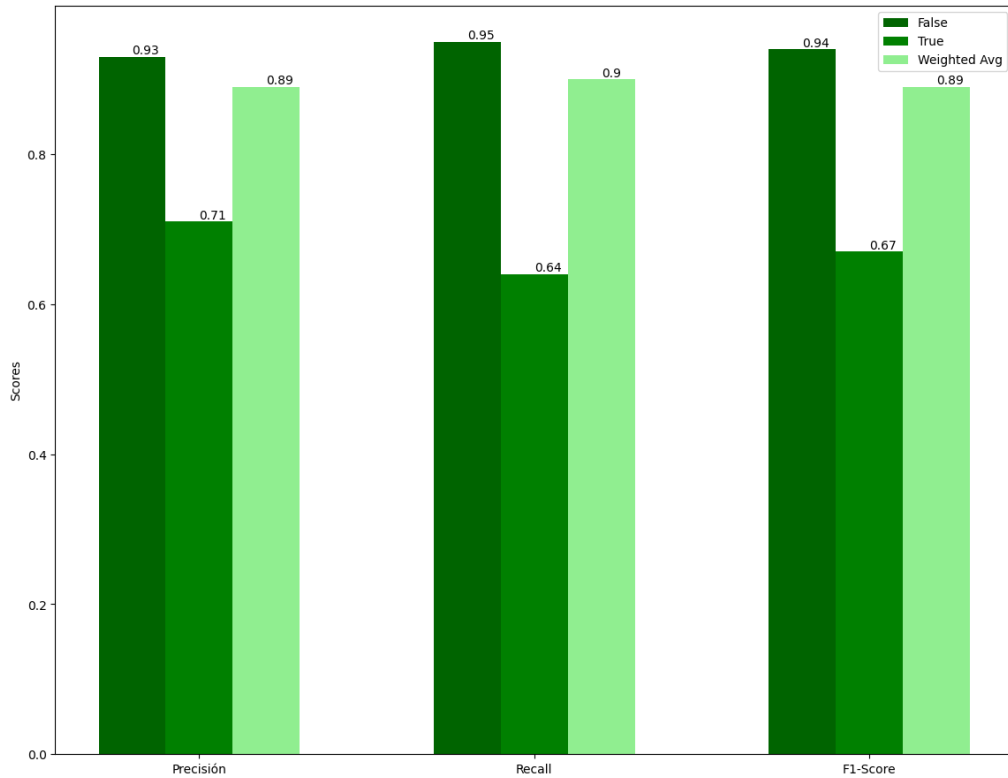
**Figura 5-5.** Características principales para XGBoost

Se identifican diez variables principales. La variable más destacada es los valores de página (PageValues), seguida por la variable mes (Month) y el visitante que regresa (Returning Visitor). Este resultado confirma la importancia de las páginas visualizadas y el tiempo dedicado a productos relacionados.



**Figura 5-6.** Matriz de confusión – Técnica XGBoost

La matriz de confusión ilustra la capacidad de la técnica para clasificar correctamente las instancias de prueba en dos categorías: usuarios que realizaron una compra (1) y usuarios que no la realizaron (0).



**Figura 5-7.** Resultados métricas de evaluación para XGBoost

La Figura 5-7 se compara la precisión, recall y F1-score de dos clases "False" y "True". Se observa que la clase "False" exhibe un rendimiento sobresaliente con una precisión del 93% y un recall del 95%, lo que se traduce en un F1-score del 94%. Esto demuestra la alta eficiencia del modelo en la identificación correcta de instancias negativas. En contraste, la clase "True" registra una precisión del 71% y un recall del 64%, con un F1-score del 67%, indicando una eficacia menor en la detección de instancias positivas. La precisión, recall y F1-score son del 89%, 90% y 89%, respectivamente, lo que evidencia un desempeño general favorable del modelo. De este modo, la técnica logra una precisión global del 89.57%.

### 5.2.4 Definición del método predictivo: ensamble de técnicas

Después de evaluar las técnicas de Bosques Aleatorios y XGBoost, se realizaron ajustes adicionales para refinar el método predictivo.

#### Ajustes en la técnica Bosques Aleatorios

##### Selección de hiperparámetros:

- Utilizando una búsqueda en cuadrícula (Grid Search) con validación cruzada (Dhali et al., 2020), se probaron un total de 108 combinaciones de hiperparámetros.
- Los hiperparámetros ajustados incluyeron el número de árboles (`n_estimators`), la profundidad máxima de los árboles (`max_depth`), el número mínimo de muestras por hoja (`min_samples_leaf`), y la fracción de características a considerar en cada división (`max_features`).

##### Implementación:

- Se utilizó la función `GridSearchCV` de `scikit-learn` para realizar la búsqueda en cuadrícula y la validación cruzada.
- El conjunto de datos se dividió en conjuntos de entrenamiento y prueba utilizando una proporción de 70:30.
- Se evaluó el rendimiento utilizando las métricas de precisión, recall y F1-Score.

#### Ajustes en a técnica XGBoost

##### Selección de hiperparámetros:

- Similar a los Bosques Aleatorios, se realizó una búsqueda en cuadrícula para ajustar los hiperparámetros del modelo XGBoost.
- Los hiperparámetros ajustados incluyeron la tasa de aprendizaje (`learning_rate`), el número de estimadores (`n_estimators`), la profundidad máxima (`max_depth`), y la fracción de muestras utilizadas para entrenar cada árbol (`subsample`).

### **Implementación:**

- Se utilizó la función GridSearchCV de scikit-learn en combinación con el modelo XGBClassifier de xgboost.
- El conjunto de datos se dividió en conjuntos de entrenamiento y prueba en una proporción de 70:30.
- Se evaluó el rendimiento utilizando las mismas métricas de evaluación que para los Bosques Aleatorios.

### **Estrategias de ensamblaje y mejoras finales**

#### **Ensamblaje de modelos:**

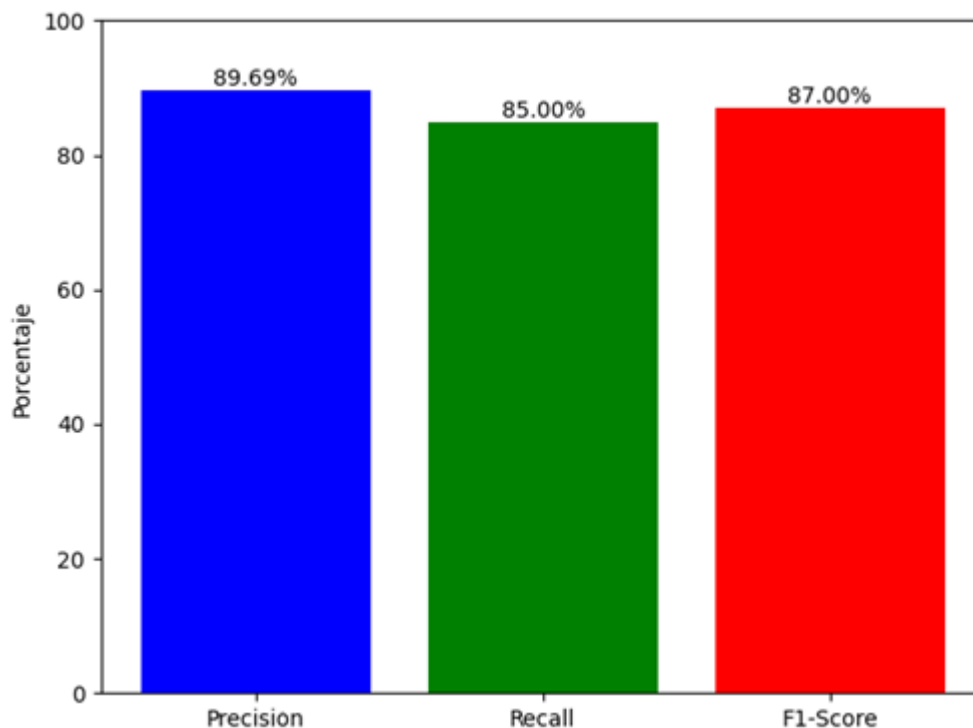
- Se combinó las predicciones de los modelos ajustados de Bosques Aleatorios y XGBoost utilizando técnicas de votación y stacking (Mootha et al., 2020)..
- La técnica de votación ponderada permitió combinar las predicciones de ambos modelos, asignando un peso basado en su rendimiento individual.
- La técnica de stacking implicó el uso de un modelo meta (Logistic Regression) que aprendió de las predicciones de los modelos base (RandomForest y XGBoost).

#### **Implementación del ensamblaje:**

- Para la votación ponderada, se utilizó el VotingClassifier de scikit-learn.
- Para el stacking, se utilizó el StackingClassifier de scikit-learn.

#### **Evaluación de la importancia de las características:**

- Se evaluó la importancia de cada característica dentro del modelo ensamblado, organizándolas en función de su relevancia.
- Las características con menor impacto en las predicciones fueron eliminadas para simplificar el modelo y optimizar su rendimiento.

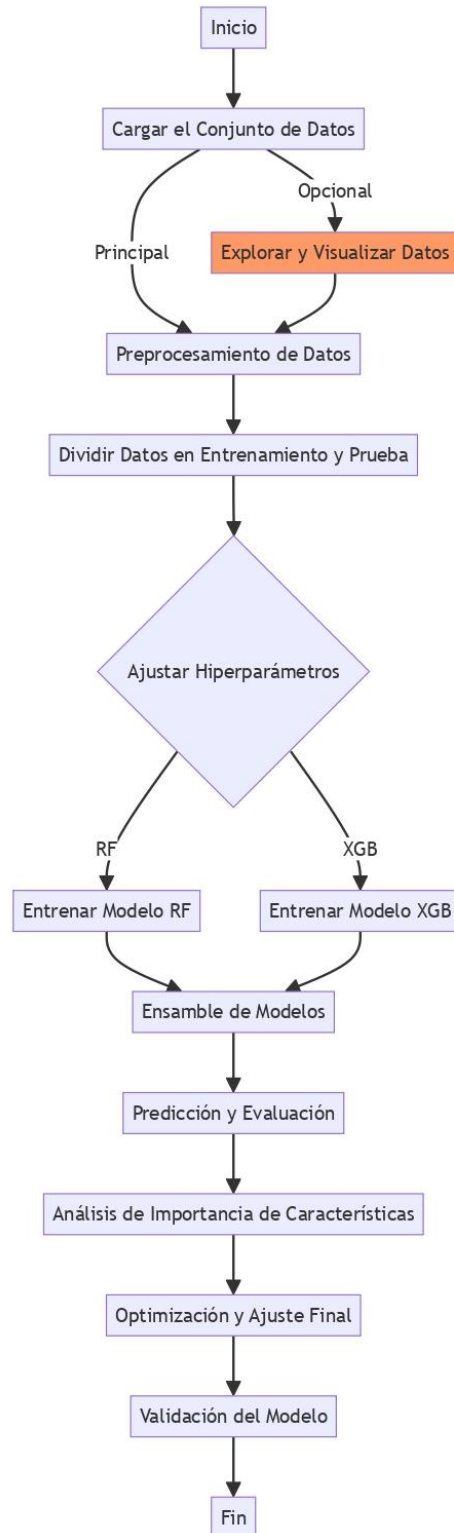
**Resultados:**

**Figura 5-8.** Resultados de métricas de evaluación del método propuesto

En la Figura 5-8 se puede observar que el ensamble de las dos técnicas ofrece un método equilibrado, con una precisión de 89.69%, manteniendo un equilibrio entre las métricas de precisión, sensibilidad y F1-Score para las diferentes clases evaluadas. Esto refleja la capacidad del modelo para identificar correctamente las instancias positivas, siendo un indicador de su eficiencia en la clasificación.

El recall, representado en color verde, es del 85.00%, lo que muestra la proporción de instancias positivas reales que fueron identificadas correctamente por el modelo. Un valor alto en esta métrica sugiere que el modelo es capaz de detectar la mayoría de las instancias positivas.

A continuación, en la Figura 5-9, se presenta el diagrama de flujo del método propuesto.



**Figura 5-9.** Diagrama de flujo del método propuesto

El proceso inicia con la carga del conjunto de datos. A continuación, y de manera opcional, se realiza una exploración inicial mediante técnicas de visualización para analizar la distribución y las relaciones entre variables, facilitando una comprensión preliminar de los datos. La etapa de preprocesamiento implica la limpieza y transformación de los datos, incluyendo la codificación de variables categóricas y la normalización de los datos. Posteriormente, los datos se dividen en conjuntos de entrenamiento y prueba.

Se procede con un procesamiento paralelo para RF y XGB, donde cada técnica recibe un tratamiento específico de escalamiento y procesamiento. El ajuste de hiperparámetros se realiza mediante búsqueda en cuadrícula con validación cruzada para ambos modelos.

El siguiente paso involucra el ensamble de modelos, donde se integran las salidas de los modelos de Bosques Aleatorios (RF) y XGBoost (XGB) para crear un modelo combinado. Este enfoque tiene como meta no solo aumentar la precisión sino también la robustez de las predicciones. Posteriormente, se procede a evaluar la importancia de cada característica dentro del modelo, organizándolas en función de su relevancia. Las características que ejercen menos impacto en las predicciones son eliminadas para simplificar el modelo y optimizar su rendimiento. Finalmente, se efectúa la predicción.

### 5.3 Trabajos reportados en la literatura

En el ámbito de la predicción de la intención de compra, diversos estudios han aportado enfoques prometedores y resultados significativos. A continuación, en la tabla 5-2, se presentan algunos de estos trabajos, todos ellos utilizando el conjunto de datos "Intención de Compradores en Línea".

**Tabla 5-2.** Trabajos similares reportados en la literatura

<b>Autores</b>	<b>Técnica</b>	<b>Precisión</b>
La presente investigación	Ensamble de técnicas (Bosques Aleatorios y XGBoost)	89.69%
Abdullah-All-Tanvir et al. (2023)	XGBoost	90.65%
Kurniawan et al. (2020)	Bosques Aleatorios	93.7%
Baati y Mohsil (2020)	Bosques Aleatorios	90%
Bansal y Vyas (2023)	Aprendizaje Profundo	85%
Sang y Wu (2022)	Regresión Logística	85%

Estos estudios muestran la diversidad de enfoques y técnicas empleadas en la predicción de la intención de compra. Desde el uso de XGBoost y Bosques Aleatorios, destacando su eficacia con precisiones notables, hasta el empleo de aprendizaje profundo y regresión logística para explorar distintos aspectos del comportamiento del consumidor.

Abdullah-All-Tanvir et al. (2023) aplicaron XGBoost en un conjunto de datos de comportamiento de navegación de clientes, logrando una precisión del 90.65%. Wang y Yang (2021) también exploraron el uso de XGBoost, destacando su eficacia con una precisión del 91.5%. Kurniawan et al. (2020) y Baati y Mohsil (2020) mejoraron la clasificación utilizando un clasificador de Bosque Aleatorio, alcanzando precisiones del 93.7% y 90%, respectivamente.

Otros estudios han empleado técnicas diferentes. Bansal y Vyas (2023) resaltaron la influencia del historial de navegación y la demografía del cliente mediante el aprendizaje profundo, alcanzando una precisión del 85%. Sang y Wu (2022) utilizaron regresión logística para identificar factores influyentes, obteniendo una precisión del 85%.

Con el método propuesto, que integra Bosques Aleatorios y XGBoost a través de la búsqueda de hiperparámetros y validación cruzada, no solo se buscó mejorar la precisión sino también favorecer una distribución equilibrada de las clases.

Esta aproximación ofrece una herramienta confiable para la predicción de la intención de compra, poniendo un énfasis particular en la interpretabilidad y la utilidad práctica del método.

## 6. Conclusiones

Este estudio ha desarrollado un método predictivo para la intención de compra en línea, empleando un enfoque integrado que combina las técnicas de aprendizaje de máquina de Bosques Aleatorios y XGBoost. Este método ha logrado una precisión general del 89.69%, demostrando ser un sistema equilibrado entre las métricas de precisión, recall y F1-Score. Este equilibrio resalta la capacidad del modelo no solo para identificar eficientemente las instancias positivas sino también para detectar la mayoría de estas, evidenciado por un recall del 85.00%.

El método propuesto ha permitido abordar de manera efectiva la complejidad del comportamiento de los usuarios en línea, ofreciendo una solución potencialmente escalable, para aplicaciones en el comercio electrónico en tiempo real.

Respecto a los objetivos específicos:

- Selección de un conjunto de datos relevante: Se ha seleccionado un conjunto de datos que incorpora información detallada sobre el comportamiento de compradores en línea, proporcionando una base sólida para el desarrollo y evaluación del método predictivo, cumpliendo con el primer objetivo específico.
- Diseño de un método de predicción: El método propuesto, que integra las técnicas de Bosques Aleatorios y XGBoost, ha sido diseñado y ajustado para identificar con precisión la intención de compra de los usuarios en línea, satisfaciendo el segundo objetivo específico.
- Validación del desempeño del método: El desempeño del método desarrollado ha sido validado y comparado con otros estudios reportados en la literatura. Aunque algunas investigaciones han mostrado tasas de precisión superiores, el método

propuesto se distingue por su equilibrio entre precisión y capacidad de generalización, así como por su adaptabilidad en el ajuste de hiperparámetros, cumpliendo con el tercer objetivo específico.

## 7. Trabajo futuro

A partir de los resultados de esta investigación, se plantean algunos escenarios para futuras investigaciones en el ámbito del comercio electrónico. Destacando:

- Incluir un rango más amplio de comportamientos de compra y perfiles de usuarios. Integrando datos adicionales que capturen patrones de navegación más detallados y las preferencias de compra específicas de los consumidores digitales.
- Reconociendo la importancia de validar el método en diferentes conjuntos de datos, esta tesis sugiere que futuras investigaciones aborden esta extensión. El trabajo futuro puede enfocarse en aplicar y evaluar el método propuesto en diversos conjuntos de datos con mayor cantidad de variables de contexto.
- Se recomienda la consolidación de un conjunto de datos diversificado, un conjunto de datos de esta naturaleza permitiría una comprensión más profunda de los patrones de compra en línea y facilitaría la identificación de tendencias emergentes en el comportamiento del consumidor digital.

## Bibliografía

- Abdullah-All-Tanvir, Ali Khandokar, I., Muzahidul Islam, A. K. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon*, 9(4), e15163. <https://doi.org/https://doi.org/10.1016/j.heliyon.2023.e15163>
- Agustyaningrum, C. I., Haris, M., Aryanti, R., & Misriati, T. (2021). Online shopper intention analysis using conventional machine learning and deep neural network classification algorithm. *Jurnal Penelitian Pos Dan Informatika*, 11(1), 89–100.
- Ahn, T., Ryu, S., & Han, I. (2004). The impact of the online and offline features on the user acceptance of Internet shopping malls. *Electronic Commerce Research and Applications*, 3(4), 405–420. <https://doi.org/10.1016/j.elerap.2004.05.001>
- Alghanam, O. A., Al-Khatib, S. N., & Hiari, M. O. (2022). Data Mining Model for Predicting Customer Purchase Behavior in E-Commerce Context. *International Journal of Advanced Computer Science and Applications*, 13(2). <https://doi.org/10.14569/IJACSA.2022.0130249>
- Alpaydin, E. (2020). *Introduction to machine learning*. MIT press.
- Alsaad, A., & Taamneh, A. (2019). The effect of international pressures on the cross-national diffusion of business-to-business e-commerce. *Technology in Society*, 59, 101158.
- Baati, K., & Mohsil, M. (2020). Real-Time Prediction of Online Shoppers' Purchasing Intention Using Random Forest. In I. Maglogiannis, L. Iliadis, & E. Pimenidis (Eds.), *Artificial Intelligence Applications and Innovations* (pp. 43–51). Springer International Publishing.
- Babenko, V., Kulczyk, Z., Perevosova, I., Syniavska, O., & Davydova, O. (2019). Factors of the development of international e-commerce under the conditions of globalization. *SHS Web of Conferences*, 65, 04016.
- Bansal, M., & Vyas, V. (2023). Analysis and Prediction of Purchase Intention of Online Customers with Deep Learning. In A. Khanna, Z. Polkowski, & O. Castillo (Eds.),

- Proceedings of Data Analytics and Management* (pp. 173–182). Springer Nature Singapore.
- Bawack, R. E., Wamba, S. F., Carillo, K. D. A., & Akter, S. (2022). Artificial intelligence in E-Commerce: a bibliometric study and literature review. *Electronic Markets*, 32(1), 297–338. <https://doi.org/10.1007/s12525-022-00537-z>
- Blackwell, R. D., Miniard, P. W., & Engel, J. F. (2006). *Consumer Behavior*. Thomson South-Western. <https://books.google.com.co/books?id=96TxAAAAMAAJ>
- Bouza-Herrera, C. N. (2021). Las curvas ROC teoría y herramientas para su uso. In *Universidad de La Habana*.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5–32.
- Chaudhuri, N., Gupta, G., Vamsi, V., & Bose, I. (2021). On the platform but will they buy? Predicting customers' purchase behavior using deep learning. *Decision Support Systems*, 149. <https://doi.org/10.1016/j.dss.2021.113622>
- Cheba, K., Kiba-Janiak, M., Baraniecka, A., & Kołakowski, T. (2021). Impact of external factors on e-commerce market in cities and its implications on environment. *Sustainable Cities and Society*, 72, 103032.
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20, 273–297.
- Dhali, S., Pati, M., Ghosh, S., & Banerjee, C. (2020). An Efficient Predictive Analysis Model of Customer Purchase Behavior using Random Forest and XGBoost Algorithm. *2020 IEEE 1st International Conference for Convergence in Engineering (ICCE)*, 416–421. <https://doi.org/10.1109/ICCE50343.2020.9290576>
- Dong, Y., & Jiang, W. (2019). Brand purchase prediction based on time-evolving user behaviors in e-commerce. *Concurrency and Computation: Practice and Experience*, 31(1), e4882. <https://doi.org/10.1002/cpe.4882>
- Du, X., Liu, B., & Zhang, J. (2019). Application of Business Intelligence Based on Big Data in E-commerce Data Analysis. *Journal of Physics: Conference Series*, 1395(1), 012011. <https://doi.org/10.1088/1742-6596/1395/1/012011>
- Esmeli, R., Bader-El-Den, M., & Abdullahi, H. (2021). Towards early purchase intention prediction in online session based retailing systems. *Electronic Markets*, 31(3), 697–715. <https://doi.org/10.1007/s12525-020-00448-x>
- Frazier, A., Maloku, F., Li, X., Chen, Y., Jung, Y., & Zohuri, B. (2022). Data Analysis of Online Shopper's Purchasing Intention Machine Learning for Prediction Analytics.

*Journal of Economics & Management Research*. SRC/JESMR-191. DOI: Doi.Org/10.47363/JESMR/2022 (3), 162, 2–8.

- Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.
- Gupta, S., Nawaz, N., Alfalah, A. A., Naveed, R. T., Muneer, S., & Ahmad, N. (2021). The Relationship of CSR Communication on Social Media with Consumer Purchase Intention and Brand Admiration. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(5), 1217–1230. <https://doi.org/10.3390/jtaer16050068>
- Harris, S., & Harris, D. (2015). *Digital design and computer architecture*. Morgan Kaufmann.
- Heidary, J., Raafat, R., Qorbani, A. R., & Daim, T. (2021). An intuitionistic fuzzy data-driven product ranking model using sentiment analysis and multi-criteria decision-making. *Technological Forecasting and Social Change*, 173, 121158. <https://doi.org/10.1016/j.techfore.2021.121158>
- Huang, Z., & Benyoucef, M. (2015). User preferences of social features on social commerce websites: An empirical study. *Technological Forecasting and Social Change*, 95, 57–72. <https://doi.org/10.1016/j.techfore.2014.03.005>
- Huseynov, F., & Özkan Yıldırım, S. (2019). Online Consumer Typologies and Their Shopping Behaviors in B2C E-Commerce Platforms. *SAGE Open*, 9(2), 215824401985463. <https://doi.org/10.1177/2158244019854639>
- Islam, M. S., Naeem, J., Emon, A. S., Baten, A., Mamun, Md. A. Al, Waliullah, G. M., Rahman, Md. S., & Mridha, M. F. (2023). Prediction of Buying Intention: Factors Affecting Online Shopping. *2023 International Conference on Next-Generation Computing, IoT and Machine Learning (NCIM)*, 1–6. <https://doi.org/10.1109/NCIM59001.2023.10212766>
- Islek, I., & Oguducu, S. G. (2022). A hierarchical recommendation system for E-commerce using online user reviews. *Electronic Commerce Research and Applications*, 52, 101131. <https://doi.org/10.1016/j.elerap.2022.101131>
- James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112). Springer.
- Jimenez, D., Valdes, S., & Salinas, M. (2019). Popularity comparison between e-commerce and traditional retail business. *International Journal of Technology for Business*, 1(1), 10–16.
- Kabir, M. R., Ashraf, F. Bin, & Ajwad, R. (2019). Analysis of Different Predicting Model for Online Shoppers' Purchase Intention from Empirical Data. *2019 22nd International*

- Conference on Computer and Information Technology (ICCIT)*, 1–6.  
<https://doi.org/10.1109/ICCIT48885.2019.9038521>
- Khan, M. M., Sohrab, M. G., & Yousuf, M. A. (2020). Customer gender prediction system on hierarchical E-commerce data. *Beni-Suef University Journal of Basic and Applied Sciences*, 9(1), 10. <https://doi.org/10.1186/s43088-020-0035-7>
- Kim, S., Shin, W., & Kim, H.-W. (2024). Predicting online customer purchase: The integration of customer characteristics and browsing patterns. *Decision Support Systems*, 177, 114105. <https://doi.org/https://doi.org/10.1016/j.dss.2023.114105>
- Kotler, P., & Keller, K. L. (2016). *A Framework for Marketing Management*. Pearson. <https://books.google.com.co/books?id=vv-yoQEACAAJ>
- Kurniawan, I., Abdussomad, Akbar, M. F., Saepudin, D. F., Azis, M. S., & Tabrani, M. (2020). Improving The Effectiveness of Classification Using The Data Level Approach and Feature Selection Techniques in Online Shoppers Purchasing Intention Prediction. *Journal of Physics: Conference Series*, 1641(1), 012083. <https://doi.org/10.1088/1742-6596/1641/1/012083>
- Laudon, K. C., & Traver, C. G. (2013). *E-commerce*. Pearson Boston, MA.
- Lilhore, U. K., Simaiya, S., Prasad, D., & Verma, D. K. (2021). Hybrid Weighted Random Forests Method for Prediction & Classification of Online Buying Customers. *Journal of Information Technology Management*, 13(2), 245–259. <https://doi.org/10.22059/jitm.2021.310062.2607>
- Liu, Y., Li, B., Yang, S., & Li, Z. (2024). Handling missing values and imbalanced classes in machine learning to predict consumer preference: Demonstrations and comparisons to prominent methods. *Expert Systems with Applications*, 237, 121694. <https://doi.org/https://doi.org/10.1016/j.eswa.2023.121694>
- Lu, C.-W., Lin, G.-H., Wu, T.-J., Hu, I.-H., & Chang, Y.-C. (2021). Influencing Factors of Cross-Border E-Commerce Consumer Purchase Intention Based on Wireless Network and Machine Learning. *Security and Communication Networks*, 2021, 9984213. <https://doi.org/10.1155/2021/9984213>
- Mannering, F., Bhat, C. R., Shankar, V., & Abdel-Aty, M. (2020). Big data, traditional data and the tradeoffs between prediction and causality in highway-safety analysis. *Analytic Methods in Accident Research*, 25, 100113.
- Marceda, T., da Silva, W. V., Mendonça Souza, A., Kudlawicz-Franco, C., & da Veiga, C. P. (2020). Online customer behavior: perceptions regarding the types of risks

- incurred through online purchases. *Palgrave Communications*, 6(1), 13. <https://doi.org/10.1057/s41599-020-0389-4>
- Miao, M., Jalees, T., Zaman, S. I., Khan, S., Hanif, N.-A., & Javed, M. K. (2021). The influence of e-customer satisfaction, e-trust and perceived value on consumer's repurchase intention in B2C e-commerce segment. *Asia Pacific Journal of Marketing and Logistics, ahead-of-print(ahead-of-print)*. <https://doi.org/10.1108/APJML-03-2021-0221>
- Micol, L., da Silveira, D. E., da Rosa Righi, R., Antunes Stoffel, R., da Costa, C. A., Victória Barbosa, J. L., Scorsatto, R., & Arcot, T. (2021). Machine learning through the lens of e-commerce initiatives: An up-to-date systematic literature review. *Computer Science Review*, 41, 100414. <https://doi.org/https://doi.org/10.1016/j.cosrev.2021.100414>
- Mokryn, O., Bogina, V., & Kuflik, T. (2019). Will this session end with a purchase? Inferring current purchase intent of anonymous visitors. *Electronic Commerce Research and Applications*, 34, 100836. <https://doi.org/10.1016/j.elerap.2019.100836>
- Mootha, S., Sridhar, S., & Devi, M. S. K. (2020). A Stacking Ensemble of Multi Layer Perceptrons to Predict Online Shoppers' Purchasing Intention. *2020 3rd International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*, 721–726. <https://doi.org/10.1109/ISRITI51436.2020.9315447>
- Naik, P., Naik, G., & Patil, M. (2022). Conceptualizing Python in Google COLAB. *India: Shashwat Publication*.
- Padmavathy, C., Swapana, M., & Paul, J. (2019). Online second-hand shopping motivation – Conceptualization, scale development, and validation. *Journal of Retailing and Consumer Services*, 51, 19–32. <https://doi.org/10.1016/j.jretconser.2019.05.014>
- Pallathadka, H., Ramirez-Asis, E. H., Loli-Poma, T. P., Kaliyaperumal, K., Ventayen, R. J. M., & Naved, M. (2023). Applications of artificial intelligence in business management, e-commerce and finance. *Materials Today: Proceedings*, 80, 2610–2613. <https://doi.org/https://doi.org/10.1016/j.matpr.2021.06.419>
- Pamuksuz, U., Yun, J. T., & Humphreys, A. (2021). A Brand-New Look at You: Predicting Brand Personality in Social Media Networks with Machine Learning. *Journal of Interactive Marketing*, 56, 55–69. <https://doi.org/10.1016/j.intmar.2021.05.001>
- Peña-García, N., Gil-Saura, I., Rodríguez-Orejuela, A., & Siqueira-Junior, J. R. (2020). Purchase intention and purchase behavior online: A cross-cultural approach. *Heliyon*, 6(6), e04284.

- Quintero, J. P. C. (2021). Conectividad de Internet en Colombia y su relación con los Objetivos de Desarrollo Sostenible (2015-2020). *Ciencia y Poder Aéreo*, 16(1), 39–54.
- Rasheed, A., Farhan, M., Zahid, M., Javed, N., & Rizwan, M. (2014). Customer's Purchase Intention of Counterfeit Mobile Phones in Pakistan. *Journal of Public Administration and Governance*, 4(3), 39. <https://doi.org/10.5296/jpag.v4i3.5848>
- Rath, M. (2020). *Machine Learning and Its Use in E-Commerce and E-Business* (pp. 111–127). <https://doi.org/10.4018/978-1-5225-9902-9.ch007>
- Ricci, P. A. G. (2022). Estado del arte de la inteligencia artificial en marketing y el comportamiento del consumidor. *Revista de Ciencias Empresariales | Universidad Blas Pascal*, 7(2022), 60–69. [https://doi.org/10.37767/2468-9785\(2022\)005](https://doi.org/10.37767/2468-9785(2022)005)
- Rico, D. F., Sayani, H. H., & Field, R. F. (2008). History of computers, electronic commerce and agile methods. *Advances in Computers*, 73, 1–55.
- Rosário, A., & Raimundo, R. (2021). Consumer Marketing Strategy and E-Commerce in the Last Decade: A Literature Review. *Journal of Theoretical and Applied Electronic Commerce Research*, 16(7), 3003–3024.
- Sakar, C., & Kastro, Y. (2018). *Online Shoppers Purchasing Intention Dataset*. <https://doi.org/10.24432/C5F88Q>
- Sakar, C. O., Polat, S. O., Katircioglu, M., & Kastro, Y. (2019). Real-time prediction of online shoppers' purchasing intention using multilayer perceptron and LSTM recurrent neural networks. *Neural Computing and Applications*, 31(10), 6893–6908. <https://doi.org/10.1007/s00521-018-3523-0>
- Sang, G., & Wu, S. (2022). Predicting the Intention of Online Shoppers' Purchasing. *2022 5th International Conference on Advanced Electronic Materials, Computers and Software Engineering (AEMCSE)*, 333–337. <https://doi.org/10.1109/AEMCSE55572.2022.00074>
- Shi, X. (2021). The Application of Machine Learning in Online Purchasing Intention Prediction. *Proceedings of the 6th International Conference on Big Data and Computing*, 21–29. <https://doi.org/10.1145/3469968.3469972>
- Sindhu Meena, K., & Suriya, S. (2020). A survey on supervised and unsupervised learning techniques. *Proceedings of International Conference on Artificial Intelligence, Smart Grid and Smart City Applications: AISGSC 2019*, 627–644.
- Sokolova, K., & Kefi, H. (2020). Instagram and YouTube bloggers promote it, why should I buy? How credibility and parasocial interaction influence purchase intentions.

*Journal of Retailing and Consumer Services*, 53, 101742.  
<https://doi.org/10.1016/j.jretconser.2019.01.011>

Solomon, M. R. (2017). *Consumer Behavior: Buying, Having, and Being*. Pearson.  
<https://books.google.com.co/books?id=kYJajwEACAAJ>

Song, Y., Li, G., & Ergu, D. (2020). Recommending Products by Fusing Online Product Scores and Objective Information Based on Prospect Theory. *IEEE Access*, 8, 58995–59006. <https://doi.org/10.1109/ACCESS.2020.2982933>

Suárez, S. J. L. (2020). El comercio electrónico (e-commerce) un aliado estratégico para las empresas en Colombia. *Revista Ibérica De Sistemas e Tecnologías De Informação*, E34, 235–251.

Svobodová, Z., & Rajchlová, J. (2020). Strategic Behavior of E-Commerce Businesses in Online Industry of Electronics from a Customer Perspective. *Administrative Sciences*, 10(4), 78. <https://doi.org/10.3390/admsci10040078>

Tahir, M., Enam, R. N., & Mustafa, S. M. N. (2021). E-commerce platform based on Machine Learning Recommendation System. *2021 6th International Multi-Topic ICT Conference (IMTIC)*, 1–4.

Trivedi, J., & Sama, R. (2020). The Effect of Influencer Marketing on Consumers' Brand Admiration and Online Purchase Intentions: An Emerging Market Perspective. *Journal of Internet Commerce*, 19(1), 103–124.  
<https://doi.org/10.1080/15332861.2019.1700741>

Turban, E., Outland, J., King, D., Lee, J. K., Liang, T.-P., & Turban, D. C. (2018). *Electronic Commerce 2018*. <https://doi.org/10.1007/978-3-319-58715-8>

Utku, A., & Akcayol, M. A. (2020). Deep Learning Based Prediction Model for the Next Purchase. *Advances in Electrical and Computer Engineering*, 20(2), 35–44.  
<https://doi.org/10.4316/AECE.2020.02005>

Viu-Roig, M., & Alvarez-Palau, E. J. (2020). The impact of E-Commerce-related last-mile logistics on cities: A systematic literature review. *Sustainability*, 12(16), 6492.

Wang, P. (2024). Impact of Brand Marketing Strategies Based on Consumer Purchase Intention Mining. *Computer-Aided Design and Applications*, 21(S12), 205–219.  
<https://doi.org/10.14733/cadaps.2024.S12.205-219>

Wang, Q., Cai, R., & Zhao, M. (2020). WITHDRAWN: E-commerce brand marketing based on FPGA and machine learning. *Microprocessors and Microsystems*, 103446.  
<https://doi.org/https://doi.org/10.1016/j.micpro.2020.103446>

- Wang, S., & Yang, Y. (2021). M-GAN-XGBOOST model for sales prediction and precision marketing strategy making of each product in online stores. *Data Technologies and Applications*, 55(5), 749–770. <https://doi.org/10.1108/DTA-11-2020-0286>
- Wu, J., Shi, L., Yang, L., Niu, X., Li, Y., Cui, X., Tsai, S.-B., & Zhang, Y. (2021). User Value Identification Based on Improved RFM Model and -Means++ Algorithm for Complex Data Analysis. *Wireless Communications and Mobile Computing*, 2021, 1–8. <https://doi.org/10.1155/2021/9982484>
- Xiao, L., Guo, F., Yu, F., & Liu, S. (2019). The Effects of Online Shopping Context Cues on Consumers' Purchase Intention for Cross-Border E-Commerce Sustainability. *Sustainability*, 11(10), 2777. <https://doi.org/10.3390/su11102777>
- Zhao, Y., Xu, X., & Wang, M. (2019). Predicting overall customer satisfaction: Big data evidence from hotel online textual reviews. *International Journal of Hospitality Management*, 76, 111–121. <https://doi.org/10.1016/j.ijhm.2018.03.017>
- Zhou, Z. H., & Liu, S. (2021). *Machine Learning*. Springer Nature Singapore. <https://books.google.com.co/books?id=ctM-EAAAQBAJ>
- Zozalbo, F., & Astuti, R. (2022). The Effect of Specific Discount Pattern and Product Type on Customers' Purchase Intention in E-commerce Platform. *Proceedings of the 4th International Conference on Economics, Business and Economic Education Science, ICE-BEES 2021, 27-28 July 2021, Semarang, Indonesia*. <https://doi.org/10.4108/eai.27-7-2021.2316875>