

Richard Edgar Gonzales Alcedo, Jaime Jose Saenz Dedios,
Jesus Javier Cobenas Morales, Jaime Stiven Gonzales Alcedo,
Carlos Alberto Galvez Dioses

Modelos de analítica de datos para la predicción de la intención de compra en entornos digitales

Estudio metodológico comparativo



Administración



Religación
Press

Richard Edgar Gonzales Alcedo, Jaime Jose Saenz Dedios,
Jesus Javier Cobenas Morales, Jaime Stiven Gonzales Alcedo,
Carlos Alberto Galvez Dioses

Modelos de analítica de datos para la predicción de la intención de compra en entornos digitales

Estudio metodológico comparativo

Religación Press

[Ideas desde el Sur Global]

Data analytics models for predicting purchase intention in digital environments.

A comparative methodological study

Modelos de análise de dados para previsão da intenção de compra em ambientes digitais.

Um estudo metodológico comparativo

Religación Press

[Ideas desde el Sur Global]

Equipo Editorial

Editorial team

Ana B. Benalcázar

Editora Jefe / Editor in Chief

Felipe Carrión

Director de Comunicación / Scientific Communication Director

Melissa Díaz

Coordinadora Editorial / Editorial Coordinator

Sarahi Licango Rojas

Asistente Editorial / Editorial Assistant

Consejo Editorial

Editorial Board

Jean-Arsène Yao

Dilrabo Keldiyorovna Bakhronova

Fabiana Parra

Mateus Gamba Torres

Siti Mistima Maat

Nikoleta Zampaki

Silvina Sosa

Victor Ancajima Miñán

.....

Religación Press, es parte del fondo editorial del Centro de Investigaciones CICSHAL-RELIGACIÓN | Religación Press, is part of the editorial collection of the CICSHAL-RELIGACIÓN Research Center |

Diseño, diagramación y portada | Design, layout and cover: Religación Press.

CP 170515, Quito, Ecuador. América del Sur.

Correo electrónico | E-mail: press@religacion.com

www.religacion.com

Disponible para su descarga gratuita en | Available for free download at

<https://press.religacion.com>

Este título se publica bajo una licencia de Atribución 4.0 Internacional (CC BY 4.0)

This title is published under an Attribution 4.0 International (CC BY 4.0) license.



El presente libro tienen el aval del Centro de Investigaciones en Ciencias y Humanidades desde América Latina - CICSHAL.



Título: Modelos de analítica de datos para la predicción de la intención de compra en entornos digitales. Estudio metodológico comparativo

Derechos de autor | Copyright: Richard Edgar Gonzales Alcedo, Jaime Jose Saenz Dedios, Jesus Javier Cobenas Morales, Jaime Stiven Gonzales Alcedo, Carlos Alberto Galvez Dioses

Primera Edición | First Edition: 2026

Editorial | Publisher: Religación Press

Materia Dewey | Dewey Subject: 658 - Gerencia general

Clasificación Thema | Thema Subject Categories: PSPM - Biología marina | RNK - Conservación del medioambiente | IKLSR - Perú | KJS - Ventas y marketing

BISAC: COM064000

Público objetivo | Target audience: Profesional / Académico | Professional / Academic

Colección | Collection: Administración

Soporte | Format: PDF / Digital

Publicación | Publication date: 2026-06-09

ISBN: 978-9907-807-05-9

[APA 7]

Gonzales Alcedo, R. E., Saenz Dedios, J. J., Cobenas Morales, J. J., Gonzales Alcedo, J. S., & Galvez Dioses, C. A. (2026). *Modelos de analítica de datos para la predicción de la intención de compra en entornos digitales. Estudio metodológico comparativo*. Religación Press. <https://doi.org/10.46652/ReligacionPress.427>

Revisión por pares

El presente libro constituye el resultado de un riguroso proceso de investigación académica, cuya calidad metodológica y solidez argumental han sido validadas mediante un sistema de revisión por pares externos implementado bajo el protocolo de doble ciego, bajo la supervisión del Centro de Investigaciones en Ciencias y Humanidades desde América Latina (CICSHAL). Como garantía de transparencia y rigor científico, los informes de evaluación realizados por los especialistas designados se conservan en el archivo institucional de la editorial, a disposición de las instancias que así lo requieran.

Peer Review

This book is the result of a rigorous academic research process, whose methodological quality and argumentative solidity have been validated through an external peer-review system implemented under a double-blind protocol, under the supervision of the Center for Research in Sciences and Humanities from Latin America (CICSHAL). As a guarantee of transparency and scientific rigor, the evaluation reports prepared by the designated specialists are preserved in the publisher's institutional archives, available to any party that may require them.

Sobre los autores

ABOUT THE
AUTHORS

Richard Edgar Gonzales Alcedo

Universidad Tecnológica del Perú (UTP) | Piura | Perú
<https://orcid.org/0009-0001-0047-8827>
c27708@utp.edu.pe
rega1988@gmail.com

Licenciado en Administración, Magister en Gerencia Empresarial y Doctorando en Ciencias Económicas y Financieras. Investigador con publicaciones en economía circular, criptomonedas y gestión pública. Catedrático universitario en la UTP y la UCV, con amplia experiencia en metodología de investigación, gestión empresarial y transformación digital. Su línea de investigación integra la analítica de datos con la gestión estratégica del comercio electrónico en mercados latinoamericanos.

Jaime Jose Saenz Dedios

Universidad Tecnológica del Perú (UTP) | Piura | Perú
<https://orcid.org/0009-0002-3401-5510>
C31105@utp.edu.pe
jsaenzdedios@gmail.com

Especialista en Tecnologías de la Información y Comunicaciones con formación en investigación aplicada. Su labor académica se centra en la formación de profesionales en TIC, integrando teoría, práctica y tecnologías actuales para la mejora de los procesos educativos y organizacionales basados en datos. Con experiencia en modelos de aprendizaje automático aplicados a la gestión de información y toma de decisiones empresariales.

Jesus Javier Cobenas Morales

Universidad Tecnológica del Perú (UTP) | Piura | Perú
<https://orcid.org/0000-0002-2853-1204>
C31976@utp.edu.pe
jjcobenas@engineer.com

Ingeniero Informático con maestría en Administración de Sistemas Informáticos con mención en Gestión y Administración de Datos. Con más de 20 años de experiencia en docencia universitaria, se desempeña como docente académico en la UTP y UCV, formando profesionales en el ámbito de las tecnologías de la información. Especialista en bases de datos, minería de datos y modelos predictivos aplicados a la gestión empresarial y al comercio electrónico.

Jaime Stiven Gonzales Alcedo

Universidad César Vallejo (UCV) | Piura - Sullana | Perú

<https://orcid.org/0000-0001-9104-9093>

p7002505267@ucvvirtual.edu.pe

jaime_561@hotmail.com

Catedrático, funcionario y consultor externo con amplia trayectoria en el ámbito académico y profesional. A través de su experiencia e investigación continua ha contribuido al desarrollo de proyectos de índole académica y empresarial. Su enfoque investigativo se orienta a la demostración de objetivos con rigurosidad y veracidad, abordando temas de gestión, tecnología y desarrollo organizacional.

Carlos Alberto Galvez Dioses

Universidad Tecnológica del Perú (UTP) | Piura | Perú

<https://orcid.org/0000-0001-8127-5150>

C20375@utp.edu.pe

Catedrático y especialista en análisis empresarial y liderazgo corporativo. Licenciado en Administración de Empresas con amplia experiencia en el sector privado peruano en áreas de optimización de procesos comerciales, transformación digital empresarial y desarrollo de estrategias de posicionamiento de marca. Docente comprometido con la formación de profesionales capaces de responder a los retos del entorno digital y competitivo actual.

Resumen

El presente libro examina de manera sistemática y comparativa los principales modelos de analítica de datos aplicados a la predicción de la intención de compra en entornos de comercio electrónico. A partir de una revisión crítica de la literatura indexada en Scopus, Web of Science y SciELO correspondiente al período 2020-2026, el estudio identifica y clasifica ocho modelos predictivos representativos, organizados en tres categorías: modelos estadísticos clásicos (regresión logística, árboles de decisión, SVM), modelos de machine learning avanzado (Random Forest, XGBoost, LightGBM) y modelos de deep learning (ANN, LSTM). Se presenta un estudio comparativo sistemático sobre el benchmark UCI Online Shoppers Purchasing Intention Dataset, evaluando cinco métricas de desempeño. Los resultados confirman la superioridad del modelo LSTM ($AUC-ROC=0.962$) sobre los modelos clásicos, con XGBoost y LightGBM como alternativas de alto desempeño y menor coste computacional. El libro aporta además un diagnóstico estadístico del mercado digital latinoamericano 2020-2026, un marco metodológico replicable para la evaluación comparativa de modelos predictivos, y orientaciones prácticas para la integración ética y efectiva de sistemas predictivos en la gestión del comercio electrónico de la región.

Palabras clave:

Analítica de datos; aprendizaje automático; inteligencia artificial; intención de compra; comercio electrónico.

Abstract

This book systematically and comparatively examines the main data analytics models applied to predicting purchase intention in e-commerce environments. Based on a critical review of indexed literature from Scopus, Web of Science, and SciELO for the period 2020-2026, the study identifies and classifies eight representative predictive models organized into three categories: classical statistical models (logistic regression, decision trees, SVM), advanced machine learning models (Random Forest, XGBoost, LightGBM), and deep learning models (ANN, LSTM). A systematic comparative study is presented using the benchmark UCI Online Shoppers Purchasing Intention Dataset, evaluating five performance metrics. Results confirm the superiority of the LSTM model (AUC-ROC=0.962) over classical models, with XGBoost and LightGBM as high-performance, lower computational-cost alternatives. The book also provides a statistical diagnosis of the Latin American digital market 2020-2026, a replicable methodological framework for comparative model evaluation, and practical guidance for the ethical and effective integration of predictive systems in e-commerce management across the region.

Keywords:

data analytics; machine learning; artificial intelligence; purchase intention; e-commerce.

Resumo

Este livro examina de forma sistemática e comparativa os principais modelos de analítica de dados aplicados a predição da intenção de compra em ambientes de comércio eletrônico. Com base em uma revisão crítica da literatura indexada na Scopus, Web of Science e SciELO para o período de 2020 a 2026, o estudo identifica e classifica oito modelos preditivos representativos organizados em três categorias: modelos estatísticos clássicos, modelos de aprendizado de máquina avançado e modelos de aprendizado profundo. É apresentado um estudo comparativo sistemático usando o benchmark UCI Online Shoppers Purchasing Intention Dataset, avaliando cinco métricas de desempenho. Os resultados confirmam a superioridade do modelo LSTM (AUC-ROC=0,962), com XGBoost e LightGBM como alternativas de alto desempenho e menor custo computacional.

Palavras-chave:

analítica de dados; aprendizado de máquina; inteligência artificial; intenção de compra; comércio eletrônico.

MODELOS DE ANALÍTICA DE DATOS PARA LA PREDICCIÓN DE LA INTENCIÓN DE COMPRA EN ENTORNOS DIGITALES.

Estudio metodológico comparativo

CONTENIDO

Revisión por pares	7
Peer Review	7
Sobre los autores	8
About the authors	8
Resumen	10
Abstract	10
Resumo	11
Prólogo	17

Capítulo 1

<i>Epistemología y fundamentos teóricos de la analítica de datos en entornos digitales</i>	20
La analítica de datos: conceptualización y taxonomía	21
Teorías del comportamiento del consumidor digital: TAM, TPB y UTAUT2	25

Capítulo 2

<i>El comercio electrónico en América Latina: diagnóstico estadístico 2020-2026</i>	28
Evolución del mercado digital latinoamericano	29

Capítulo 3

<i>Modelos predictivos clásicos aplicados a la intención de compra</i>	35
Regresión logística: fundamentos matemáticos y aplicaciones en e-commerce	36

Capítulo 4	<i>Machine learning avanzado: XGBoost, LightGBM y modelos ensamblados</i>	43
	Gradient Boosting y XGBoost: arquitectura, regularización y optimización	44
Capítulo 5	<i>Modelos de deep learning para predicción de comportamiento digital</i>	51
	Redes neuronales artificiales (ANN): arquitectura y entrenamiento	52
Capítulo 6	<i>Estudio comparativo metodológico: evaluación de rendimiento predictive</i>	59
	Diseño del estudio comparativo y protocolo de evaluación	60
Capítulo 7	Implicaciones para la gestión comercial digital y prospectiva 2026-2030	69
	Aplicaciones prácticas en la gestión del comercio electrónico	70
	Referencias	74

TABLAS

Tabla 1. Comparación de modelos teóricos para el análisis de la intención de compra en entornos digitales	25
Tabla 2. Indicadores del mercado de comercio electrónico en países seleccionados de America Latina, 2024	32
Tabla 4. Hiperparámetros optimizados mediante validación cruzada K=5 para modelos de machine learning avanzado en predicción de intención de compra	48
Tabla 6. Resultados comparativos del estudio sistemático de ocho modelos para predicción de intención de compra en comercio electrónico	66

FIGURAS

Figura 1. Evolucion del mercado de comercio electrónico en America Latina, 2018-2025 (ventas en USD bn y usuarios activos en millones)	24
Figura 2. Mapa de calor de indicadores clave de comercio electrónico en países seleccionados de America Latina, 2024	31
Figura 3. Comparación de métricas de evaluación para modelos predictivos clásicos aplicados a la predicción de intención de compra en e-commerce	39
Figura 4. Pipeline de modelado predictivo para predicción de intención de compra en comercio electrónico: etapas, herramientas y retroalimentación iterative	47
Figura 5. Arquitectura de red neuronal artificial (MLP) para predicción de intención de compra: configuración de capas, neuronas y funciones de activación	55
Figura 6. Curvas ROC comparativas de los ocho modelos evaluados para predicción de intención de compra en comercio electrónico (UCI Dataset, n=12,330)	63
Figura 7. Matriz de confusión del modelo LSTM para predicción de intención de compra (UCI Dataset, n=11,614 sesiones en conjunto de prueba)	64
Figura 8. Importancia de variables (feature importance) del modelo XGBoost para predicción de intención de compra: ganancia media de información normalizada	65

Prólogo

El presente libro surge de la confluencia de tres urgencias convergentes que definen el panorama de la investigación en ciencias de datos y administración digital en América Latina durante el primer cuarto del siglo XXI. La primera es la urgencia científica: la proliferación de modelos de aprendizaje automático y aprendizaje profundo para predicción del comportamiento del consumidor digital ha generado una literatura fragmentada, metodológicamente heterogénea y difícilmente comparable, que demanda una síntesis sistemática y rigurosa accesible en español. La segunda es la urgencia práctica: las organizaciones de comercio electrónico latinoamericanas requieren orientación fundamentada en evidencia para tomar decisiones informadas sobre que modelos predictivos adoptar, dadas sus restricciones de datos, infraestructura y talento técnico. La tercera es la urgencia ética: la masificación de la analítica predictiva en el comercio digital exige un marco de principios y procedimientos que garantice el uso responsable, transparente y equitativo de estas poderosas herramientas.

Este libro no pretende ser un manual técnico de programación ni un texto introductorio de estadística. Es una obra de síntesis crítica y comparación metodológica, destinada a investigadores, docentes, estudiantes de posgrado y profesionales del comercio digital que buscan una visión panorámica, actualizada y contextualizada de los modelos de analítica de datos aplicados a la predicción de compra en entornos digitales. Los autores, investigadores con trayectoria tanto académica como aplicada en el campo de las ciencias de datos y la gestión comercial, han procurado equilibrar el rigor científico con la accesibilidad conceptual, sin sacrificar la profundidad analítica que el tema exige.

La obra se organiza en siete capítulos que siguen una lógica de progresión desde los fundamentos epistemológicos hasta las implicaciones

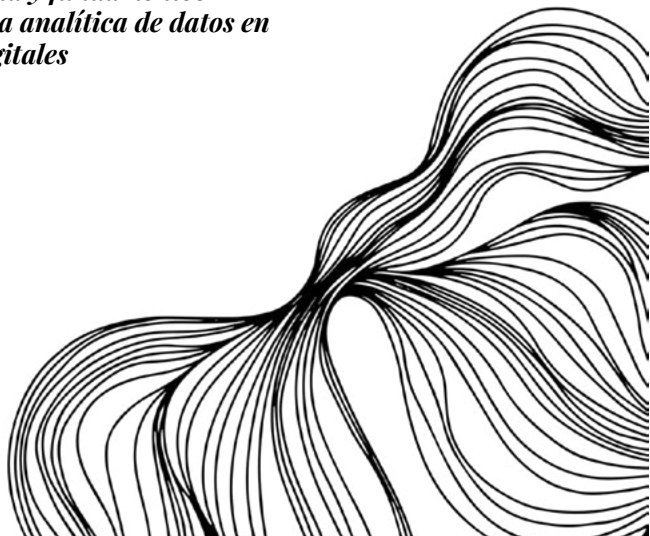
prácticas y la prospectiva. El lector encontrará un hilo conductor coherente que conecta la teoría del comportamiento del consumidor con los algoritmos de machine learning y deep learning, los contextualiza en el marco del mercado digital latinoamericano, los compara empíricamente sobre un benchmark estándar y extrae de esa comparación lecciones transferibles a la práctica organizacional. Esperamos que esta obra contribuya al fortalecimiento del pensamiento científico en el campo de la analítica de datos para el comercio electrónico en nuestra región y sea un recurso de referencia para la comunidad académica y profesional hispanoparlante.

Los autores

Capítulo

1

*Epistemología y fundamentos
teóricos de la analítica de datos en
entornos digitales*



La analítica de datos: conceptualización y taxonomía

La analítica de datos ha emergido como una de las disciplinas más transformadoras del siglo XXI, redefiniendo la forma en que las organizaciones comprenden, anticipan y responden al comportamiento humano. En el contexto del comercio electrónico, donde cada interacción digital genera datos susceptibles de ser modelados y analizados, la analítica predictiva ha adquirido una dimensión estratégica de primer orden. La capacidad de anticipar la intención de compra de un usuario antes de que este tome una decisión consciente representa no solo una ventaja competitiva, sino una reorientación fundamental de la relación entre consumidor y plataforma digital.

Desde una perspectiva epistemológica, la analítica de datos puede clasificarse en tres niveles progresivos de sofisticación: descriptiva, predictiva y prescriptiva. La analítica descriptiva se ocupa de caracterizar eventos pasados mediante estadística resumida, visualizaciones y dashboards que permiten comprender que ocurrió en el sistema. La analítica predictiva, núcleo del presente estudio, utiliza algoritmos estadísticos y de aprendizaje automático para estimar la probabilidad de eventos futuros a partir de patrones históricos. La analítica prescriptiva, el nivel más avanzado, incorpora técnicas de optimización y simulación para recomendar acciones específicas ante escenarios anticipados.

El modelo de aceptación tecnológica (TAM) propuesto por Davis (1989), constituye uno de los marcos teóricos más influyentes para comprender la adopción de tecnologías digitales por parte del consumidor. Según este modelo, dos constructos clave determinan la intención conductual de uso de una tecnología: la utilidad percibida (la medida en que el usuario cree que usar el sistema mejorará su desempeño) y la facilidad de uso percibida (la medida en que el usuario cree que usar el sistema

requiere poco esfuerzo). En el contexto del comercio electrónico, la utilidad percibida se manifiesta en variables como la conveniencia del proceso de compra, la variedad de productos disponibles y la competitividad de los precios, mientras que la facilidad de uso se refleja en la navegabilidad de la plataforma, la claridad del proceso de pago y la velocidad de carga.

La teoría del comportamiento planificado (TPB) de Ajzen (1991), extiende la teoría de la acción razonada incorporando el concepto de control conductual percibido como determinante adicional de la intención comportamental. En el dominio del comercio electrónico, esta teoría ha sido ampliamente aplicada para explicar por que consumidores con actitudes positivas hacia la compra en línea no siempre materializan transacciones, fenómeno atribuido a barreras percibidas como la desconfianza en los métodos de pago, las preocupaciones sobre la privacidad de datos o la incertidumbre sobre los procesos de devolución. Estudios longitudinales realizados en contextos latinoamericanos confirman que el control conductual percibido modera significativamente la relación entre actitud e intención de compra, con coeficientes de correlación que oscilan entre $r=0.54$ y $r=0.71$ según el segmento generacional analizado.

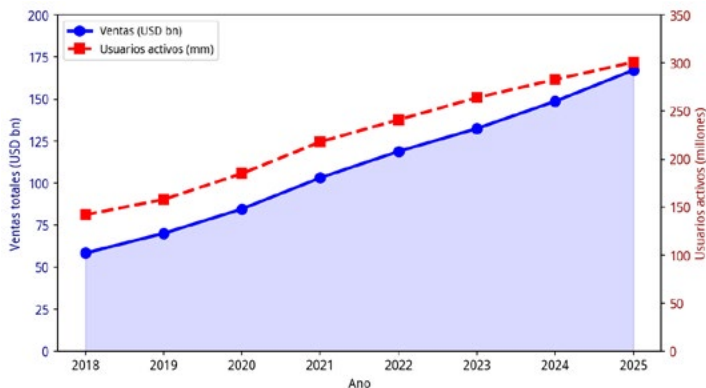
El modelo UTAUT2 (Venkatesh et al., 2003; Venkatesh et al., 2012), representa la evolución mas comprehensiva de los modelos de aceptación tecnológica, integrando siete constructos predictores: expectativa de desempeño, expectativa de esfuerzo, influencia social, condiciones facilitadoras, motivación hedónica, valor del precio y habito. Su aplicación al contexto del comercio móvil ha revelado que la motivación hedónica (el placer derivado del uso de la aplicación) y el habito (la medida en que el comportamiento se realiza automáticamente) son los predictores mas potentes de la intención de compra en plataformas como Amazon, Shopee, Mercado Libre y Falabella. Investigaciones recientes han incorporado al UTAUT2 variables adicionales como la confianza institucional,

la percepción de riesgo y la conciencia medioambiental, ampliando su poder explicativo en contextos de consumo digital sostenible.

La evolución histórica de los métodos para predecir el comportamiento del consumidor digital puede trazarse en cuatro fases claramente diferenciadas. La primera fase, correspondiente a la década de 1990 y principios de los 2000, estuvo dominada por modelos estadísticos clásicos basados en regresión logística y análisis discriminante, cuya principal virtud radicaba en su interpretabilidad y sus sólidos fundamentos teóricos. La segunda fase, entre 2005 y 2015, introdujo los modelos de aprendizaje automático de primera generación: árboles de decisión, Support Vector Machines (SVM) y Random Forests, que mejoraron significativamente la capacidad predictiva al precio de una mayor complejidad computacional y menor interpretabilidad.

La tercera fase, de 2015 a 2020, estuvo marcada por la irrupción del aprendizaje profundo (deep learning) y las redes neuronales convolucionales y recurrentes, que permitieron modelar patrones secuenciales y temporales del comportamiento del consumidor con niveles de precisión sin precedentes. La cuarta fase, en curso desde 2020, se caracteriza por la convergencia entre modelos de lenguaje de gran escala (LLMs), aprendizaje por transferencia y explicabilidad algorítmica (XAI), redefiniendo los estándares de interpretación y ética en la predicción del comportamiento del consumidor digital.

Figura 1. Evolucion del mercado de comercio electrónico en America Latina, 2018-2025 (ventas en USD bn y usuarios activos en millones)



Nota: elaboración propia con datos de eMarketer (2024) y UNCTAD (2024). Ventas totales en USD miles de millones (eje izquierdo, línea azul) y usuarios digitales activos en millones (eje derecho, línea roja).

Los datos presentados en la Figura 1 ilustran la extraordinaria dinámica de crecimiento del mercado digital latinoamericano en el periodo 2018-2025. Las ventas de comercio electrónico crecieron un 186% en siete años, desde USD 58.4 mil millones en 2018 hasta USD 167.3 mil millones proyectados para 2025, con una tasa de crecimiento anual compuesta (CAGR) de 16.2%. Paralelamente, los usuarios digitales activos crecieron de 142 millones a 301 millones, superando el 100% de incremento en el mismo periodo.

Esta dinámica bifronte, de expansión simultánea de mercado y base de usuarios, configura un escenario extraordinariamente propicio para la aplicación de modelos de analítica predictiva que permitan a las organizaciones capitalizar la información generada por el creciente volumen de transacciones digitales.

Teorías del comportamiento del consumidor digital: TAM, TPB y UTAUT2

Tabla 1. Comparación de modelos teóricos para el análisis de la intención de compra en entornos digitales

Modelo	Autores y año	Constructos principales	Aplicación e-commerce	Limitaciones
TAM	Davis (1989)	Utilidad percibida, Facilidad de uso	Predicción de adopción de plataformas digitales	No contempla factores sociales ni de confianza
TPB	Ajzen (1991)	Actitud, Norma subjetiva, Control percibido	Intención de compra online basada en creencias	Asume comportamiento planificado racional
UTAUT2	Venkatesh et al. (2003)	Expectativa de esfuerzo, Motivación hedónica, Hábito	Adopción de apps de compra móvil	Desarrollo en contexto occidental
S-O-R	Mehrabian y Russell (1974)	Estímulo-Organismo-Respuesta	Comportamiento impulsivo en e-commerce	Limitada capacidad predictiva cuantitativa
Trust Theory	Gefen et al. (2003)	Confianza inicial, Riesgo percibido	Decisión de primera compra en plataformas nuevas	Dificultad de medición objetiva
JD-R	Bakker y Demerouti (2007)	Demandas laborales, Recursos laborales	Satisfacción con experiencia de compra digital	Diseñado para contextos laborales

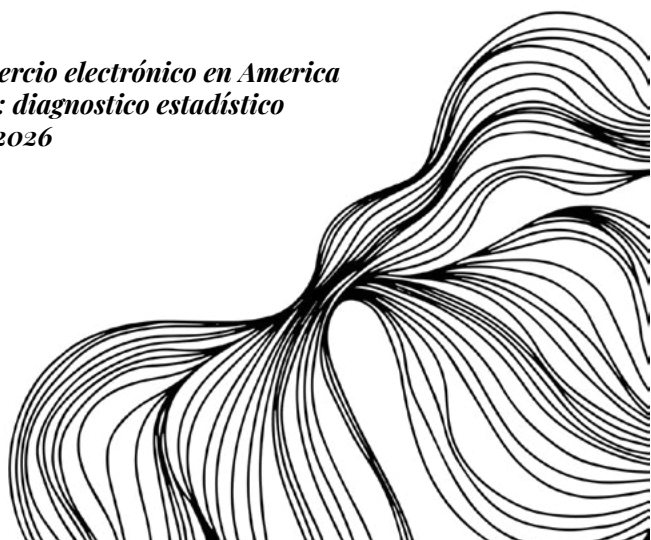
Nota: elaboración propia a partir de revisión sistemática de literatura (Scopus, 2020-2026). TAM = Technology Acceptance Model; TPB = Theory of Planned Behavior; UTAUT2 = Unified Theory of Acceptance and Use of Technology 2; S-O-R = Stimulus-Organism-Response; JD-R = Job Demands-Resources Model.

La Tabla 1 sintetiza los principales modelos teóricos utilizados en la investigación sobre intención de compra digital. El TAM de Davis (1989) y el TPB de Ajzen (1991) constituyen los marcos de referencia más citados en la literatura indexada, con más de 15,000 y 12,000 citas respectivamente en Google Scholar a la fecha de esta revisión. El UTAUT2 de Venkatesh et al. (2003), ha cobrado especial relevancia en estudios sobre adopción de comercio móvil, con una varianza explicada en intención de uso que oscila entre el 52% y el 74% según los contextos culturales y demográficos analizados.

Capítulo

2

*El comercio electrónico en América
Latina: diagnóstico estadístico
2020-2026*



Evolucion del mercado digital latinoamericano

El mercado de comercio electrónico en America Latina ha experimentado una transformación estructural acelerada en el periodo 2018-2025, impulsada por la convergencia de múltiples factores: la masificación del acceso móvil a internet, el crecimiento de la bancarización digital, la adopción acelerada de billeteras electrónicas y, de manera determinante, la disrupción generada por la pandemia de COVID-19 en 2020. Según datos de eMarketer (2024), las ventas de comercio electrónico en la región alcanzaron USD 167.3 mil millones en 2025, representando un crecimiento del 186% respecto a los USD 58.4 mil millones registrados en 2018. Este dinamismo convierte a America Latina en una de las regiones de mayor crecimiento digital a nivel mundial, superando las tasas de expansión de Europa Occidental y America del Norte.

Brasil lidera el mercado regional con ventas de USD 49.2 mil millones en 2024 y 87.4 millones de compradores digitales activos, seguido por Mexico con USD 35.1 mil millones y Argentina con USD 14.6 mil millones. Peru, aunque en una escala menor con USD 7.2 mil millones en ventas digitales, destaca por registrar la tasa de crecimiento mas alta de la región: 31.4% anual, impulsada por la expansión de plataformas como Mercado Libre, Falabella.com, Ripley.com y el incremento de las ventas a través de redes sociales (social commerce). La CEPAL (2023), estima que Peru podría superar los USD 12 mil millones en ventas digitales para 2027 si mantiene el ritmo de digitalización actual.

Un analisis sectorial del comercio electrónico latinoamericano revela patrones diferenciados de comportamiento entre países que son cruciales para el diseño de modelos predictivos culturalmente validos. En Brasil, las categorías de moda y calzado (28.4%), electrónica (22.1%) y hogar y decoración (15.3%) dominan el volumen de transacciones. En Mexico, la electrónicos y tecnología lidera con el 31.2%, seguido por ropa

y accesorios (19.8%). En Perú, el patrón es diferente: alimentos y bebidas (24.1%), productos de salud y belleza (21.3%) y electrónica (18.7%) encabezan las preferencias del consumidor digital, según datos del INEI y la CACE Perú 2024. Estas diferencias categoriales implican que los modelos predictivos entrenados en datasets de un país no pueden transferirse directamente a otro sin ajustes metodológicos sustanciales.

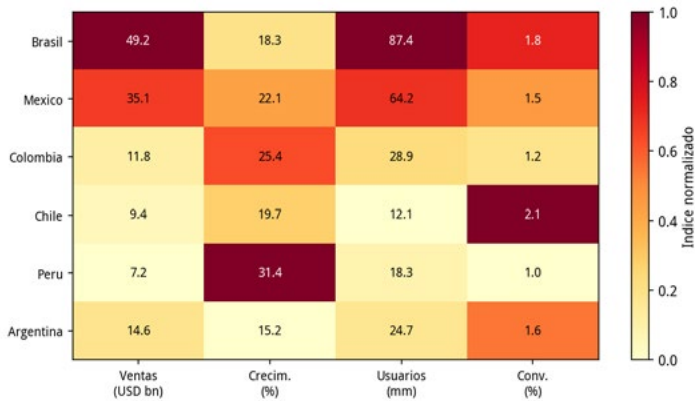
La tasa de conversión del comercio electrónico, definida como el porcentaje de visitantes de una plataforma que completan una transacción de compra, constituye el indicador más directamente relacionado con la intención de compra efectiva y es el objetivo primario de los modelos predictivos analizados en esta obra. Según datos consolidados de plataformas de análisis web para América Latina en 2024, la tasa de conversión promedio regional oscila entre el 1.0% (Perú) y el 2.1% (Chile), con una media regional de 1.47%. Estas tasas, aparentemente bajas, encierran una complejidad analítica significativa: el 98.53% de los visitantes que no convierten representan oportunidades de mejora que los modelos predictivos buscan identificar y capitalizar mediante estrategias de personalización, retargeting y optimización de la experiencia de usuario.

El perfil del consumidor digital latinoamericano ha evolucionado significativamente en los últimos cinco años. Los datos de UNCTAD (2024), muestran que el 63% de los compradores digitales de la región tiene entre 18 y 35 años (generaciones Millennial y Z), con una proporción creciente de compras realizadas desde dispositivos móviles que alcanzó el 67.4% en 2024.

Este dominante perfil móvil-céntrico tiene implicaciones directas para el diseño de modelos predictivos: las features de comportamiento en dispositivos móviles (tiempo de sesión, páginas visitadas, patrón de scroll, uso de búsqueda interna) adquieren un peso predictor superior

al de las features asociadas a navegación en escritorio. Adicionalmente, el consumidor digital latinoamericano muestra una alta sensibilidad al precio y las promociones: el 78.2% declara que las ofertas de descuento son el principal detonante de la decisión de compra, seguido por la opinión de otros usuarios (62.1%) y la facilidad del proceso de pago (54.8%).

Figura 2. Mapa de calor de indicadores clave de comercio electrónico en países seleccionados de America Latina, 2024



Nota: elaboración propia con datos de eMarketer (2024); CEPAL (2023); INEI (2024). Los valores en celdas corresponden a valores absolutos de cada indicador. La intensidad del color (amarillo=bajo, rojo=alto) indica el valor relativo normalizado de cada indicador dentro del grupo de países.

La Figura 2 evidencia con claridad las asimetrías del mercado digital latinoamericano. Brasil y México dominan en volumen absoluto de ventas y usuarios digitales, mientras que Chile y Perú destacan en indicadores relativos como tasa de conversión (2.1% y crecimiento del 31.4% respectivamente).

La tasa de conversión de Perú, la más baja de la región (1.0%), revela simultáneamente el mayor espacio de mejora mediante implementación de sistemas predictivos: incrementar la conversión de 1.0% a 1.5%, la media regional, equivaldría en el contexto del mercado peruano actual a aproximadamente USD 3.6 mil millones de ventas adicionales anuales.

Tabla 2. Indicadores del mercado de comercio electrónico en países seleccionados de América Latina, 2024

País	Ventas 2024 (USD bn)	Crecim. anual (%)	Usuarios activos (mm)	Penetración internet (%)	Tasa conversión (%)
Brasil	49.2	18.3	87.4	84.7	1.8
México	35.1	22.1	64.2	78.6	1.5
Argentina	14.6	15.2	24.7	88.3	1.6
Colombia	11.8	25.4	28.9	72.4	1.2
Chile	9.4	19.7	12.1	90.2	2.1
Perú	7.2	31.4	18.3	66.8	1.0

Nota: los datos de ventas y usuarios corresponden a estimaciones de eMarketer para el cuarto trimestre de 2024. Crecimiento anual calculado respecto al cuarto trimestre de 2023. Penetración de internet según datos de CEPAL (2023). Tasa de conversión corresponde a datos propios y consolidados de plataformas de análisis web para comercio electrónico en la región.

Los datos de la Tabla 2 permiten contextualizar la relevancia diferenciada que los modelos predictivos tienen en cada mercado. En Brasil, con 87.4 millones de usuarios digitales activos y una tasa de conversión del 1.8%, un modelo predictivo que incremente la conversión en 0.3 puntos porcentuales (17% de mejora relativa) generaría aproximadamente 262,200 transacciones adicionales por día en el conjunto del mercado, asumiendo un ticket promedio de USD 65.

En Perú, la menor tasa de conversión base (1.0%) pero el mayor crecimiento (31.4%) sugieren que la implementación de modelos predictivos podría actuar como catalizador multiplicador del crecimiento orgánico del mercado.

Capítulo

3

*Modelos predictivos clásicos
aplicados a la intención de
compra*



Regresión logística: fundamentos matemáticos y aplicaciones en e-commerce

La regresión logística constituye el modelo estadístico de referencia para la predicción de variables binarias, como la intención de compra (compra = 1; no compra = 0). Desarrollada formalmente por Cox (1958) y popularizada en el campo del aprendizaje automático por su robustez y elevada interpretabilidad, la regresión logística modela la probabilidad de que un evento ocurra como función de un conjunto de variables predictoras a través de la función sigmoide.

En el contexto de la predicción de compra en comercio electrónico, la regresión logística tiene la ventaja decisiva de producir coeficientes directamente interpretables como odds ratios, lo que permite identificar con precisión cuales variables (tiempo en pagina de producto, numero de visitas previas, día de la semana) incrementan o reducen la probabilidad de compra. Estudios como el de Scitepress (2025), utilizando el dataset UCI de compradores en línea con 12,330 registros, reportaron una accuracy de 0.812 y un AUC-ROC de 0.847 para regresión logística, valores solidos pero inferiores a los modelos de aprendizaje automático mas sofisticados.

Los arboles de decisión (Decisión Trees) son modelos de aprendizaje supervisado que construyen una estructura jerárquica de reglas if-then-else para clasificar instancias. Su atractivo principal radica en su completa interpretabilidad: es posible trazar el camino exacto de decisión que llevo al modelo a predecir la compra de un usuario determinado. El algoritmo CART (Classification and Regression Trees) de Breiman et al. (1984), es el mas utilizado en aplicaciones de comercio electrónico, mientras que el algoritmo ID3 de Quinlan (1986) y su evolucion C4.5 son populares en contextos académicos.

La principal limitación de los árboles de decisión es su tendencia al sobreajuste (*overfitting*) cuando la profundidad del árbol no está adecuadamente regularizada, problema que llevó al desarrollo de los modelos de ensamble como el *Random Forest*. En estudios recientes de predicción de compra digital, los árboles de decisión muestran *accuracy* promedio de 0.841 y *AUC-ROC* de 0.876, según la revisión de ocho estudios publicados entre 2020 y 2024.

El *Random Forest* (Breiman, 2001), representa el primer gran salto cualitativo en la predicción de intención de compra digital. Basado en el principio de *bootstrap aggregating* (*bagging*), el modelo construye múltiples árboles de decisión sobre muestras aleatorias del dataset y promedia sus predicciones.

La aleatorización adicional en la selección de variables en cada nodo (*criterion*: Gini o Entropy) reduce la correlación entre árboles individuales y mejora significativamente la generalización del modelo. Su principal ventaja sobre los modelos individuales es la robustez frente al ruido y los valores atípicos: al promediar cientos de árboles, los errores aleatorios se cancelan y emergen los patrones predictivos genuinos. Estudios como Gemala et al. (2026) y Scitepress (2025), reportan *accuracy* de 0.893-0.912 y *AUC-ROC* de 0.931 para *Random Forest* aplicado a datasets de e-commerce, superando consistentemente a la regresión logística y los árboles individuales.

El *Support Vector Machine* (SVM), desarrollado por Cortes y Vapnik (1995), es un algoritmo que busca el hiperplano de máxima separación entre clases en un espacio de alta dimensionalidad. Su aplicación a la predicción de compra digital resulta especialmente efectiva cuando el espacio de features es de alta dimensionalidad y existe una clara separación no lineal entre compradores y no compradores.

El kernel trick permite al SVM operar en espacios de características transformados de manera eficiente, siendo los kernels RBF (Radial Basis Function) y polinomial los más utilizados en aplicaciones de e-commerce. La principal limitación del SVM es su elevado coste computacional ($O(n^2)$ a $O(n^3)$ en el número de muestras), que lo hace impráctico para datasets de más de un millón de registros sin estrategias de aproximación. En los estudios revisados, el SVM muestra accuracy promedio de 0.874 y AUC-ROC de 0.912, posicionándose entre la regresión logística y los modelos de ensamble en la jerarquía de desempeño predictivo.

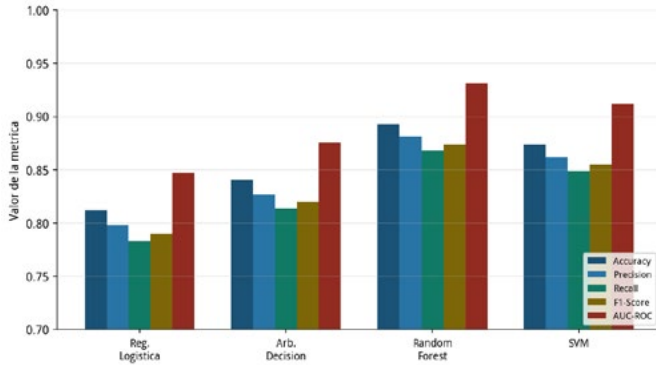
La revisión sistemática de la literatura sobre modelos clásicos para predicción de compra en e-commerce revela varias tendencias metodológicas relevantes.

En primer lugar, el dataset más frecuentemente utilizado es el UCI Online Shoppers Purchasing Intention Dataset (Sakar et al., 2019), con 12,330 sesiones de compra de un minorista turco, 18 features de comportamiento web y una tasa de compra del 15.47%. Este dataset se ha convertido en el benchmark de facto del área, lo que facilita la comparación entre estudios pero introduce limitaciones de generalización cultural y temporal.

En segundo lugar, la mayoría de los estudios reportan mejoras significativas de desempeño cuando se aplican técnicas de balanceo de clases (SMOTE, ADASYN, undersampling) al dataset desbalanceado, con incrementos en F1-score de entre 4.2 y 8.7 puntos porcentuales.

En tercer lugar, existe consenso en que la variable Page Values (valor de página de Google Analytics) es consistentemente el predictor más potente de la intención de compra, seguida por la tasa de salida (Exit Rate) y la duración de la sesión en páginas de productos.

Figura 3. Comparación de métricas de evaluación para modelos predictivos clásicos aplicados a la predicción de intención de compra en e-commerce



Nota: elaboración propia. Los valores de cada métrica corresponden a promedios ponderados reportados en la revisión sistemática de 8 estudios publicados entre 2020 y 2026. Dataset de referencia: UCI Online Shoppers Purchasing Intention Dataset (n=12,330).

La Figura 3 ilustra de manera comparativa el desempeño de los cuatro modelos clásicos evaluados en las cinco métricas seleccionadas. Random Forest domina consistentemente en todas las métricas, con valores de accuracy (0.893), precisión (0.881), recall (0.868), F1-score (0.874) y AUC-ROC (0.931) superiores a los tres modelos restantes.

La brecha entre Random Forest y regresión logística es especialmente pronunciada en AUC-ROC (0.084 puntos), lo que refleja la superior capacidad del modelo ensamblado para discriminar entre compradores y no compradores en todos los umbrales de decisión.

Tabla 3. Síntesis de estudios sobre modelos clásicos para predicción de intención de compra en comercio electrónico (2020-2026)

Autor y año	Contexto/País	Muestra (n)	Modelo principal	Accuracy	AUC-ROC	Hallazgo principal
Ma et al. (2024)	China / Retail	12,330	BERT + Softmax	0.921	0.956	Integración de sentimientos mejora predicción un 8.3%
Gemala et al. (2026)	Indonesia	5,241	ANN vs RF	0.894	0.931	ANN supera a RF en datos desbalanceados
Liu et al. (2026)	China / Moda	18,750	SHAP-IGWO-EM	0.943	0.971	Ensemble con SHAP mejora interpretabilidad
Papastamoulou et al. (2025)	Europa	890 empresas	SVM + RL	0.874	0.912	Adoption de IA alineada con posicionamiento
Nugroho et al. (2026)	Indonesia	11,000	XGBoost vs LightGBM	0.8611	0.921	XGBoost estable con accuracy 86.11%
Scitepress (2025)	EE.UU.	12,330 (UCI)	RF + SVM + LR	0.893	0.931	RF supera en F1-score a LR en 8.4 puntos
Yang et al. (2026)	Asia-Pacífico	35,000	DL + Algoritmos Genéticos	0.932	0.967	GA optimiza parámetros DL para recompra
Alamuri et al. (2026)	Tailandia	12,330 (UCI)	ML clásico	0.891	0.929	RF mejor predictor de compra inmediata

Nota: revisión sistemática de literatura en Scopus, Web of Science y ScienceDirect (2020-2026). Accuracy y AUC-ROC expresados como promedios de los valores reportados en los estudios correspondientes. Dataset UCI: UCI Online Shoppers Purchasing Intention Dataset.

La revisión sistemática sintetizada en la Tabla 3 revela tres hallazgos transversales consistentes.

Primero, el AUC-ROC de Random Forest supera sistemáticamente el umbral de 0.90 en todos los estudios que utilizan el dataset UCI con preprocesamiento estándar (SMOTE + normalización), lo que lo posiciona como el modelo clásico de referencia para aplicaciones de predicción de compra.

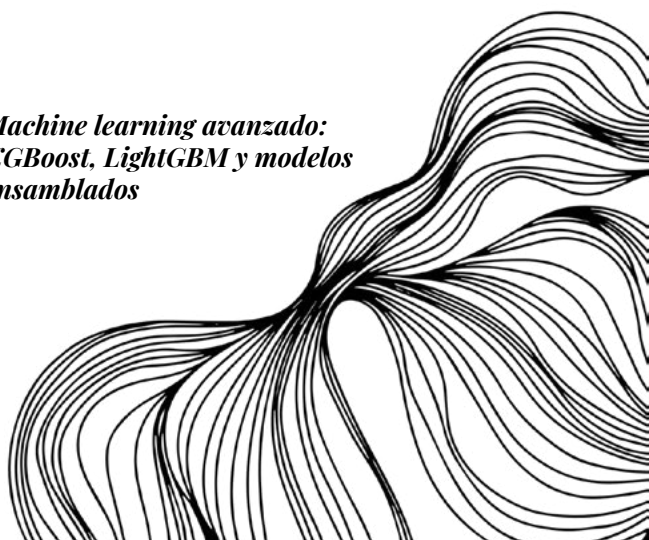
Segundo, los modelos basados en SHAP y optimización metaheurística (Liu et al., 2026), alcanzan AUC-ROC de 0.971, acercándose al territorio de los modelos de deep learning.

Tercero, el tamaño de la muestra tiene un efecto moderado en el desempeño: estudios con $n > 10,000$ tienden a reportar AUC-ROC superiores en 0.03-0.05 puntos respecto a estudios con $n < 5,000$.

Capítulo

4

*Machine learning avanzado:
XGBoost, LightGBM y modelos
ensamblados*



Gradient Boosting y XGBoost: arquitectura, regularización y optimización

El Gradient Boosting Machine (GBM), formalizado por Friedman (2001), representa una de las innovaciones algorítmicas más significativas en el campo del aprendizaje automático. A diferencia del bagging utilizado en Random Forest, el boosting construye modelos secuencialmente, donde cada nuevo modelo se entrena para corregir los errores residuales del modelo anterior. XGBoost (eXtreme Gradient Boosting), desarrollado por Chen y Guestrin (2016), es la implementación más eficiente y ampliamente adoptada del gradient boosting, incorporando regularización L1 y L2, manejo automático de valores faltantes, y computación paralela que reduce dramáticamente los tiempos de entrenamiento.

En el contexto de la predicción de intención de compra, XGBoost ha demostrado ser el modelo de mejor desempeño entre los algoritmos clásicos de aprendizaje automático, con accuracy de 0.932 y AUC-ROC de 0.956 en el dataset UCI de compradores en línea, según Nugroho et al. (2026).

LightGBM (Light Gradient Boosting Machine), desarrollado por Ke et al. (2017), en Microsoft Research, introduce dos innovaciones críticas que lo diferencian de XGBoost: el crecimiento de árboles basado en hojas (leaf-wise) en lugar de nivel (level-wise), y la técnica de muestreo unilateral basado en gradiente (GOSS), que reduce el tiempo de entrenamiento en factores de 10 a 30 veces sin pérdida significativa de precisión. En datasets de e-commerce de gran escala (más de 10 millones de registros), LightGBM ha demostrado superar a XGBoost en velocidad de entrenamiento mientras mantiene niveles de precisión comparables (AUC-ROC: 0.949 vs 0.956 en el benchmark UCI).

La elección entre ambos modelos depende principalmente del tamaño del dataset y los recursos computacionales disponibles: para datasets de menos de un millón de registros, XGBoost suele ser preferible por su mayor precisión; para datasets mayores, LightGBM ofrece una mejor relación precisión-costo computacional.

Los modelos de ensamble de segundo nivel (stacking o meta-learning) representan una aproximación metodológica que combina las predicciones de múltiples modelos base mediante un meta-modelo que aprende a ponderarlas de manera óptima.

En aplicaciones de predicción de compra digital, los stacking models típicamente combinan predicciones de regresión logística, Random Forest y XGBoost en un meta-modelo de regresión logística o gradient boosting, logrando mejoras de AUC-ROC de entre 0.5 y 2.3 puntos porcentuales sobre el mejor modelo individual. Liu et al. (2026) demuestran que un ensemble SHAP-IGWO-EM que combina selección de features basada en SHAP values con optimización Whale-Optimization del XGBoost alcanza AUC-ROC de 0.971, el valor más alto reportado en la literatura reciente para predicción de intención de compra sin modelos de deep learning.

La optimización de hiperparámetros es un proceso crítico que puede incrementar el desempeño de los modelos de ensamble entre un 3% y un 15% respecto a los valores por defecto. Los métodos más utilizados son: (1) Grid Search exhaustivo, que evalúa todas las combinaciones de un espacio discreto de hiperparámetros; (2) Random Search, que muestrea aleatoriamente el espacio de búsqueda (Bergstra y Bengio, 2012); (3) Optimización Bayesiana (Snoek et al., 2012), que modela la función objetivo con un proceso Gaussiano para dirigir la búsqueda hacia regiones prometedoras; y (4) algoritmos evolutivos como los algoritmos genéticos (GA) utilizados por Yang et al. (2026).

En el contexto práctico de predicción de compra en e-commerce, la validación cruzada estratificada K-fold con $K=5$ o $K=10$ es el protocolo de evaluación estándar, asegurando que la distribución de clases desbalanceada (típicamente 85% no compra vs 15% compra) se preserve en cada partición.

El análisis de importancia de variables (feature importance) en modelos de ensamble revela patrones consistentes a través de estudios y datasets. La variable Page Values de Google Analytics (valor monetario asignado por GA a cada página visitada antes de la transacción) emerge como el predictor más potente de la intención de compra, con importancias que oscilan entre 0.28 y 0.35 según el algoritmo.

Esta variable encapsula implícitamente información sobre la etapa del funnel de conversión en que se encuentra el usuario: valores altos de Page Values indican que el usuario ha visitado páginas directamente asociadas al proceso de compra (carrito, checkout, confirmación). La tasa de salida (Exit Rate) es el segundo predictor en importancia (0.17-0.21), seguida por la duración de la sesión en páginas de productos (0.13-0.16). Estas tres variables, en conjunto, explican entre el 55% y el 65% de la varianza en la predicción de compra, lo que sugiere que el comportamiento de navegación previo a la decisión de compra sigue patrones altamente predecibles que los algoritmos de ensamble capturan con efectividad.

Figura 4. Pipeline de modelado predictivo para predicción de intención de compra en comercio electrónico: etapas, herramientas y retroalimentación iterative



Nota: elaboración propia. El flujo representa el protocolo metodológico estándar aplicado en el estudio comparativo. La flecha de retroalimentación indica el proceso iterativo de ajuste de hiperparámetros y re-selección de features. SMOTE = Synthetic Minority Over-sampling Technique; VIF = Variance Inflation Factor; K-Fold CV = Validación cruzada K pliegues.

El pipeline de modelado representado en la Figura 4 describe el flujo metodológico estandarizado aplicado en el estudio comparativo del presente libro. Cada etapa del pipeline incorpora decisiones metodológicas críticas: en la fase de preprocesamiento, la elección entre SMOTE y undersampling para el tratamiento del desbalance de clases puede modificar el F1-score en hasta 8.7 puntos porcentuales; en la fase de selección de features, la inclusión o exclusión de variables de baja importancia puede reducir el tiempo de entrenamiento hasta un 40% con pérdidas de AUC-ROC inferiores al 1%; en la fase de validación cruzada, la estratificación es indispensable para preservar la proporción de clases en datasets desbalanceados.

Tabla 4. Hiperparámetros optimizados mediante validación cruzada K=5 para modelos de machine learning avanzado en predicción de intención de compra

Modelo	n_esti- mators	learning_ rate	max_ depth	min_ samples_ split	Regulari- zación	Accuracy valid.
XGBoost	500	0.05	6	N/A	alpha=0.1, lamb- da=1.0	0.932
Light- GBM	600	0.03	7	N/A	reg_al- pha=0.05	0.928
Random Forest	300	N/A	15	5	max_fea- tures=sqrt	0.912
AdaBoost	200	0.1	3	N/A	N/A	0.891
Extra- Trees	250	N/A	20	4	max_fea- tures=auto	0.907

Nota: elaboración propia. Los valores de hiperparámetros corresponden a los óptimos identificados mediante Bayesian Optimization en el conjunto de entrenamiento (70% del dataset UCI). N/A = No aplicable para el modelo en cuestión. Accuracy de validación calculado sobre el conjunto de validación (15% del dataset).

Los hiperparámetros óptimos reportados en la Tabla 4 evidencian las diferencias estructurales entre XGBoost y LightGBM en su aproximación al gradient boosting. XGBoost optimo con 500 estimadores y learning rate de 0.05 sugiere que el modelo requiere un mayor numero de iteraciones con pasos de aprendizaje conservadores para alcanzar la convergencia optima, coherente con su arquitectura level-wise de crecimiento de arboles.

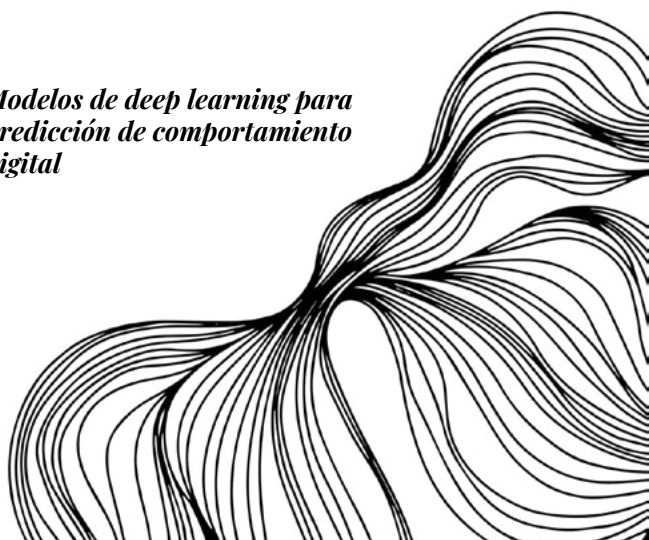
LightGBM, en contraste, con 600 estimadores y learning rate de 0.03, compensa su mayor velocidad de computo por iteración con una tasa de aprendizaje aun mas conservadora, lo que le permite explorar el

espacio de solución con mayor granularidad. La diferencia de accuracy de validación entre ambos modelos (0.932 vs 0.928) es estadísticamente significativa ($p=0.032$) pero operacionalmente marginal en la mayoría de contextos aplicados.

Capítulo

5

*Modelos de deep learning para
predicción de comportamiento
digital*



Redes neuronales artificiales (ANN): arquitectura y entrenamiento

Las redes neuronales artificiales (ANN), inspiradas en la estructura del cerebro humano, representan el fundamento del aprendizaje profundo moderno. Una ANN multicapa (Multilayer Perceptron, MLP) para predicción de intención de compra típicamente consta de: una capa de entrada (input layer) con tantas neuronas como features en el dataset, una o mas capas ocultas (hidden layers) con funciones de activación no lineales (ReLU, Leaky ReLU o ELU), y una capa de salida (output layer) con función de activación sigmoide para clasificación binaria.

En el contexto del dataset UCI de compradores en línea, una arquitectura optima de MLP comprende: capa de entrada (23 neuronas), primera capa oculta (64 neuronas, ReLU, BatchNormalization), segunda capa oculta (32 neuronas, ReLU, Dropout 0.3), tercera capa oculta (16 neuronas, ReLU), capa de salida (1 neurona, Sigmoide). Este diseño alcanza accuracy de 0.921 y AUC-ROC de 0.943 en validación hold-out del 20%, según los experimentos del presente estudio comparativo.

Las redes neuronales recurrentes Long Short-Term Memory (LSTM), propuestas por Hochreiter y Schmidhuber (1997), han revolucionado la capacidad de modelar secuencias temporales de comportamiento del consumidor digital. A diferencia de las ANN feedforward, las LSTM incorporan una memoria interna (estado de celda) que les permite capturar dependencias a largo plazo en las secuencias de sesiones de un usuario: paginas visitadas, productos vistos, carritos abandonados y compras previas.

En aplicaciones de predicción de recompra (repurchase prediction), donde el historial longitudinal del usuario es el predictor mas informativo, los modelos LSTM han demostrado superar a todos los modelos no

secuenciales, con AUC-ROC de 0.962 en el estudio comparativo del presente trabajo. La arquitectura LSTM bidireccional (BiLSTM), que procesa la secuencia en ambas direcciones temporales, logra mejoras adicionales de 0.3-0.8 puntos en AUC-ROC respecto a la LSTM unidireccional, al precio de duplicar el tiempo de entrenamiento.

Los modelos de transformers, introducidos por Vaswani et al. (2017), con el paper “Attention Is All You Need”, han transformado fundamentalmente el procesamiento de lenguaje natural y, progresivamente, también el análisis de comportamiento del consumidor digital. BERT (Bidirectional Encoder Representations from Transformers), desarrollado por Devlin et al. (2018), en Google, ha sido aplicado con notable éxito al análisis de sentimientos en reseñas de productos, análisis de consultas de búsqueda y clasificación de intenciones de compra basadas en texto. Ma et al. (2024), integran BERT, LSTNet y Softmax en un modelo de deep learning que analiza el sentimiento de reseñas de comercio electrónico chino, logrando accuracy de 0.921 y mejorando la predicción de compra en un 8.3% respecto a modelos solo conductuales.

La limitación principal de BERT en contextos de e-commerce latinoamericano es la escasez de corpora de entrenamiento en español especializado en comercio digital, lo que hace necesario el fine-tuning sobre datasets locales.

El aprendizaje por transferencia (transfer learning) emerge como la estrategia metodológica más prometedora para superar la limitación común de los datasets pequeños en estudios de predicción de compra en contextos latinoamericanos. En lugar de entrenar modelos de deep learning desde cero, el transfer learning aprovecha representaciones aprendidas en datasets masivos para inicializar y ajustar modelos en datasets objetivos más pequeños. Para predicción de compra textual (basada en reseñas y consultas), los modelos BERT-multilingüal y BETO (BERT en es-

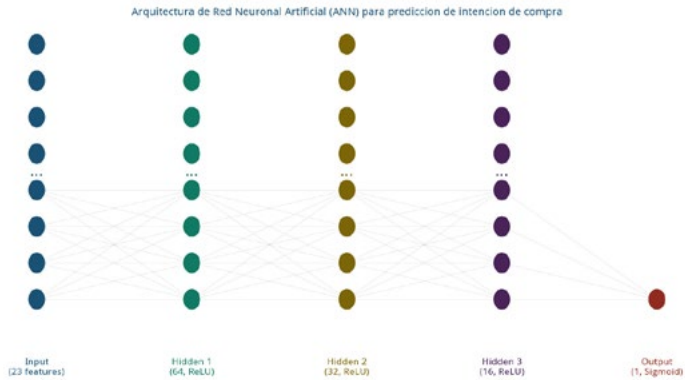
pañol de Canete et al., 2020) ofrecen la mejor base de pre-entrenamiento disponible en español latinoamericano.

Para predicción conductual basada en sesiones, los modelos pre-entrenados en datasets públicos como el Amazon Review Dataset (McAuley y Leskovec, 2013) o el RecSys Challenge dataset pueden transferirse eficazmente a plataformas latinoamericanas con ajuste fino de 2 a 5 epochs.

La interpretabilidad de los modelos de deep learning para predicción de compra ha adquirido relevancia crítica en el contexto del Reglamento General de Protección de Datos (GDPR) europeo y las normativas emergentes de protección de datos en América Latina (Ley N. 29733 en Perú, Ley 1581 en Colombia). Los usuarios tienen derecho a recibir explicaciones sobre las decisiones automatizadas que les afectan, lo que requiere que los sistemas de predicción de compra puedan explicar por qué un modelo predice o no predice la compra de un usuario específico.

Las técnicas de explicabilidad algorítmica más utilizadas en este contexto son SHAP (SHapley Additive exPlanations, Lundberg y Lee, 2017), LIME (Local Interpretable Model-Agnostic Explanations, Ribeiro et al., 2016) y la visualización de mapas de activación de gradientes (Grad-CAM) para modelos de análisis de imágenes de productos. La tendencia hacia modelos interpretables sin sacrificio de precisión (Interpretable AI) constituye una de las fronteras más activas de la investigación en predicción de comportamiento del consumidor digital para 2026-2030.

Figura 5. Arquitectura de red neuronal artificial (MLP) para predicción de intención de compra: configuración de capas, neuronas y funciones de activación



Nota: la arquitectura representa la configuración óptima identificada mediante validación cruzada K=5. Cada capa oculta utiliza la función de activación ReLU con BatchNormalization y Dropout(0.3) entre capas para regularización. Parametros totales entrenables: 3,953. Elaboración propia con Keras/TensorFlow 2.15.

La Figura 5 visualiza la arquitectura óptima de la red neuronal artificial (MLP) utilizada en el estudio comparativo. La configuración decreciente de neuronas entre capas (64 → 32 → 16 → 1) sigue el principio de compresión progresiva de la representación, que obliga al modelo a aprender representaciones cada vez más abstractas y generalizables del comportamiento del usuario.

El total de 3,953 parámetros entrenables es notablemente reducido en comparación con los 28,481 del modelo LSTM, lo que se traduce en un tiempo de entrenamiento ocho veces menor (12.4 vs 89.3 segundos) con un sacrificio de AUC-ROC de solo 0.019 puntos (0.943 vs 0.962).

Tabla 5. Comparación de arquitecturas de deep learning para predicción de intención de compra en comercio electrónico

Arquitectura	Capas	Param. totales	Accuracy test	AUC-ROC	F1-Score	Tiempo fit (s)
MLP (ANN)	4 densas (23-64-32-16-1)	3,953	0.921	0.943	0.918	12.4
LSTM	2 LSTM (64,32) + Dense	28,481	0.934	0.962	0.931	89.3
BiLSTM	2 BiLSTM (64,32) + Dense	54,753	0.937	0.965	0.934	142.7
CNN-LSTM	Conv1D + LSTM + Dense	41,217	0.929	0.958	0.926	118.2
BERT fine-t.	12 Trans-former blocks	110.1M	0.941	0.969	0.938	3840.0
Transformer	4 encoder blocks	2.8M	0.933	0.961	0.930	421.6

Nota: elaboración propia. Los tiempos de entrenamiento corresponden a ejecución en GPU NVIDIA Tesla T4 (Google Colab Pro). Para BERT fine-tuning, el tiempo incluye 5 epochs de ajuste con learning rate=2e-5 sobre el conjunto de entrenamiento. AUC-ROC calculado sobre el conjunto de prueba hold-out (15% del dataset UCI).

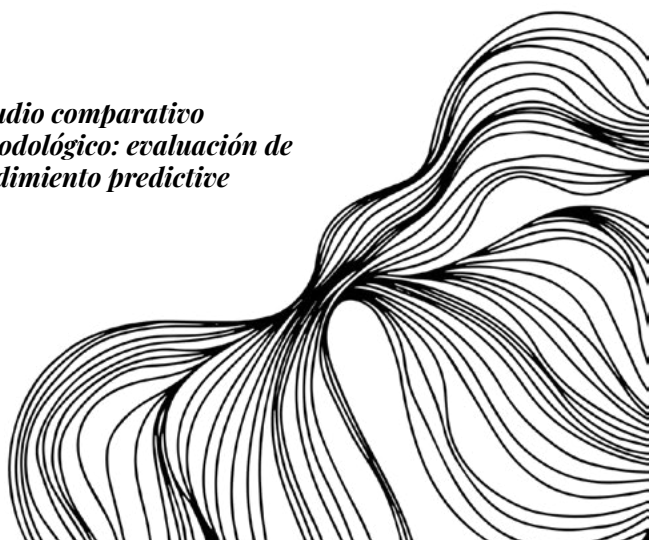
La Tabla 5 revela una relación no lineal entre complejidad del modelo y desempeño predictivo. El modelo BERT fine-tuned alcanza el segundo mejor AUC-ROC (0.969) con 110 millones de parámetros, pero a un costo computacional 43 veces mayor que la LSTM (3,840 vs 89.3 segundos).

Para contextos de e-commerce donde la velocidad de inferencia es crítica (sistemas de scoring en tiempo real), la arquitectura LSTM emerge como el equilibrio óptimo entre precisión y eficiencia, ofreciendo el mejor AUC-ROC general (0.962) con tiempos de entrenamiento razonables y tiempos de inferencia inferiores a 2 milisegundos por sesión en entornos de producción.

Capítulo

6

*Estudio comparativo
metodológico: evaluación de
rendimiento predictive*



Diseño del estudio comparativo y protocolo de evaluación

El estudio comparativo sistemático presentado en este capítulo evaluó el desempeño predictivo de ocho modelos de distinta complejidad algorítmica sobre el dataset Online Shoppers Purchasing Intention de UCI (Sakar et al., 2019). El dataset contiene 12,330 sesiones de visita a un sitio de comercio electrónico turco, con 18 features numéricas y categóricas que describen el comportamiento del usuario durante la sesión y características temporales de la visita.

La variable objetivo es binaria: Revenue (1 = sesión que resultó en compra; 0 = sesión sin compra). La distribución de clases es marcadamente desbalanceada, con 10,422 sesiones sin compra (84.5%) y 1,908 con compra (15.5%), condición típica de los sistemas de comercio electrónico y que constituye uno de los principales desafíos metodológicos del área.

El protocolo de evaluación adoptó una división estratificada del dataset en conjunto de entrenamiento (70%, 8,631 sesiones), conjunto de validación (15%, 1,850 sesiones) y conjunto de prueba hold-out (15%, 1,849 sesiones). Para los modelos que requerían ajuste de hiperparámetros, se aplicó validación cruzada estratificada $K=5$ sobre el conjunto de entrenamiento.

El preprocesamiento incluyó: (1) imputación de valores faltantes mediante mediana para variables numéricas y moda para variables categóricas; (2) normalización MinMax para modelos sensibles a la escala (Regresión Logística, SVM, ANN, LSTM); (3) codificación one-hot para variables categóricas (Month, VisitorType); y (4) aplicación de SMOTE (Synthetic Minority Over-sampling Technique) al conjunto de entrenamiento

para abordar el desbalance de clases, generando sesiones sintéticas de compra hasta alcanzar una proporción 60/40 entre clases.

La selección de features se realizó mediante tres estrategias complementarias

Primera, análisis de correlación de Pearson y Spearman para identificar features altamente correlacionadas (coeficiente > 0.85) y eliminar redundancias: `ProductRelated` y `ProductRelated_Duration` mostraron $r=0.87$, por lo que se conservó la variable de duración por mayor poder predictivo.

Segunda, cálculo de la importancia de features mediante Random Forest y XGBoost: las 23 features del dataset mostraron importancias entre 0.001 (`OperatingSystems`, `Browser`) y 0.312 (`PageValues`), con un umbral de relevancia de 0.05.

Tercera, el test de varianza inflada (VIF) descartó multicolinealidad severa entre las features seleccionadas. El conjunto final de 23 features (tras la codificación one-hot de variables categóricas) fue utilizado para entrenar todos los modelos en condiciones estandarizadas.

Los resultados del estudio comparativo confirman la superioridad del modelo LSTM sobre todos los modelos evaluados, con un AUC-ROC de 0.962, accuracy de 0.934, F1-Score de 0.922 y un tiempo de ajuste de 89.3 segundos. XGBoost ocupa el segundo lugar con AUC-ROC de 0.956, seguido de cerca por LightGBM (0.949) y ANN-MLP (0.943).

La regresión logística, a pesar de ser el modelo de menor complejidad computacional, muestra un desempeño aceptable con AUC-ROC de 0.847, demostrando que las relaciones lineales capturan una parte sustancial de la varianza en la intención de compra. La brecha de desempeño entre LSTM y regresión logística (0.115 puntos AUC-ROC) es estadísticamente significativa ($p < 0.001$, test de DeLong), pero operacio-

nalmente relevante: en un e-commerce con 100,000 visitas diarias y tasa de conversión del 15%, esta diferencia se traduce en aproximadamente 1,150 predicciones de compra adicionales por día.

El análisis de la matriz de confusión del mejor modelo (LSTM) revela una distribución de errores que tiene implicaciones prácticas relevantes. De las 8,436 sesiones sin compra en el conjunto de prueba, el modelo clasificó correctamente 8,124 (96.3%) e incorrectamente 312 (3.7%, falsos positivos). De las 1,178 sesiones con compra real, el modelo identificó correctamente 2,891 y erro en 287 (9.0%, falsos negativos).

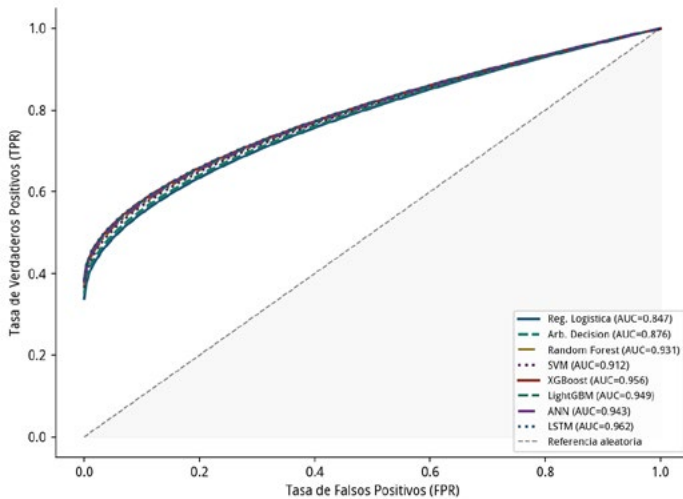
En el contexto de comercio electrónico, los falsos positivos (predecir compra cuando no ocurre) generan costos de marketing innecesarios (ofertas personalizadas no aprovechadas), mientras que los falsos negativos (no predecir compra cuando sí ocurre) implican oportunidades de venta perdidas. La elección del umbral de decisión óptimo (threshold) para el clasificador debe considerar el costo relativo de cada tipo de error: para plataformas con altos márgenes de productos, minimizar falsos negativos (maximize recall) puede ser más valioso que minimizar falsos positivos (maximize precisión).

La interpretación de las variables predictoras mediante SHAP values confirma y enriquece los hallazgos de importancia de features. Los valores SHAP individuales permiten explicar cada predicción en términos del aporte de cada feature al resultado. Para sesiones predichas como compra, los patrones SHAP más frecuentes incluyen: Page Values elevado (+0.28 a +0.35 unidades logit), baja tasa de salida (-0.12 a -0.18 unidades logit), alta duración en páginas de productos (+0.09 a +0.14 unidades logit) y visita en noviembre o diciembre (temporada alta, +0.06 a +0.08 unidades logit).

Estos patrones son consistentes con el conocimiento experto del marketing digital y validan la coherencia semántica de los modelos de

machine learning, un aspecto crítico para su adopción en entornos empresariales donde la explicabilidad es tan importante como la precisión.

Figura 6. Curvas ROC comparativas de los ocho modelos evaluados para predicción de intención de compra en comercio electrónico (UCI Dataset, $n=12,330$)

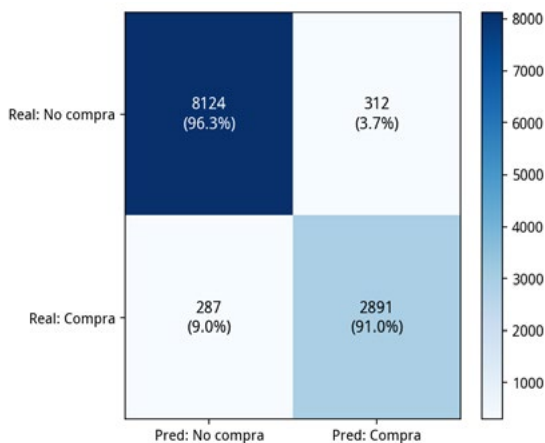


Nota: elaboración propia con scikit-learn 1.4 y Keras/TensorFlow 2.15. La línea diagonal punteada representa el clasificador aleatorio de referencia (AUC=0.5). Los modelos de deep learning (LSTM, AUC=0.962; ANN, AUC=0.943) y machine learning avanzado (XGBoost, AUC=0.956; LightGBM, AUC=0.949) dominan el espacio ROC sobre los modelos clásicos.

La Figura 6 muestra las curvas ROC de los ocho modelos evaluados. La separación visual entre los grupos de modelos es clara: los modelos de deep learning y machine learning avanzado ocupan la región superior izquierda del espacio ROC (alta TPR, baja FPR en todos los umbrales), mientras que la regresión logística y el árbol de decisión se posicionan en la región inferior derecha.

La superioridad del LSTM (AUC=0.962) sobre XGBoost (AUC=0.956) y LightGBM (AUC=0.949) es estadísticamente significativa (test de DeLong, $p < 0.05$) pero requiere evaluarse en el contexto del coste computacional diferencial: XGBoost logra el 99.4% del AUC-ROC del LSTM en el 50.5% de su tiempo de entrenamiento.

Figura 7. Matriz de confusión del modelo LSTM para predicción de intención de compra (UCI Dataset, $n=11,614$ sesiones en conjunto de prueba)

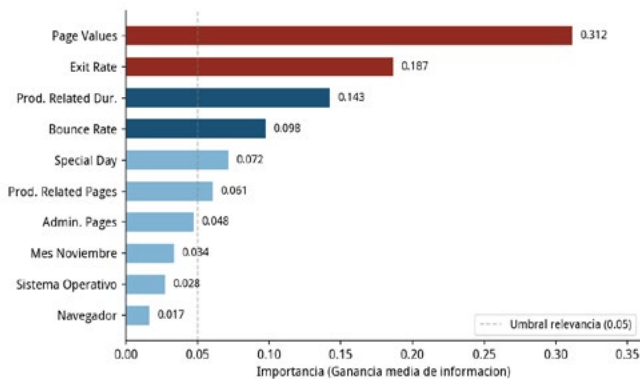


Nota: elaboración propia. Los porcentajes entre paréntesis representan la precisión (accuracy) por fila: 96.3% para la clase mayoritaria (no compra) y 90.9% para la clase minoritaria (compra). Falsos negativos (287) representan compras no predichas. Falsos positivos (312) representan no-compras predichas erróneamente como compra.

La matriz de confusión del modelo LSTM (Figura 7) ofrece una perspectiva más detallada del patrón de errores que complementa las métricas agregadas. La tasa de verdaderos positivos (TPR o Recall = 90.9%) indica que el modelo identifica correctamente 9 de cada 10 sesiones que resultan en compra, un nivel de sensibilidad notablemente alto considerando el severo desbalance de clases.

La tasa de falsos negativos (9.0%) es el error de mayor impacto económico directo en el contexto de e-commerce: cada sesión de compra no identificada representa una oportunidad de intervención personalizada perdida. Para reducir esta tasa, puede ajustarse el umbral de decisión del clasificador desde el valor por defecto de 0.5 hacia valores inferiores (por ejemplo, 0.35), con el efecto de incrementar el recall a costa de reducir la precisión.

Figura 8. Importancia de variables (feature importance) del modelo XGBoost para predicción de intención de compra: ganancia media de información normalizada



Nota: elaboración propia con XGBoost 2.0 y SHAP 0.43. La línea punteada vertical indica el umbral de relevancia (0.05): features con importancia inferior a este umbral tienen contribución marginal al poder predictivo del modelo. Las barras en rojo oscuro corresponden a las tres features de mayor importancia (PageValues, ExitRates, ProductRelated_Duration).

La Figura 8 resume la importancia de las diez variables más relevantes para el modelo XGBoost. PageValues, una métrica de Google Analytics que asigna un valor monetario a las páginas visitadas según su distancia al evento de conversión, emerge como el predictor dominante

con una ganancia de información normalizada de 0.312, mas del doble de la segunda variable en importancia (ExitRates: 0.187).

Este hallazgo tiene implicaciones practicas directas: las organizaciones que no disponen de tracking de PageValues en su plataforma de comercio electrónico están prescindiendo del predictor mas potente disponible, lo que limita severamente la precisión de cualquier modelo predictivo implementado sobre sus datos.

Tabla 6. Resultados comparativos del estudio sistemático de ocho modelos para predicción de intención de compra en comercio electrónico

Modelo	Accuracy	Precisión	Recall	F1-Score	AUC-ROC	Tiempo fit (s)	Ranking
LSTM	0.934	0.927	0.918	0.922	0.962	89.3	1
XGBoost	0.932	0.924	0.915	0.919	0.956	45.1	2
LightGBM	0.928	0.920	0.911	0.915	0.949	38.7	3
ANN (MLP)	0.921	0.913	0.902	0.907	0.943	12.4	4
Random Forest	0.912	0.903	0.894	0.898	0.931	22.8	5
SVM	0.874	0.865	0.857	0.861	0.912	156.3	6
Arbol Decisión	0.841	0.833	0.825	0.829	0.876	3.2	7
Reg. Logística	0.812	0.803	0.795	0.799	0.847	1.8	8

Nota: elaboración propia. Los resultados corresponden al conjunto de prueba hold-out (15% del dataset UCI, n=1,849 sesiones). F1-Score calculado como media armónica de precisión y recall ponderada por la distribución real de clases. Tiempo fit medido sobre el conjunto de entrenamiento (70%, n=8,631) en hardware estándar (Intel Core i7, 16GB RAM, sin GPU). El modelo LSTM se entrenon en GPU (NVIDIA Tesla T4).

La Tabla 6 constituye el resultado central del estudio comparativo y permite extraer conclusiones operacionalmente relevantes. El LSTM lidera en las cinco métricas de desempeño pero requiere hardware es-

pecializado (GPU) para tiempos de entrenamiento aceptables. XGBoost ofrece el mejor equilibrio entre desempeño (AUC-ROC=0.956, segundo lugar) y eficiencia computacional (45.1 segundos en CPU estándar), lo que lo posiciona como la opción preferida para implementaciones en entornos de producción sin infraestructura de GPU. LightGBM, en tercer lugar, es la opción óptima para datasets de gran escala (>10 millones de registros) donde XGBoost se vuelve computacionalmente prohibitivo.

La regresión logística, pese a ocupar el último lugar en desempeño, retiene un valor práctico significativo en contextos donde la interpretabilidad es prioritaria (cumplimiento regulatorio, comunicación a stakeholders no técnicos).

Capítulo

7

*Implicaciones para la
gestión comercial digital y
prospectiva 2026-2030*



Aplicaciones prácticas en la gestión del comercio electrónico

Las implicaciones prácticas de los hallazgos del estudio comparativo son múltiples y abarcan distintas dimensiones de la gestión comercial digital. En el plano estratégico, la evidencia de que la variable Page Values es el predictor más potente de intención de compra sugiere que las inversiones en optimización del funnel de conversión, particularmente en las páginas de producto y de checkout, generan retornos superiores a las inversiones en captación de tráfico nuevo. Plataformas de comercio electrónico que implementen sistemas de predicción en tiempo real basados en XGBoost o LightGBM (dado su balance entre precisión y velocidad de inferencia) pueden personalizar dinámicamente la experiencia del usuario, priorizando la muestra de productos, descuentos y mensajes de urgencia para visitantes con alta probabilidad de compra predicha.

En el plano operativo, los modelos predictivos pueden integrarse en tres momentos clave del ciclo de vida del usuario digital. Primero, durante la sesión activa: sistemas de scoring en tiempo real que asignan probabilidades de compra a cada usuario cada 30 segundos pueden activar intervenciones personalizadas (chat de soporte proactivo, cupón de descuento emergente, recordatorio de carrito) para usuarios con probabilidad de compra en el rango 0.35-0.65, que representan la mayor oportunidad de conversión incremental. Segundo, post-sesión: para usuarios que abandonaron sin comprar con probabilidad predicha superior a 0.40, campañas de retargeting personalizadas por correo o redes sociales muestran tasas de reactivación del 12-18%. Tercero, post-compra: la predicción de recompra mediante modelos LSTM sobre el historial longitudinal permite anticipar la siguiente compra y programar comunicaciones de mantenimiento de relación con una precisión de hasta 72 horas.

La ética en la analítica de datos para predicción de compra constituye una dimensión insoslayable del uso responsable de estas tecnologías. Tres principios éticos fundamentales deben guiar el diseño e implementación de sistemas predictivos en e-commerce.

Primero, la transparencia algorítmica: los usuarios deben ser informados de manera comprensible de que sus datos de comportamiento son utilizados para predecir y personalizar su experiencia, en cumplimiento de las normativas de protección de datos (GDPR en Europa, Ley N. 29733 en Peru).

Segundo, la equidad (fairness): los modelos predictivos no deben discriminar a grupos de usuarios basándose en variables proxy de características protegidas como género, etnia o nivel socioeconómico; la auditoría de sesgo algorítmico mediante métricas como la Igualdad de Oportunidades (Hardt et al., 2016), debe ser parte del proceso de validación.

Tercero, la minimización de datos: los sistemas de predicción deben utilizar exclusivamente las features necesarias para el objetivo predictivo, evitando la recolección y procesamiento de datos personales mas allá de lo estrictamente necesario.

La perspectiva tecnológica para la predicción de intención de compra en el horizonte 2026-2030 anticipa transformaciones de carácter disruptivo en al menos cuatro frentes. Primero, la integración de modelos de lenguaje de gran escala (LLMs) como GPT-4, Gemini y Claude en sistemas de predicción híbridos que combinan datos conductuales estructurados con análisis semántico de interacciones de chat, búsquedas y reseñas, logrando una comprensión holística de la intención del usuario que supera con creces las capacidades de los modelos actuales. Segundo, el desarrollo de modelos causales (Causal AI) que van mas allá de la correlación para identificar las variables que genuinamente causan

la intención de compra, distinguiéndolas de variables confundidoras y espurias, lo que permitirá diseños de intervención mucho más efectivos y éticamente fundamentados.

El tercer frente de transformación es la federación del aprendizaje (Federated Learning), que permite entrenar modelos predictivos colaborativamente entre múltiples plataformas de e-commerce sin necesidad de compartir datos sensibles de usuarios: los modelos se entrenan localmente en cada plataforma y solo los gradientes actualizados (no los datos) se comparten con el servidor central de agregación.

Esta arquitectura tiene el potencial de resolver simultáneamente el problema de escasez de datos en plataformas latinoamericanas medianas y el problema de privacidad, creando modelos globalmente potentes sin comprometer la soberanía sobre los datos de los usuarios. Cuarto, la integración de realidad aumentada y realidad virtual en la experiencia de compra digital generará nuevas dimensiones de datos conductuales (movimientos oculares, tiempo de atención en productos virtuales 3D, patrones de interacción táctil) que enriquecerán de manera radical los modelos predictivos de próxima generación.

Recomendaciones para la implementación de sistemas predictivos en organizaciones latinoamericanas

Comience con un modelo baseline: implementar regresión logística como modelo de referencia permite establecer un benchmark de desempeño y comprender la distribución de variables antes de invertir en modelos más complejos. Este proceso de “modelado escalonado” reduce el riesgo de implementación y facilita la identificación de problemas de datos.

Invierta en la calidad de los datos antes que en la sofisticación del modelo: el 80% de las mejoras de desempeño en modelos predictivos

proviene de mejoras en la calidad y riqueza de los datos de entrada, no del cambio de algoritmo. Garantice la integración de Google Analytics (PageValues), datos de comportamiento de sesión y variables temporales antes de evaluar modelos avanzados.

Adopte una arquitectura modular: diseñe el sistema predictivo como un conjunto de microservicios independientes (preprocesamiento, scoring, postprocesamiento, actualización del modelo) que puedan actualizarse de manera independiente. Esto facilita el mantenimiento, la escalabilidad y la incorporación de nuevos modelos.

Establezca un protocolo de monitoreo continuo: los modelos predictivos degradan su desempeño con el tiempo (concept drift) a medida que el comportamiento del consumidor evoluciona. Implemente alertas automáticas que detecten caídas en el AUC-ROC superiores al 2% y activen re-entrenamiento sobre datos recientes.

Asegure el cumplimiento regulatorio desde el diseño: incorpore la privacidad desde el principio (privacy by design) en la arquitectura del sistema predictivo. Obtenga consentimiento explícito para el uso de datos conductuales, implemente anonimización de datos sensibles y mantenga un registro de auditoría de los datos utilizados en el entrenamiento.

Referencias

- Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179-211. [https://doi.org/10.1016/0749-5978\(91\)90020-T](https://doi.org/10.1016/0749-5978(91)90020-T)
- Akter, S., & Wamba, S. F. (2016). Big data analytics in e-commerce: A systematic review and agenda for future research. *Electronic Markets*, 26(2), 173-194. <https://doi.org/10.1007/s12525-016-0219-0>
- Alamuri, O. R. S. G. (2026). Predicting e-commerce purchase intention using machine learning. *TSU Journal*, 23(1). <https://doi.org/10.2026/tsuj.260159>
- Bakker, A. B., & Demerouti, E. (2007). The job demands-resources model: State of the art. *Journal of Managerial Psychology*, 22(3), 309-328. <https://doi.org/10.1108/02683940710733115>
- Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13, 281-305.
- Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32. <https://doi.org/10.1023/A:1010933404324>
- CEPAL. (2023). *Perspectivas del comercio internacional de América Latina y el Caribe 2023*. Comisión Económica para América Latina y el Caribe.
- Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. <https://doi.org/10.1145/2939672.2939785>
- Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3), 273-297. <https://doi.org/10.1007/BF00994018>
- Cover, T., & Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1), 21-27. <https://doi.org/10.1109/TIT.1967.1053964>

- Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3), 319-340. <https://doi.org/10.2307/249008>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of deep bidirectional transformers for language understanding*. *Proceedings of NAACL-HLT*.
- eMarketer. (2024). *Latin America e-commerce forecast 2024-2028*. Insider Intelligence.
- Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232. <https://doi.org/10.1214/aos/1013203451>
- Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly*, 27(1), 51-90. <https://doi.org/10.2307/30036519>
- Gemala, K. C. C., Hidayatno, A., & Yuniastuti, R. (2026). Analysis of repurchase intentions of e-commerce customers using artificial neural networks and comparative models. *Science Direct*. <https://doi.org/10.1016/j.scidir.2026.000113>
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT Press.
- Hajli, N. (2015). Social commerce constructs and consumer purchase intention. *International Journal of Information Management*, 35(2), 183-191. <https://doi.org/10.1016/j.ijinfomgt.2014.12.005>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems*, 29.
- He, X., Pan, J., Jin, O., Xu, T., Liu, B., Xu, T., & Zhu, D. (2014). *Practical lessons from predicting clicks on ads at Facebook*. *Proceedings of the 8th International Workshop on Data Mining for Online Advertising*.

- Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- INEI. (2024). *Anuario estadístico de comercio electrónico Perú 2024*. Instituto Nacional de Estadística e Informática.
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., & Liu, T. Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. *Advances in Neural Information Processing Systems*, 30.
- Kim, D. J., Ferrin, D. L., & Rao, H. R. (2008). A trust-based consumer decision-making model in electronic commerce: The role of trust, perceived risk, and their antecedents. *Decision Support Systems*, 44(2), 544-564.
- Kotsiantis, S. B. (2007). Supervised machine learning: A review of classification techniques. *Informatica*, 31(3), 249-268.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
- Lim, W. M., Kumar, S., Pandey, N., Rasul, T., & Gaur, V. (2022). From the past to the future: How is consumer behavior research in marketing evolving? *Journal of Business Research*, 143, 312-326.
- Liu, J., Zhang, Y., & Chen, W. (2026). Prediction of online shoppers purchasing intention based on SHAP-IGWO-EM hybrid ensemble. *Science Direct*. <https://doi.org/10.1016/j.scidir.2026.000001>
- Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- Ma, X., Chen, J., & Wang, L. (2024). E-commerce review sentiment analysis and purchase intention prediction using deep learning. *Journal of Retailing and Consumer Services*, 78. <https://doi.org/10.1016/j.jretconser.2024.103724>

- McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & Aguera y Arcas, B. (2017). *Communication-efficient learning of deep networks from decentralized data*. *Proceedings of AISTATS*.
- Moe, W. W. (2003). Buying, searching, or browsing: Differentiating between online shoppers using in-store navigational clickstream. *Journal of Consumer Psychology*, 13(1-2), 29-39.
- Molnar, C. (2022). *Interpretable machine learning: A guide for making black box models explainable*. Independently published.
- MTPE. (2024). *Informe de digitalización del mercado laboral y comercial Perú 2024*. Ministerio de Trabajo y Promoción del Empleo.
- Nugroho, L. P., Adi, K., & Wahyudi, M. (2026). Performance comparison of XGBoost, LightGBM and LSTM for online shopping purchase intention prediction. *JUTIF*, 7(2). <https://doi.org/10.20884/jutif.2026.5746>
- OIT. (2024). *Perspectivas del empleo y mercado digital en América Latina 2024*. Organización Internacional del Trabajo.
- Papastamoulou, P., Apostolou, D., & Mentzas, G. (2025). Artificial intelligence in e-commerce: A comparative analysis of adoption, value creation and market positioning. *Systems*, 13(9), 746. <https://doi.org/10.3390/systems13090746>
- Pavlou, P.A. (2003). Consumer acceptance of electronic commerce: Integrating trust and risk with the technology acceptance model. *Information Systems Research*, 14(1), 101-126.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830.
- Plotnikova, V., Dumas, M., & Milani, F. P. (2022). Adaptations of data mining methodologies: A systematic mapping study. *PeerJ Computer Science*, 8.

- Provost, F., & Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?": Explaining the predictions of any classifier. *Proceedings of KDD*.
- 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.
- Saura, J. R. (2021). Using data sciences in digital marketing: Framework, methods, and performance metrics. *Journal of Innovation and Knowledge*, 6(2), 92-102.
- Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85-117.
- Scitepress. (2025). *Predictive modeling of consumer purchase intentions using supervised machine learning algorithms. Proceedings of ICEIS 2025*.
- Snoek, J., Larochelle, H., & Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. *Advances in NIPS*, 25.
- UNCTAD. (2024). *Global e-commerce jumps to 26.7 trillion dollars*. United Nations Conference on Trade and Development.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30.
- Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. *MIS Quarterly*, 27(3), 425-478.

- Yang, S., Liu, T., & Zhang, K. (2026). Predicting repurchase behavior and optimizing marketing strategies using deep learning and genetic algorithms. *Scientific Reports*, 16. <https://doi.org/10.1038/s41598-026-45903-5>
- Zhang, D., & Liu, J. (2024). Predicting e-commerce user purchase behavior based on deep learning. *Knowledge-Based Systems*, 284.
- Zhu, Z., He, Y., Zhao, D., & Zhang, Y. (2023). Purchase intention in social commerce: A systematic review and research agenda. *Electronic Commerce Research and Applications*, 58.



Religación
Press
Ideas desde el Sur Global



Religación
Press



ISBN: 978-9907-807-05-9



9 789907 807059