Campo / Field	Contenido / Content
Tipo de contribución / Article type	Artículo original de investigación basado en análisis secundario predictivo de datos abiertos
Título en español	Analítica de sesiones de comercio electrónico para decisiones de conversión: desempeño predictivo, calibración y dependencia de métricas próximas al resultado
Title in English	E-Commerce Session Analytics for Conversion Decisions: Predictive Performance, Calibration, and Dependence on Outcome-Proximal Metrics
Autorías / Authors	LENIN HOCHIMIN TENECELA-CALDERON — Escuela Superior Politécnica del Litoral (ESPOL), Ecuador — ORCID: https://orcid.org/0009-0006-7086-878X
Correspondencia / Correspondence	LENIN HOCHIMIN TENECELA-CALDERON — lenin.tenecela@asesoreseducativos-ec.com

Resumen

La analítica de comercio electrónico suele presentar la predicción de compra como un problema de clasificación, pero una métrica elevada no garantiza que el modelo pueda apoyar una decisión en el momento requerido. Este estudio evaluó cuánto cambia el desempeño al incorporar grupos de variables plausiblemente disponibles en distintas fases de una sesión y, en particular, una métrica próxima al resultado: PageValues. Se realizó un análisis secundario de 12.330 sesiones anónimas del conjunto Online Shoppers Purchasing Intention, con 1.908 conversiones (15,47 %). Se compararon regresión logística penalizada y árboles potenciados mediante validación cruzada estratificada de cinco pliegues repetida tres veces. Los predictores se organizaron en cuatro paneles acumulativos: contexto (P1), navegación (P2), salida (P3) y completo con PageValues (P4). Se evaluaron ROC AUC, precisión promedio, Brier, log-loss, calibración, escenarios de umbral y beneficio neto. Con P3, la regresión logística obtuvo ROC AUC de 0,751 (IC 95 %: 0,740–0,761) y precisión promedio de 0,325 (0,306–0,344); con P4 alcanzó 0,896 (0,888–0,904) y 0,642 (0,615–0,666). En árboles potenciados, el cambio fue de 0,788 a 0,936 en ROC AUC y de 0,374 a 0,757 en precisión promedio. La calibración interna fue cercana a la ideal, pero no sustituye una validación externa. En el umbral 0,20, P4 redujo la proporción de sesiones accionables de 30,9 % a 19,4 % respecto de P3 y elevó el valor predictivo positivo de 29,7 % a 55,4 %. El incremento desproporcionado al añadir PageValues demuestra dependencia predictiva de una variable cuya disponibilidad temporal no puede confirmarse con el conjunto. Por tanto, el uso empresarial responsable exige registrar el instante de cálculo de cada métrica, validar temporalmente el modelo, definir costos y probar el efecto de las intervenciones antes de atribuir mejoras de conversión.

Palabras clave: analítica web; comercio electrónico; intención de compra; calibración; aprendizaje automático; toma de decisiones; conversión.

Abstract

E-commerce analytics often frames purchase prediction as a classification problem, yet a high performance metric does not ensure that a model can support a decision at the required time. This study assessed how predictive performance changes when adding groups of variables plausibly available at different stages of a session, with particular attention to an outcome-proximal metric, PageValues. We conducted a secondary analysis of 12,330 anonymous sessions from the Online Shoppers Purchasing Intention dataset, including 1,908 conversions (15.47%). Penalized logistic regression and boosted trees were evaluated through five-fold stratified cross-validation repeated three times. Predictors were organized into four cumulative panels: context (P1), navigation (P2), exit (P3), and complete with PageValues (P4). We assessed ROC AUC, average precision, Brier score, log-loss, calibration,

threshold scenarios, and net benefit. With P3, logistic regression achieved a ROC AUC of 0.751 (95% CI: 0.740–0.761) and average precision of 0.325 (0.306–0.344); with P4, these increased to 0.896 (0.888–0.904) and 0.642 (0.615–0.666). For boosted trees, ROC AUC increased from 0.788 to 0.936 and average precision from 0.374 to 0.757. Internal calibration was close to ideal but does not replace external validation. At a 0.20 threshold, P4 reduced the actionable share of sessions from 30.9% to 19.4% compared with P3 and increased positive predictive value from 29.7% to 55.4%. The disproportionate gain after adding PageValues reveals predictive dependence on a feature whose temporal availability cannot be confirmed from the dataset. Responsible business use therefore requires recording when each metric is computed, temporal validation, explicit cost definitions, and intervention testing before claiming conversion improvements.

Keywords: web analytics; e-commerce; purchase intention; calibration; machine learning; decision-making; conversion.

1. Introducción

Las plataformas de comercio electrónico producen señales continuas sobre procedencia del tráfico, dispositivos, páginas visitadas, duración, rebotes, salidas y transacciones. Esta abundancia favorece una narrativa atractiva: si una organización identifica anticipadamente las sesiones con mayor probabilidad de compra, podría priorizar recomendaciones, asistencia, incentivos o recuperación. Sin embargo, entre una predicción técnicamente correcta y una decisión empresarial útil existen varias condiciones intermedias. La variable debe estar disponible antes de ejecutar la acción; la probabilidad debe conservar una relación adecuada con la frecuencia observada; el umbral debe reflejar recursos y costos; y la intervención desencadenada debe producir un beneficio mayor que sus efectos adversos.

La literatura ha desarrollado modelos de intención de compra con redes neuronales, métodos secuenciales, selección de variables y ensambles. El conjunto Online Shoppers Purchasing Intention, presentado por Sakar et al. (2019) y distribuido por UCI Machine Learning Repository (2018), se convirtió en un referente porque ofrece 12.330 sesiones con variables administrativas, informativas, de producto, contexto técnico y métricas derivadas de analítica web. Estudios posteriores han informado buen desempeño con gradiente potenciado y otras familias de aprendizaje automático (Abdullah-All-Tanvir et al., 2023; Trivedi et al., 2024). Los enfoques secuenciales también subrayan que el orden y el momento de las interacciones importan para reconocer intención (Liu et al., 2021).

El énfasis habitual en exactitud o discriminación puede ocultar tres problemas. Primero, las conversiones representan una minoría; en este conjunto, solo 15,47 % de las sesiones concluyen con compra. Bajo desbalance, una exactitud alta puede coexistir con baja utilidad para identificar el evento, y la precisión promedio resulta informativa porque relaciona precisión y sensibilidad con la prevalencia positiva (Saito & Rehmsmeier, 2015). Segundo, dos modelos con ROC AUC similar pueden entregar probabilidades de distinta confiabilidad. La calibración evalúa si, por ejemplo, un grupo con riesgo estimado de 20 % convierte aproximadamente en esa proporción; esta propiedad es esencial cuando la predicción sustenta un umbral (Van Calster et al., 2019). Tercero, una variable puede contener información que no existiría al momento de la decisión. Kapoor y Narayanan (2023) identifican las variables ilegítimas y la filtración temporal como amenazas capaces de inflar el desempeño. En escenarios con variables que evolucionan, la disponibilidad y el costo de adquisición deben formar parte de la evaluación (von Kleist et al., 2025).

PageValues requiere atención particular. La documentación del conjunto la define como el valor promedio de las páginas visitadas antes de completar una transacción. En herramientas de analítica, el valor de página se deriva del valor de transacciones y objetivos atribuido a páginas vistas. Por construcción, puede incorporar información acumulada de toda la sesión o estadísticas históricas estrechamente relacionadas con la conversión. El archivo no registra el instante exacto en que se calculó ni garantiza que estuviera disponible cuando una intervención temprana habría sido posible. Por ello, incluirla sin contraste puede confundir capacidad de clasificación retrospectiva con capacidad de decisión prospectiva.

La analítica empresarial también enfrenta restricciones organizacionales. Los tableros y modelos no generan valor automáticamente: necesitan definiciones estables, calidad de datos, competencias, gobernanza y alineación con

decisiones concretas (Gupta et al., 2020; Gudigantala et al., 2023). Estudios sobre pequeñas y medianas empresas describen dificultades para convertir datos en decisiones repetibles, como fragmentación de fuentes, escasez de capacidades y falta de integración estratégica (Tawil et al., 2024). La asociación entre analítica y desempeño organizacional tampoco autoriza a atribuir causalidad a un algoritmo aislado (Gupta et al., 2021; Almatrafi & Alharbi, 2023).

Este estudio no busca declarar un modelo listo para producción ni demostrar que una intervención aumenta ventas. Su objetivo general es evaluar el desempeño predictivo, la calibración y las implicaciones de decisión de modelos construidos con paneles acumulativos de variables de sesión. Los objetivos específicos son: describir la conversión por segmentos disponibles; comparar una regresión logística penalizada y árboles potenciados; cuantificar la ganancia al añadir variables de navegación, salida y PageValues; explorar escenarios de umbral y beneficio neto; y formular condiciones mínimas para una validación empresarial responsable.

La pregunta central es: ¿cuánto cambia la capacidad predictiva cuando se incorporan variables progresivamente próximas al resultado y qué riesgos introduce ese cambio para decisiones de conversión? La contribución reside en desplazar la comparación desde “qué algoritmo logra la cifra más alta” hacia “qué información estaba disponible, qué tan confiable es la probabilidad y qué decisión podría sostener”.

2. Metodología

2.1. Diseño y fuente

Se realizó un estudio cuantitativo, observacional y secundario con datos abiertos. La fuente fue Online Shoppers Purchasing Intention Dataset, alojada por UCI bajo DOI 10.24432/C5F88Q. El archivo conserva 12.330 observaciones y 18 columnas: 17 predictores y Revenue como resultado binario. Cada fila representa una sesión; no se incluye un identificador de usuario, de cliente ni de sesión original. La licencia declarada por la fuente es CC BY 4.0.

El corte efectivo de evidencia bibliográfica y documental fue el 31 de agosto de 2025. Ninguna fuente posterior a esa fecha se utilizó para formular el método, interpretar resultados o sostener conclusiones.

El archivo descargado se preservó sin modificaciones y se registró una huella SHA-256 de 2972E6184D3AD7BEAAA831D9FC2B059DC3EE29DF69D1EC593C466A5CD8485D14. La copia analítica añadió únicamente un número secuencial row_id para trazabilidad. No se imputaron datos porque el conjunto no presentaba valores faltantes.

2.2. Variables y paneles de disponibilidad

Revenue tomó valor 1 cuando la sesión finalizó con compra y 0 en caso contrario. Los predictores se agruparon antes del modelado en cuatro paneles acumulativos:

Panel	Variables incorporadas	Interpretación temporal
P1 Contexto	SpecialDay, Month, OperatingSystems, Browser, Region, TrafficType, VisitorType, Weekend	Señales de contexto potencialmente disponibles al inicio o muy temprano
P2 Navegación	P1 + conteos y duraciones de páginas Administrative, Informational y ProductRelated	Señales acumuladas durante la navegación
P3 Salida	P2 + BounceRates y ExitRates	Métricas derivadas del comportamiento de páginas, potencialmente tardías o históricas
P4 Completo	P3 + PageValues	Métrica próxima al resultado; momento de disponibilidad no verificable

La clasificación no afirma que todas las variables de un panel estén disponibles exactamente al mismo tiempo. Es una prueba de sensibilidad conceptual basada en su proximidad plausible al desenlace. En particular, P3 no se denomina “temprano”, porque las tasas de rebote y salida pueden depender de procesos históricos o de la finalización de la visita. P4 se separó para medir cuánto desempeño dependía de PageValues.

2.3. Auditoría y preparación

Se comprobaron dimensiones, tipos, valores faltantes, frecuencia del resultado, categorías y duplicación exacta. Se encontraron 125 filas adicionales idénticas a una observación previa considerando predictores y resultado. Como el conjunto no ofrece identificadores, no fue posible establecer si representaban duplicados técnicos o sesiones legítimas con el mismo vector. El análisis principal retuvo todas las filas, conforme a la fuente. Como sensibilidad se repitió la evaluación logística de P3 y P4 con 12.205 vectores únicos.

Las variables numéricas se estandarizaron dentro de cada partición de entrenamiento para la regresión logística. Las categóricas se codificaron mediante indicadores, permitiendo categorías desconocidas en evaluación. Todas las transformaciones se ajustaron exclusivamente con datos de entrenamiento para evitar contaminación. En árboles potenciados se aplicó una representación compatible de variables categóricas y numéricas dentro del mismo flujo.

2.4. Modelos y validación

Se prespecificaron dos familias complementarias. La regresión logística con penalización L2 ofreció una línea base interpretable y estable; el parámetro C se seleccionó dentro de cada conjunto de entrenamiento. Los árboles potenciados capturaron relaciones no lineales e interacciones sin presuponer una forma funcional lineal. El objetivo no fue agotar algoritmos, sino contrastar una familia regularizada y otra flexible.

La evaluación utilizó validación cruzada estratificada de cinco pliegues repetida tres veces. En cada repetición, cada sesión fue evaluada fuera de la partición usada para ajustar el modelo. Para la regresión logística, la selección de C ocurrió dentro de los datos de entrenamiento, separada de la evaluación exterior. Se promediaron las probabilidades fuera de muestra de las tres repeticiones para obtener una estimación por sesión.

El diseño se alineó, de forma adaptada al ámbito comercial, con principios de reporte transparente de estudios predictivos: descripción de datos, preprocesamiento, validación, métricas, subgrupos, limitaciones y materiales reproducibles (Collins et al., 2024). La referencia metodológica no convierte el análisis en un estudio clínico.

2.5. Métricas

La discriminación se evaluó mediante ROC AUC. Debido al desbalance, se añadió precisión promedio (average precision, AP), cuya referencia no informativa se aproxima a la prevalencia de 0,155. El desempeño probabilístico global se cuantificó con Brier y log-loss; valores menores indican mejor ajuste.

La calibración débil se resumió con intercepto y pendiente estimados sobre probabilidades fuera de muestra. Un intercepto cercano a 0 indica ausencia de sobreestimación o subestimación global, y una pendiente cercana a 1 indica dispersión apropiada. Se construyeron además diez grupos de igual tamaño para comparar probabilidad media y frecuencia observada. Estas estimaciones son internas y no demuestran calibración en otra empresa o periodo.

Se calcularon intervalos de confianza de 95 % mediante 1.000 remuestreos bootstrap de sesiones. Se informaron cambios absolutos entre paneles. Para la regresión logística se exploraron umbrales de 0,10, 0,20 y 0,30, reportando sensibilidad, especificidad, valor predictivo positivo (VPP), valor predictivo negativo (VPN) y proporción de sesiones clasificadas como accionables.

El beneficio neto se calculó entre umbrales de 0,05 y 0,50 y se comparó con actuar sobre todas las sesiones y no actuar sobre ninguna, siguiendo el marco de Vickers y Elkin (2006). Esta curva expresa intercambio relativo entre verdaderos y falsos positivos; no sustituye una función monetaria ni demuestra el efecto de la acción.

2.6. Análisis descriptivo y subgrupos

Se estimaron tasas de conversión e intervalos aproximados de 95 % por mes, tipo de visitante y fin de semana. Para P4 se exploró el desempeño por tipo de visitante y mes cuando existían eventos suficientes. El propósito fue detectar heterogeneidad visible, no certificar equidad ni estabilidad. No había atributos demográficos protegidos ni geografía interpretable.

2.7. Reproducibilidad y ética

El expediente conserva la fuente, el script, parámetros, resultados tabulares, predicciones fuera de muestra, figuras y un cuaderno analítico. La semilla de aleatoriedad se fijó para reproducibilidad. El análisis utilizó datos públicos anónimos y no reclutó participantes ni recopiló información nueva. No fue aplicable una nueva revisión ética para este análisis secundario de datos anónimos. El uso de herramientas de inteligencia artificial se declara al final y no reemplaza la verificación científica autoral.

3. Resultados

3.1. Muestra y conversión

De 12.330 sesiones, 1.908 finalizaron con compra y 10.422 no lo hicieron, para una conversión global de 15,47 %. No hubo datos faltantes. Los visitantes nuevos representaron 1.694 sesiones y convirtieron en 24,91 %; los recurrentes sumaron 10.551 sesiones y convirtieron en 13,93 %. La categoría Other fue pequeña (n = 85; 18,82 %), por lo que no se usó para evaluar desempeño de subgrupo.

Las sesiones de fin de semana mostraron una conversión de 17,40 % frente a 14,89 % en días laborables. La variación mensual fue amplia: 1,63 % en febrero, alrededor de 10 % en marzo, mayo y junio, 20,95 % en octubre y 25,35 % en noviembre. Estas diferencias son descriptivas. El archivo no incluye año, promociones, inventario o adquisición de clientes, por lo que no permiten atribuir estacionalidad causal.

Tabla 1. Conversión por segmentos seleccionados

Segmento	Sesiones	Conversiones	Tasa	IC 95 %
Total	12.330	1.908	15,47 %	—
Visitante nuevo	1.694	422	24,91 %	22,85–26,97 %
Visitante recurrente	10.551	1.470	13,93 %	13,27–14,59 %
Otro visitante	85	16	18,82 %	10,51–27,13 %
Día laborable	9.462	1.409	14,89 %	14,17–15,61 %
Fin de semana	2.868	499	17,40 %	16,01–18,79 %

ISSN 2838-5527 - DOI: https://doi.org/10.64747/dxcy5h52 - p. 17

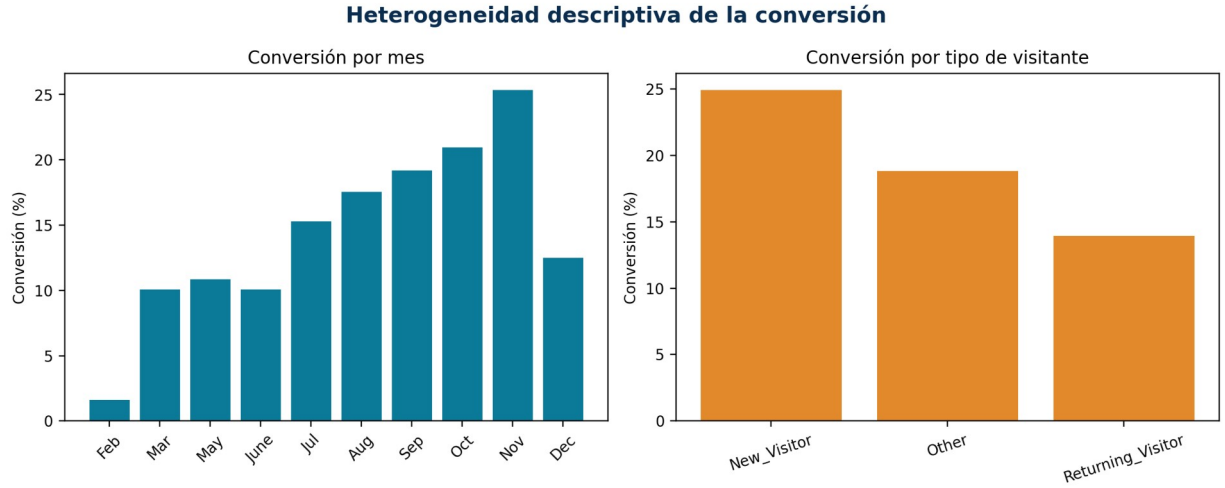

Figura 1. Conversión observada por mes, tipo de visitante y fin de semana. Las barras representan tasas descriptivas; no implican efectos causales.

3.2. Desempeño por panel

En la regresión logística, P1 produjo ROC AUC de 0,692 y AP de 0,267. Añadir navegación en P2 elevó estas métricas a 0,717 y 0,296. P3 alcanzó 0,751 (IC 95 %: 0,740–0,761) y 0,325 (0,306–0,344). El avance acumulado de P1 a P3 fue moderado: 0,059 en ROC AUC y 0,058 en AP.

Al incorporar PageValues, P4 alcanzó ROC AUC de 0,896 (0,888–0,904) y AP de 0,642 (0,615–0,666). El salto exclusivo P3–P4 fue 0,145 en ROC AUC y 0,317 en AP. También mejoraron Brier, de 0,118 a 0,085, y log-loss, de 0,375 a 0,292.

Los árboles potenciados obtuvieron mayor discriminación. P1 fue comparable con la logística (ROC AUC 0,693; AP 0,264), P2 alcanzó 0,769 y 0,350, y P3 0,788 (0,777–0,798) y 0,374 (0,352–0,396). P4 llegó a 0,936 (0,930–0,941) y 0,757 (0,737–0,774). El salto P3–P4 fue 0,148 en ROC AUC y 0,383 en AP. Brier cayó de 0,112 a 0,068.

Tabla 2. Desempeño predictivo fuera de muestra

Modelo	Panel	ROC AUC (IC 95 %)	AP (IC 95 %)	Brier	Log-loss	Intercepto	Pendiente
Logística L2	P1 Contexto	0,692 (0,680–0,705)	0,267 (0,251–0,284)	0,123	0,401	-0,017	0,989
Logística L2	P2 Navegación	0,717 (0,706–0,729)	0,296 (0,278–0,315)	0,121	0,394	-0,028	0,981
Logística L2	P3 Salida	0,751 (0,740–0,761)	0,325 (0,306–0,344)	0,118	0,375	-0,009	0,994
Logística L2	P4 Completo	0,896 (0,888–0,904)	0,642 (0,615–0,666)	0,085	0,292	0,020	1,012
Árboles potenciados	P1 Contexto	0,693 (0,681–0,706)	0,264 (0,249–0,280)	0,123	0,400	0,066	1,044
Árboles potenciados	P2 Navegación	0,769 (0,759–0,780)	0,350 (0,331–0,370)	0,115	0,366	0,126	1,092
Árboles potenciados	P3 Salida	0,788 (0,777–0,798)	0,374 (0,352–0,396)	0,112	0,356	0,086	1,062
Árboles potenciados	P4 Completo	0,936 (0,930–0,941)	0,757 (0,737–0,774)	0,068	0,221	0,055	1,060

ISSN 2838-5527 - DOI: https://doi.org/10.64747/dxcy5h52 - p. 18

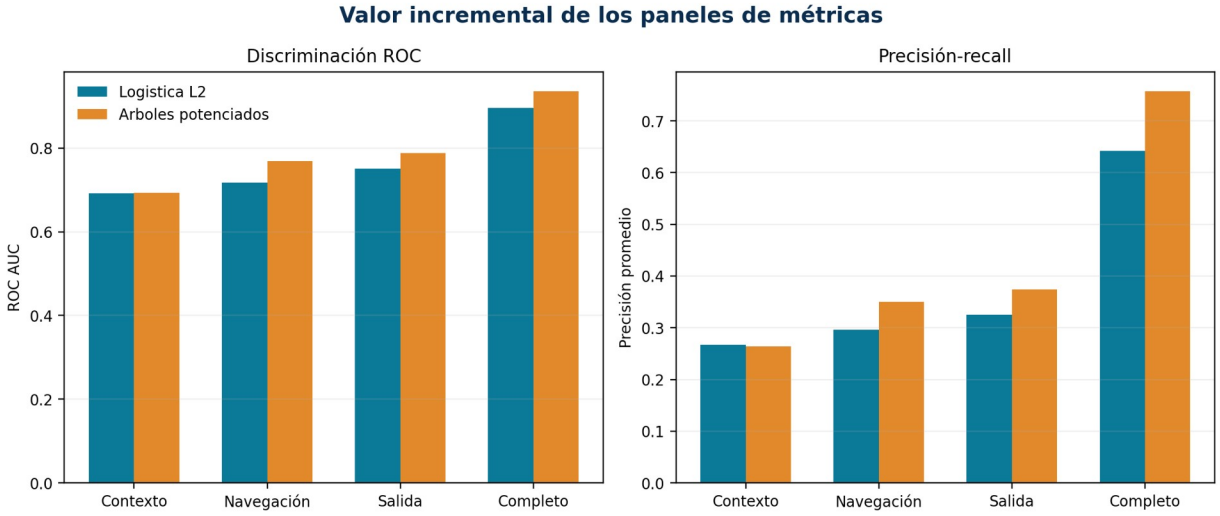

Figura 2. Desempeño por modelo y panel de variables. P4 incorpora PageValues y concentra la mayor ganancia incremental.

3.3. Calibración

Las estimaciones internas de calibración fueron próximas a los valores ideales. En logística, los interceptos oscilaron entre -0,028 y 0,020 y las pendientes entre 0,981 y 1,012. En árboles potenciados, los interceptos fueron positivos entre 0,055 y 0,126 y las pendientes entre 1,044 y 1,092, compatibles con probabilidades algo moderadas y subestimación global pequeña dentro de la muestra evaluada.

Las curvas agrupadas mostraron correspondencia razonable entre riesgo previsto y observado, con discrepancias localizadas en extremos donde existían menos observaciones efectivas. La aparente calibración favorable debe interpretarse junto con el diseño: todas las particiones proceden del mismo conjunto histórico. Un cambio de país, empresa, interfaz, canal, mezcla de visitantes o prevalencia podría alterar el intercepto, la pendiente y la utilidad de los umbrales.

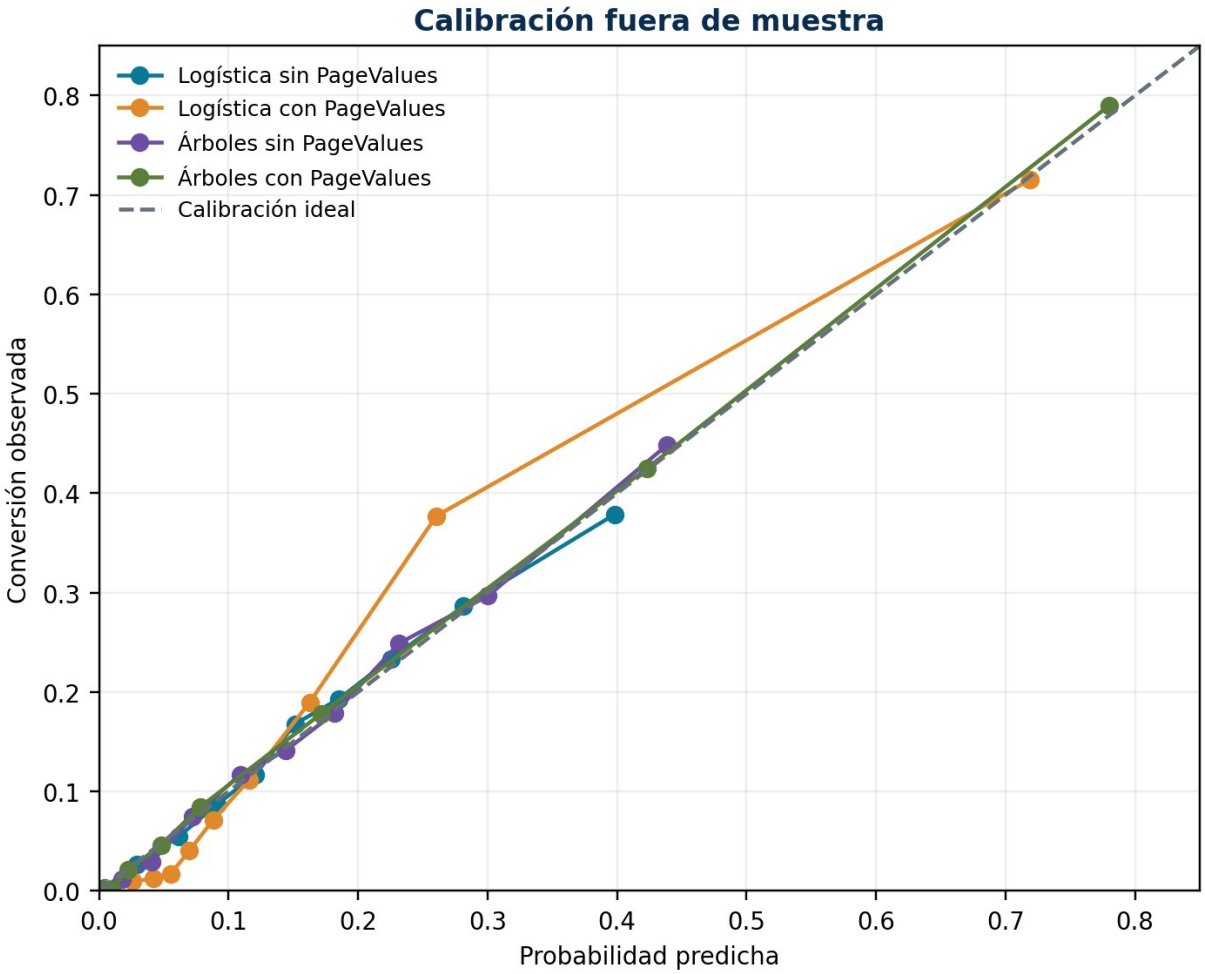

Figura 3. Calibración fuera de muestra de la regresión logística por panel. La diagonal representa correspondencia ideal.

3.4. Escenarios de decisión

Los escenarios de umbral mostraron que el panel elegido afecta tanto el número de sesiones intervenidas como la concentración de conversiones. Con P3 y umbral 0,10, el modelo habría marcado 61,9 % de las sesiones, con sensibilidad 90,3 % y VPP 22,6 %. P4 mantuvo sensibilidad similar (90,0 %), redujo la proporción accionable a 39,9 % y elevó el VPP a 34,9 %.

En umbral 0,20, P3 marcó 30,9 % de las sesiones: sensibilidad 59,3 %, especificidad 74,3 % y VPP 29,7 %. P4 marcó 19,4 %, con sensibilidad 69,4 %, especificidad 89,8 % y VPP 55,4 %. En umbral 0,30, ambos paneles marcaron cerca de 12 %; P4 presentó sensibilidad 54,8 % y VPP 68,3 %, frente a 28,8 % y 36,5 % con P3.

Tabla 3. Escenarios de umbral para regresión logística

Panel	Umbral	Sensibilidad	Especificidad	VPP	VPN	Sesiones accionables
P3 Salida	0,10	90,3 %	43,3 %	22,6 %	96,0 %	61,9 %
P3 Salida	0,20	59,3 %	74,3 %	29,7 %	90,9 %	30,9 %
P3 Salida	0,30	28,8 %	90,8 %	36,5 %	87,5 %	12,2 %
P4 Completo	0,10	90,0 %	69,2 %	34,9 %	97,4 %	39,9 %
P4 Completo	0,20	69,4 %	89,8 %	55,4 %	94,1 %	19,4 %
P4 Completo	0,30	54,8 %	95,3 %	68,3 %	92,0 %	12,4 %

ISSN 2838-5527 - DOI: https://doi.org/10.64747/dxcy5h52 - p. 20

La curva de decisión indicó beneficio neto positivo para modelos en una franja de umbrales, con ventaja más amplia de P4. Sin embargo, esta comparación asume que identificar un verdadero positivo tiene valor y que el costo de un falso positivo puede representarse por el umbral. No incorpora precio del incentivo, margen, fatiga por mensajes, canibalización, privacidad ni capacidad del equipo.

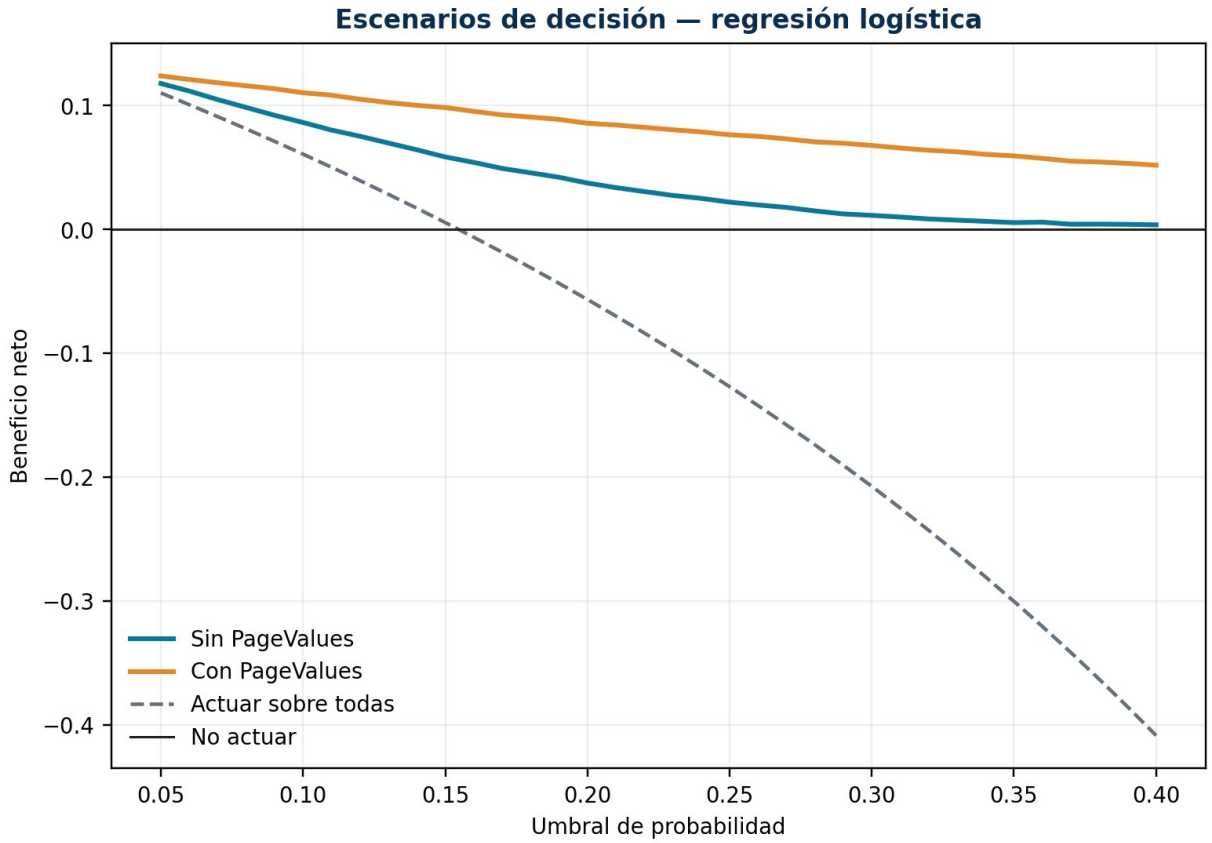

Figura 4. Curvas de decisión para regresión logística. El beneficio neto es exploratorio y no equivale a retorno económico.

3.5. Sensibilidad y subgrupos

Al excluir vectores exactos repetidos, P3 obtuvo ROC AUC 0,746, AP 0,326, Brier 0,119 y log-loss 0,380; P4 alcanzó 0,894, 0,642, 0,086 y 0,295, respectivamente. La conclusión sobre la dependencia de PageValues permaneció.

En P4, la ROC AUC fue 0,897 en visitantes nuevos y 0,893 en recurrentes. La AP fue mayor en nuevos (0,809) que en recurrentes (0,582), en parte porque la prevalencia de conversión también fue mayor. Por mes, la ROC AUC varió y algunas estimaciones extremas correspondieron a grupos pequeños o con muy pocos eventos; febrero, con tres conversiones, no permite una interpretación estable. Estas diferencias justifican validación prospectiva por periodos y segmentos, no una conclusión de inequidad o superioridad estructural.

ISSN 2838-5527 - DOI: https://doi.org/10.64747/dxcy5h52 - p. 21

4. Discusión

4.1. Hallazgo principal

El resultado central no es simplemente que los árboles potenciados superaron a la regresión logística. La observación más relevante es la forma del incremento: los paneles de contexto, navegación y salida añadieron capacidad de manera gradual, mientras que PageValues produjo un salto mucho mayor en ambas familias. En logística, P3–P4 aumentó 0,145 la ROC AUC y 0,317 la AP; en árboles, 0,148 y 0,383. La consistencia entre algoritmos y el análisis sin vectores repetidos indica que el patrón no depende de una sola especificación.

Este hallazgo puede leerse de dos maneras complementarias. Retrospectivamente, PageValues contiene señal muy informativa y mejora la clasificación. Prospectivamente, esa misma cercanía al resultado plantea una pregunta de legitimidad temporal: ¿la variable existiría, con el mismo valor, antes de elegir la intervención? El conjunto no permite responderla. Por ello, reportar únicamente P4 generaría una imagen incompleta de la capacidad operativa.

La literatura específica ha documentado alto desempeño con gradiente potenciado en el mismo conjunto (Abdullah-All-Tanvir et al., 2023), y estudios comparativos han destacado el potencial de aprendizaje automático para identificar compra (Trivedi et al., 2024). Nuestros resultados son compatibles con ese potencial, pero añaden una restricción: la comparación debe definir el momento de predicción. Los métodos secuenciales reconocen explícitamente que la intención evoluciona durante la sesión (Liu et al., 2021); una matriz final agregada no conserva esa secuencia.

4.2. Discriminación, calibración y utilidad

La discriminación responde si las sesiones compradoras tienden a recibir mayor puntaje que las no compradoras. No garantiza que un riesgo de 0,20 signifique 20 % de conversión ni que actuar sea rentable. La calibración interna cercana a intercepto 0 y pendiente 1 es favorable, especialmente en logística, pero es esperable que se deteriore ante cambios de prevalencia y medición (Van Calster et al., 2019). La ausencia de fecha exacta impidió una división temporal, y no existe una segunda empresa para validación externa.

La precisión promedio aportó una lectura necesaria. La AP de P3 en logística fue 0,325 frente a prevalencia 0,155; el modelo ordena casos mejor que una referencia aleatoria, pero una gran proporción de alertas seguiría siendo falsa. P4 elevó la AP a 0,642. Esta métrica depende de la frecuencia de conversión; no debe compararse sin contexto entre empresas con prevalencias distintas.

Los escenarios muestran el vínculo entre probabilidad y carga operativa. Si una empresa pudiera contactar como máximo a una quinta parte de las sesiones, P4 con umbral 0,20 habría marcado 19,4 % y concentrado 1.324 de las 1.908 conversiones. P3 habría requerido actuar sobre 30,9 % para detectar 1.131 conversiones. Pero esa superioridad solo es accionable si P4 está calculado antes de contactar, si la predicción se actualiza con latencia aceptable y si la acción produce valor.

La curva de decisión evita fijar un único umbral universal y reconoce el costo relativo de falsos positivos (Vickers & Elkin, 2006). Aun así, su beneficio neto es una medida abstracta. En marketing, un falso positivo puede significar descuento innecesario, mensaje no deseado o gasto de atención; un falso negativo puede ser una oportunidad no tratada. Los costos varían por margen, canal y objetivo. Convertir beneficio neto en decisión financiera exige una función de valor local.

4.3. De la predicción a una arquitectura de decisión

Una implementación responsable debería separar cuatro momentos. Primero, al inicio de la sesión, P1 puede orientar decisiones de baja fricción, aunque su discriminación es limitada. Segundo, durante la navegación, P2 puede actualizar el riesgo a medida que aparecen interacciones. Tercero, variables de salida o métricas históricas deben etiquetarse con su latencia y ventana de cálculo. Cuarto, el resultado final se usa para entrenamiento y evaluación, no como señal contemporánea.

Esta separación coincide con una visión de analítica digital orientada a decisiones y no solo a reportes (Gupta et al., 2020). También responde a marcos que vinculan IA con valor empresarial mediante problema, decisión, acción y resultado medible (Gudigantala et al., 2023). Sin esa cadena, un tablero puede exhibir métricas correctas sin cambiar prácticas.

La analítica de marketing aporta valor cuando conecta seguimiento, interpretación y ajuste estratégico, no cuando acumula indicadores sin una decisión asociada (Al Adwan et al., 2023). Del mismo modo, los sistemas de recomendación basados en aprendizaje automático requieren convertir una predicción en una regla verificable y monitoreada, con supuestos y límites explícitos (Gangadharan et al., 2025).

El registro operativo mínimo debería incluir timestamp del evento, timestamp de cálculo de cada predictor, versión del modelo, probabilidad emitida, umbral, acción, costo, elegibilidad y resultado. Esto permitiría validar si la característica era conocida antes de actuar y reduciría el riesgo de filtración descrito por Kapoor y Narayanan (2023). Para variables costosas o tardías, la evaluación puede tratar la adquisición como parte de la política, en lugar de asumir disponibilidad gratuita (von Kleist et al., 2025).

4.4. Implicaciones para organizaciones con capacidades limitadas

Las pequeñas organizaciones pueden beneficiarse de una secuencia incremental. Un primer nivel no requiere el modelo más complejo: necesita definiciones de eventos, control de duplicados, monitoreo de conversión y segmentación reproducible. Un segundo nivel añade probabilidades calibradas y umbrales vinculados a capacidad. Un tercer nivel prueba intervenciones mediante aleatorización o diseños cuasiexperimentales.

Esta progresión es importante porque la adopción de analítica en pymes enfrenta barreras de calidad, integración y competencias (Tawil et al., 2024). La asociación observada entre herramientas de analítica y desempeño no basta para elegir una arquitectura ni estimar retorno local (Almatrafi & Alharbi, 2023). Asimismo, la evidencia sobre marketing basado en datos sitúa el conocimiento organizacional como mecanismo entre infraestructura y desempeño (Gupta et al., 2021). Un modelo sin procesos de aprendizaje puede convertirse en una capa tecnológica aislada.

La regresión logística merece permanecer como referencia. En P4 fue inferior a árboles, pero entregó probabilidades bien calibradas internamente y desempeño sustancial. La diferencia de complejidad debe justificarse por valor adicional, mantenimiento, explicabilidad y estabilidad. El propósito no es defender un algoritmo único, sino evitar que una ganancia marginal o una característica tardía determine la solución sin evaluación de uso.

4.5. Propuesta de validación prospectiva

El siguiente estudio debería capturar eventos con identificadores seudonimizados y tiempo. Antes de entrenar, se definirían cortes de decisión, por ejemplo, tras la primera página, después de tres interacciones y antes del checkout. Cada predictor tendría un contrato de disponibilidad. La partición de evaluación sería temporal: entrenar en periodos anteriores y evaluar en periodos posteriores.

Después de validar discriminación y calibración, se seleccionarían umbrales con costos explícitos. Un experimento controlado asignaría de manera aleatoria una acción elegible —asistencia, recomendación o incentivo— frente a control, estratificando por riesgo. Los resultados incluirían conversión incremental, margen, costo por compra adicional, cancelación, devolución y señales de experiencia. Solo esa fase permitiría afirmar impacto causal.

El modelo tendría monitoreo de prevalencia, distribución de predictores, calibración y carga de alertas. Los segmentos con bajo volumen no deberían evaluarse mediante una cifra aislada. La documentación seguiría principios de transparencia como los sintetizados por Collins et al. (2024), adaptados al contexto empresarial.

5. Conclusiones

Los registros de sesión permiten construir modelos que discriminan conversiones mejor que el azar, pero la magnitud del desempeño depende críticamente del panel de variables. Con contexto, navegación y salida, los

resultados fueron moderados. Al incorporar PageValues, la discriminación, precisión promedio y error probabilístico mejoraron de forma desproporcionada en regresión logística y árboles potenciados.

Esa mejora no debe confundirse con preparación para uso. El conjunto no documenta cuándo se calculó PageValues, carece de secuencia temporal e impide comprobar que la señal estuviera disponible antes de una decisión. La conclusión responsable es doble: PageValues es altamente predictiva en el archivo histórico y, precisamente por ello, requiere auditoría de temporalidad antes de cualquier implementación.

La calibración interna y los escenarios de umbral muestran cómo traducir probabilidades en carga de trabajo, pero no estiman retorno ni efecto causal. Una decisión de conversión necesita datos con tiempo, validación temporal y externa, costos, capacidad y una prueba de intervención. La pregunta empresarial correcta no es solo "¿qué modelo predice mejor?", sino "¿qué información existe ahora, qué acción es posible y qué valor incremental produce?".

6. Limitaciones

Primero, el conjunto corresponde a una empresa anónima y no identifica país, sector, tamaño, año exacto, campañas ni condiciones de mercado. No representa automáticamente a comercios ecuatorianos ni a pymes. Segundo, cada fila resume una sesión terminada; no hay secuencia de eventos ni timestamps para reconstruir predicciones tempranas. Tercero, PageValues puede incorporar información histórica o contemporánea muy próxima a Revenue; su disponibilidad real es desconocida.

Cuarto, no existe identificador de sesión o usuario. Los 125 vectores repetidos pueden ser duplicados técnicos o sesiones legítimas; la sensibilidad por vectores únicos no resuelve su origen. Quinto, toda validación fue interna. La repetición reduce variabilidad de partición, pero no evalúa cambio temporal, transporte entre organizaciones ni impacto de una interfaz distinta.

Sexto, los subgrupos se limitaron a variables disponibles y algunos meses tuvieron pocos eventos. No se evaluaron atributos demográficos, equidad, privacidad de un sistema en producción ni deriva. Séptimo, el beneficio neto no incorporó valores monetarios, costos de incentivos, margen, capacidad, fatiga ni canibalización. Octavo, el estudio compara dos familias de modelos y no prueba todas las alternativas. Noveno, no se evaluó una intervención; ninguna diferencia puede interpretarse como aumento causal de conversión.

Agradecimientos

Se reconoce a los responsables del conjunto Online Shoppers Purchasing Intention y a UCI Machine Learning Repository por su disponibilidad pública.

Declaraciones

Campo / Field	Contenido / Content
Contribuciones CRediT	Conceptualización; metodología; software; validación; análisis formal; investigación; curación de datos; visualización; redacción del borrador original; redacción, revisión y edición; administración del proyecto.
Financiamiento	La investigación no recibió financiación externa específica.
Conflictos de interés	La persona autora declara no tener conflictos de interés.

Campo / Field	Contenido / Content
Aprobación ética	Análisis secundario de un conjunto público de sesiones anónimas, sin reclutamiento ni recopilación de datos nuevos. No fue aplicable una nueva revisión ética de investigación con seres humanos.
Consentimiento	No aplicable; el estudio no reclutó participantes ni recopiló datos nuevos.
Disponibilidad de datos	UCI: https://doi.org/10.24432/C5F88Q. Código, resultados y cuaderno analítico en el expediente.
Uso de inteligencia artificial	Codex de OpenAI apoyó la programación reproducible, la organización del manuscrito, la traducción académica y los controles editoriales. La persona autora revisó el contenido y asume la responsabilidad íntegra por la exactitud, originalidad, interpretación y versión final del artículo.

Referencias

Abdullah-All-Tanvir, Khandokar, I. A., Islam, A. K. M. M., Islam, S., & Shatabda, S. (2023). A gradient boosting classifier for purchase intention prediction of online shoppers. *Heliyon, 9*(4), e15163. https://doi.org/10.1016/j.heliyon.2023.e15163

Al Adwan, A., Kokash, H. A., Al Adwan, R., & Khattak, A. (2023). Data analytics in digital marketing for tracking the effectiveness of campaigns and informing strategy. *International Journal of Data and Network Science, 7*. https://doi.org/10.5267/j.ijdns.2023.3.015

Almatrafi, A. M., & Alharbi, Z. H. (2023). The impact of web analytics tools on the performance of small and medium enterprises. *Engineering, Technology & Applied Science Research, 13*(6). https://doi.org/10.48084/etasr.6261

Collins, G. S., Moons, K. G. M., Dhiman, P., et al. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078378. https://doi.org/10.1136/bmj-2023-078378

Gangadharan, K., Purandaran, A., Malathi, K., Subramanian, B., Jeyaraj, R., & Jung, S. K. (2025). From data to decisions: The power of machine learning in business recommendations. *IEEE Access, 13*. https://doi.org/10.1109/ACCESS.2025.3532697

Gudigantala, N., Madhavaram, S., & Bicen, P. (2023). An AI decision-making framework for business value maximization. *AI Magazine, 44*(1), 67–84. https://doi.org/10.1002/aaai.12076

Gupta, S., Justy, T., Kamboj, S., Kumar, A., & Kristoffersen, E. (2021). Big data and firm marketing performance: Findings from knowledge-based view. *Technological Forecasting and Social Change, 165*, 120986. https://doi.org/10.1016/j.techfore.2021.120986

Gupta, S., Leszkiewicz, A., Kumar, V., Bijmolt, T. H. A., & Potapov, D. (2020). Digital analytics: Modeling for insights and new methods. *Journal of Interactive Marketing, 51*, 26–43. https://doi.org/10.1016/j.intmar.2020.04.003

Kapoor, S., & Narayanan, A. (2023). Leakage and the reproducibility crisis in machine-learning-based science. *Patterns, 4*(9), 100804. https://doi.org/10.1016/j.patter.2023.100804

Kumar, B., Roy, S., Sinha, A., Iwendi, C., & Strážovská, Ľ. (2023). E-commerce website usability analysis using association rule mining and machine learning algorithm. *Mathematics, 11*(1), 25. https://doi.org/10.3390/math11010025

Liu, Y., Tian, Y., Xu, Y., Zhao, S., Huang, Y., Fan, Y., Duan, F., & Guo, P. (2021). TPGN: A Time-Preference Gate Network for e-commerce purchase intention recognition. *Knowledge-Based Systems, 220*, 106920. https://doi.org/10.1016/j.knosys.2021.106920

Saito, T., & Rehmsmeier, M. (2015). The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE, 10*(3), e0118432. https://doi.org/10.1371/journal.pone.0118432

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*, 6893–6908. https://doi.org/10.1007/s00521-018-3523-0

Tawil, A.-R. H., Mohamed, M., Schmoor, X., Vlachos, K., & Haidar, D. (2024). Trends and challenges towards effective data-driven decision making in UK small and medium-sized enterprises. *Big Data and Cognitive Computing, 8*(7), 79. https://doi.org/10.3390/bdcc8070079

Trivedi, S. K., Patra, P., Srivastava, P. R., Zhang, J. Z., & Zheng, L. J. (2024). What prompts consumers to purchase online? A machine learning approach. *Electronic Commerce Research, 24*(4), 2953–2989. https://doi.org/10.1007/s10660-022-09624-x

UCI Machine Learning Repository. (2018). *Online Shoppers Purchasing Intention Dataset*. https://doi.org/10.24432/C5F88Q

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. *BMC Medicine, 17*, 230. https://doi.org/10.1186/s12916-019-1466-7

Vickers, A. J., & Elkin, E. B. (2006). Decision curve analysis: A novel method for evaluating prediction models. *Medical Decision Making, 26*(6), 565–574. https://doi.org/10.1177/0272989X06295361

von Kleist, H., Zamanian, A., Shpitser, I., & Ahmidi, N. (2025). Evaluation of active feature acquisition methods for time-varying feature settings. *Journal of Machine Learning Research, 26*(60), 1–84. https://jmlr.org/papers/v26/23-1635.html

Wang, T., Li, N., Wang, H., Xian, J., & Guo, J. (2022). Visual analysis of e-commerce user behavior based on log mining. *Computational Intelligence and Neuroscience, 2022*, 4291978. https://doi.org/10.1155/2022/4291978

Anexos

No se incluyen anexos en el manuscrito editorial.

Información editorial

Campo / Field	Contenido / Content
Recepción / Received	10 de agosto de 2025 / August 10, 2025
Revisión / Revised	5 de septiembre de 2025 / September 5, 2025
Aceptación / Accepted	28 de septiembre de 2025 / September 28, 2025
Publicación / Published	30 de diciembre de 2025 / December 30, 2025
Volumen, número y año / Volume, issue and year	Vol. 1, Núm. 2, julio-diciembre de 2025