Modelo predictivo para determinar el nivel de venta en transacciones E-commerce

Jenny Marcela Amado Cortes
Diana Carolina Mendieta Rojas

22 de mayo 2025

Resumen

Este informe construye un modelo que permita predecir si una transacción de un cliente corresponde a una venta de monto alto (ventas = 1) o o venta de monto bajo (ventas = 0) en un entorno de comercio electrónico.

Índice

Índice de figuras

Índice de cuadros

1. Introducción

El E-commerce o comercio electrónico a presentado un crecimiento continuo en la economía, teniendo como hito importante la llegada de la pandemia o COVID 19 en el año 2020, a partir del confinamiento obligatorio los consumidores tuvieron la necesidad de probar nuevos canales comerciales, el comercio electrónico permitió al consumidor comprar bienes y servicios a través de medios electrónicos, páginas web online o por medio de plataformas como redes sociales. A partir de allí los hábitos de compra, consumo y venta se transformaron radicalmente obligando a las empresas a adaptarse para seguir siendo competitivas en el mercado.

El E-commerce se convirtió en una estrategia clave para las empresas y consumidores debido a que ofrece ventajas como procesos de compra agiles, múltiples modelos de negocio, reducción de costos en las operaciones, variedad de herramientas de publicidad y marketing, flexibilidad en la prestación del servicio, comparación de precios, ofrece diversas de formas de pago y se adapta a las necesidades del cliente. El comercio electrónico no es solo un catálogo sino un servicio completo en el cual se ofrece una experiencia digital en donde se encuentra una plataforma de venta, una pasarela de pagos, un sistema de gestión de inventarios y un servicio de logística y envíos.

Sin embargo, este es un modelo de negocio muy competitivo que requiere elementos diferenciadores que permitan captar la atención del cliente y que logren moldearse a sus necesidades. Mejorar la experiencia digital del usuario o cliente requiere potenciar aspectos como el marketing digital como factor clave, la publicidad en línea y las estrategias de captación son esenciales para el crecimiento del mercado e-commerce.

Para la sostenibilidad el mercado E-commerce es necesario conocer las tendencias o patrones de comportamiento de distintos elementos que podrían afectar la dirección de este mercado, factores como comportamiento de navegación del cliente, el análisis de consumo, la cultura, los avances tecnológicos entre otros, serán herramientas fundamentales para predecir y establecer estrategias dirigidas hacia el comportamiento futuro del mercado.

El uso de modelos predictivos ofrece beneficios que pueden optimizar la gestión, planificación y optimización de inventarios, ayuda a identificar cambios en la demanda, permite identificar tendencias de productos a largo plazo, permite conocer el comportamiento a futuro de los clientes ayudan a anticiparse y generar estrategias planificadas que permitan tomar decisiones

acertadas y eficaces que generen fidelización en el cliente; Conocer los montos de transacción de las ventas E-commerce permite prever cambios que lleven a anticiparse y generar campañas publicitarias y de marketing direccionadas a las necesidades específicas de los clientes y usar esto como ventaja competitiva ante un mercado cambiante.

Para el modelo predictivo se tomó una base de datos mediante "Kaggle" la cual contiene información del comportamiento de los consumidore de comercio electrónico, esta incluye información sobre patrones de compra, datos demográficos, precios de compra, entre otros, la base de datos es publica y contiene 5000 datos con 9 variables.

2. Desarrollo

2.1. Construcción de la Variable Objetivo

La variable objetivo `ventas` se definió como una clasificación binaria que representa si una transacción corresponde a una venta de alto valor (1) o bajo valor (0). Se construyó un *score* compuesto a partir de las siguientes variables relevantes:

- `unit_price`: precio unitario del producto
- `quantity`: cantidad de productos comprados
- `website_visits`: visitas al sitio web
- `app_usage_score`: nivel de compromiso digital del cliente (0=Low, 1=Medium, 2=High)
- `loyalty_program`: 50 puntos adicionales si el cliente pertenece al programa de lealtad
- `discount_applied`: 30 puntos adicionales si se utilizó un descuento

La clasificación final se determinó asignando ventas = 1 a las transacciones cuyo *score* estuviera por encima del percentil 75 del total. Esto permite segmentar de forma objetiva las ventas consideradas altas.

2.2. Visualización Exploratoria

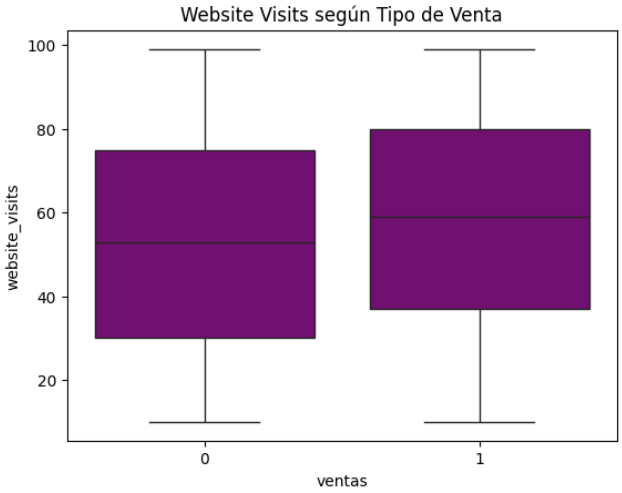

Figura 1: Distribución de visitas al sitio web según tipo de venta.

La Figura 1 muestra un diagrama de caja que compara el número de visitas al sitio web entre los dos tipos de transacciones: ventas de monto bajo (ventas = 0) y ventas de monto alto (ventas = 1). Se puede observar que, en promedio, los clientes que realizaron compras de mayor valor tendieron a visitar con mayor frecuencia el sitio web. Esto se evidencia en la mediana más alta y una mayor concentración de valores altos en el grupo de ventas altas.

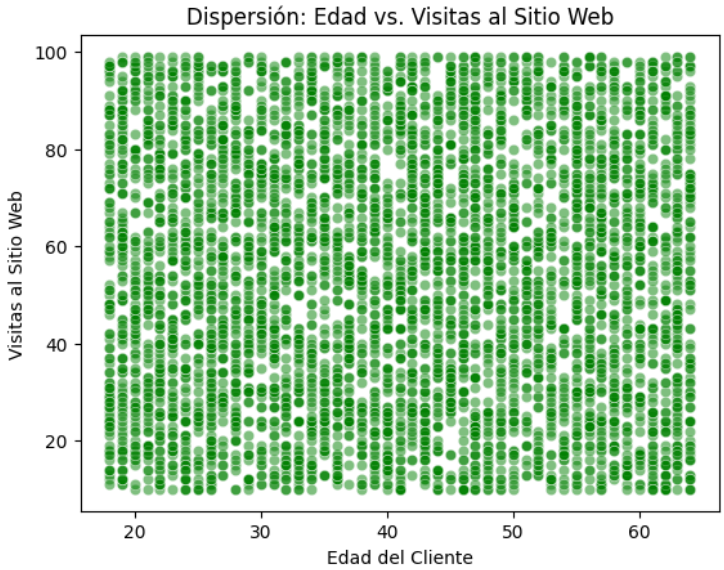

Figura 2: Dispersión: Edad vs. Visitas al Sitio Web

La Figura 2 presenta un gráfico de dispersión que muestra la relación entre la edad del cliente y la cantidad de visitas al sitio web. A simple vista, se evidencia una distribución uniforme de puntos en todas las edades, lo que indica que no existe una correlación significativa entre la edad del cliente y la frecuencia de visitas al sitio.

2.3. Evaluación Inicial del Modelo

Antes de aplicar transformaciones al dataset, se evaluó un modelo base de regresión logística utilizando solo las variables `unit_price`, `quantity`, `website_visits`, `app_usage_score` y `loyalty_program`. El modelo obtuvo una exactitud del 96.67 %, lo que indicó una buena capacidad predictiva incluso sin preprocesamiento avanzado.

2.4. Limpieza y Preparación del Dataset

Se imputaron los valores faltantes con la mediana y se escalaron las variables numéricas utilizando `MinMaxScaler`. Además, se codificó la variable categórica `gender` a formato binario. Esto preparó el dataset para una evaluación más rigurosa de los modelos.

2.5. Evaluación de Modelos Post-Limpieza

Se entrenaron y compararon tres modelos: regresión logística, KNN y árbol de decisión.

- **Regresión Logística:** Exactitud del 96.2 %. Excelente equilibrio entre precision y recall (0.94 y 0.91 respectivamente para clase 1). AUC = 1.00, lo que indica una distinción perfecta entre clases.
- **KNN:** Exactitud del 85.6 %. Menor recall para ventas altas (0.68) y menor consistencia entre pliegues.
- **Árbol de Decisión:** Exactitud del 95 %, con recall de 91 % para ventas altas. Las variables `unit_price` y `quantity` resultaron ser los predictores más fuertes.

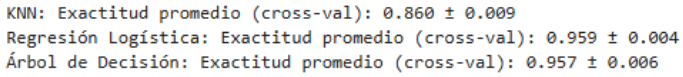

Figura 3: Resultado de validación cruzada

Figura 3 La validación cruzada aplicada a los tres modelos considerados —*Regresión Logística*, *Árbol de Decisión* y *K-Nearest Neighbors (KNN)*— permite comparar su rendimiento en términos de exactitud promedio y variabilidad.

El modelo de **Regresión Logística** obtuvo la mayor exactitud promedio (**0.959**) y la menor desviación estándar (± **0.004**), lo que demuestra un excelente desempeño y gran estabilidad ante diferentes particiones de los datos.

El **Árbol de Decisión** mostró un rendimiento muy cercano (**0.957** ± **0.006**), lo que lo convierte en una opción igualmente válida, especialmente por su capacidad interpretativa.

Por su parte, el modelo **KNN** presentó la exactitud más baja (**0.860**) y la mayor variabilidad (± **0.009**), evidenciando menor consistencia y rendimiento, posiblemente debido a la alta dimensionalidad o la necesidad de mayor ajuste de hiperparámetros.

En conjunto, se concluye que la Regresión Logística es el modelo más robusto, preciso y confiable para abordar el problema de predicción de ventas altas en comercio electrónico.

2.6. Comparación con y sin Poda en el *Decision Tree*

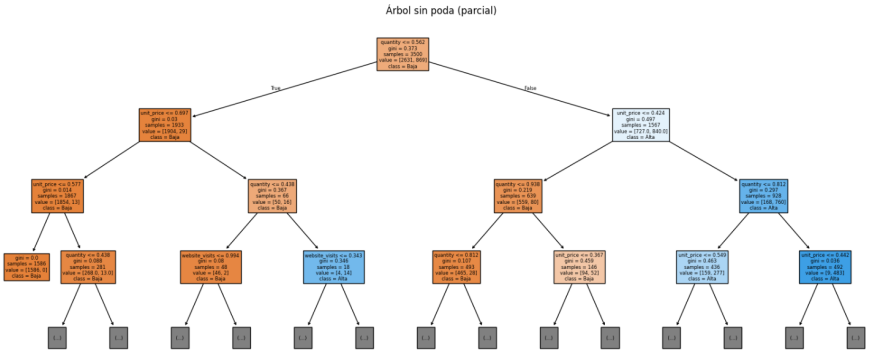

Figura 4: Árbol de Decisión Sin Poda

Como se observa en la Figura 4, las variables clave para la predicción son `unit_price`, que aparece como nodo raíz del árbol de decisión, seguido por `quantity`, ubicada en las ramas del primer nivel y que demuestra ser altamente predictiva. Finalmente, la variable `website_visits` aparece en niveles más profundos del árbol, lo que sugiere que cumple un rol de refinamiento en la clasificación. Esta estructura jerárquica evidencia la importancia relativa de cada variable en el proceso de toma de decisiones del modelo.

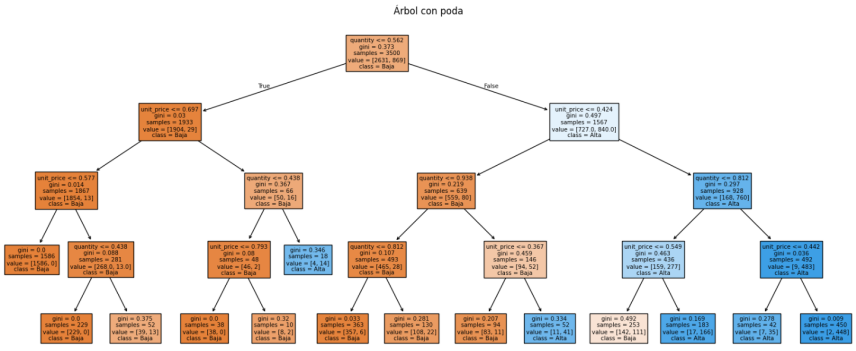

Figura 5: Árbol de Decisión Con Poda

El árbol de decisión podado, mostrado en la Figura 5, representa una versión simplificada del modelo completo. Si bien se observa una ligera disminución en la capacidad predictiva, especialmente en la clase de ventas altas

(recall de 77 % frente al 91 % del árbol sin poda), esta versión reduce el riesgo de sobreajuste y mejora la interpretabilidad del modelo.

- **Sin Poda:** Accuracy = 95 %, recall clase 1 = 91 %, F1-score clase 1 = 0.90
- **Con Poda:** Accuracy = 93 %, recall clase 1 = 77 %, F1-score clase 1 = 0.84

Esto evidencia que el **modelo sin poda es más eficaz para detectar ventas altas**, aunque potencialmente más complejo. En contraste, el modelo podado es más simple y generalizable pero con menor sensibilidad.

2.7. Conclusión Comparativa

Los resultados muestran que la regresión logística fue el modelo con mejor desempeño global, alcanzando una exactitud superior al 96Por otro lado el árbol de decisión sin poda ofreció una interpretación clara y precisión comparable, mientras que su versión podada mostró mayor simplicidad, aunque con una leve pérdida de sensibilidad para detectar ventas altas. Por su parte, el modelo KNN, si bien funcional, presentó un rendimiento inferior y mayor variabilidad.

En conjunto, se concluye que el enfoque basado en regresión logística es el más adecuado para este problema específico, logrando un equilibrio óptimo entre desempeño, estabilidad y eficiencia computacional. Además, la matriz de correlación y las visualizaciones exploratorias evidenciaron que variables como el precio unitario, la cantidad de productos, las visitas al sitio y el programa de lealtad son altamente relevantes para predecir el comportamiento del cliente. Estos hallazgos pueden ser utilizados por áreas comerciales y de marketing para diseñar estrategias más efectivas de fidelización y segmentación de clientes.