

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 a venta de monto bajo (ventas = 0) en un entorno de comercio electrónico.

Índice

1	Introducción	3
2	Desarrollo	4
2.1	Construcción de la Variable Objetivo	4
2.2	Revisar estructura del dataset	5
2.3	Visualización Exploratoria	6
2.4	Evaluación Inicial del Modelo	8
2.5	Limpieza y Preparación del Dataset	8
2.6	Evaluación de Modelos Post-Limpieza	8
2.6.1	Regresión Logística	9
2.6.2	KNN	12

2.6.3	Árbol de Decisión	12
2.7	Validación cruzada	13
2.8	Comparación con y sin Poda en el <i>Decision Tree</i>	14
2.9	Conclusión	15
A	Anexos	16
A.1	Matriz de Confusión - Árbol de Decisión	16
A.2	Curva ROC- Árbol de Decisión	18

Índice de figuras

1	Estructura del dataset	5
2	Matriz de correlación de ventas	6
3	Distribución de visitas al sitio web según tipo de venta.	7
4	Dispersión: Edad vs. Visitas al Sitio Web	7
5	Modelo inicial	8
6	Modelo entrenado	9
7	Curva ROC	10
8	Matriz de Confusión - Regresión Logística	11
9	Modelo KNN	12
10	Árbol de Decisión	13
11	Resultado de validación cruzada	13
12	Árbol de Decisión Sin Poda	14
13	Árbol de Decisión Con Poda	14
14	Matriz de Confusión - Árbol de Decisión sin poda	16
15	Matriz de Confusión - Árbol de Decisión podado	17
16	Curva ROC - Árbol de Decisión sin poda	18
17	Curva ROC - Árbol de Decisión podado	19

1. Introducción

El E-commerce o comercio electrónico ha 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 ágiles, 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 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 consumidores de comercio electrónico, esta incluye información sobre patrones de compra, datos demográficos, precios de compra, entre otros, la base de datos es pública y contiene 5000 datos con 10 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 temp* 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 temp* estuviera por encima del percentil 75 del total. Esto permite segmentar de forma objetiva las ventas consideradas altas.

2.2. Revisar estructura del dataset

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5000 entries, 0 to 4999
Data columns (total 10 columns):
#   Column                Non-Null Count  Dtype
---  -
0   unit_price            5000 non-null   float64
1   quantity              5000 non-null   int64
2   website_visits        5000 non-null   int64
3   membership_years      5000 non-null   int64
4   app_usage_score       5000 non-null   int64
5   discount_applied      5000 non-null   int64
6   age                   5000 non-null   int64
7   loyalty_program       5000 non-null   int64
8   gender                 5000 non-null   int64
9   ventas                5000 non-null   int64
dtypes: float64(1), int64(9)
memory usage: 390.8 KB
```

Figura 1: Estructura del dataset

Tenemos 10 variables donde:

- 1 de esas variables son de tipo `float64`, lo que significa que contienen números decimales.
- Las otras 9 variables son de tipo `int64`, lo que indica que contienen números enteros.

2.3. Visualización Exploratoria

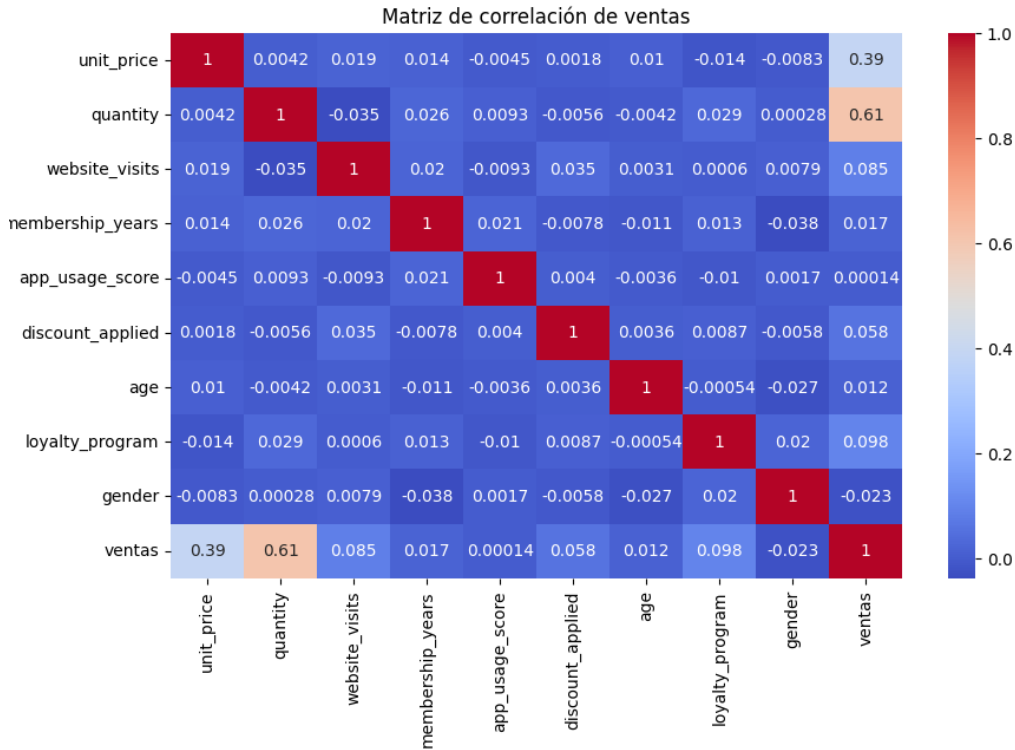


Figura 2: Matriz de correlación de ventas

El análisis de correlación 2 revela que la variable **quantity** presenta una fuerte correlación positiva con el nivel de ventas, lo que sugiere que un mayor número de unidades vendidas incrementa la probabilidad de una transacción clasificada como venta alta. De manera similar, **unit_price** también muestra una correlación relevante con las ventas.

En contraste, la variable **app_usage_score** no presenta una correlación significativa, lo que indica que su influencia directa en el volumen de ventas es limitada.

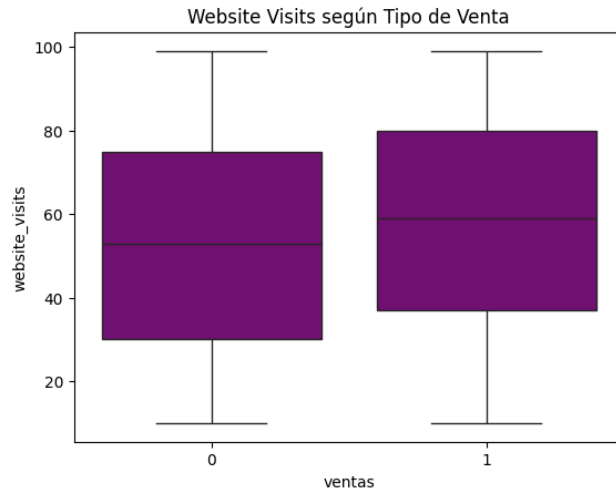


Figura 3: Distribución de visitas al sitio web según tipo de venta.

La Figura 3 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.

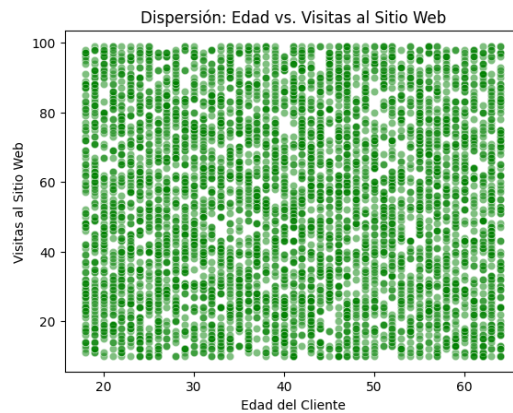


Figura 4: Dispersión: Edad vs. Visitas al Sitio Web

La Figura 4 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.

2.4. 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`.

```
# Evaluación
yb_pred = model_before.predict(Xb_test)
acc_before = accuracy_score(yb_test, yb_pred)
print("✅ Exactitud antes de limpieza:", round(acc_before, 4))
```

✅ Exactitud antes de limpieza: 0.9667

Figura 5: Modelo inicial

Este valor indica que el modelo de regresión logística logró una exactitud del 96.67% al predecir si una transacción corresponde a una venta alta (`ventas = 1`) o baja (`ventas = 0`), sin aplicar limpieza ni transformación previa sobre los datos.

2.5. 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` y `loyalty program` a formato binario. Esto preparó el dataset para una evaluación más rigurosa de los modelos.

2.6. Evaluación de Modelos Post-Limpieza

En esta parte se selecciona la variable (X) y la variable objetivo (y), se divide los datos en conjuntos de entrenamiento y prueba, entrena un modelo de regresión logística con los datos de entrenamiento, y luego evalúa el rendimiento del modelo en el conjunto de prueba.


```
# Evaluación
yc_pred = model_after.predict(Xc_test)
acc_after = accuracy_score(yc_test, yc_pred)
print("✅ Exactitud después de la limpieza:", round(acc_after, 4))
```

✅ Exactitud después de la limpieza: 0.962

Figura 6: Modelo entrenado

Por otro lado se entrenaron y compararon tres modelos: regresión logística, KNN y árbol de decisión.

2.6.1. Regresión Logística

■ Clase 0 (Ventas bajas)

La precisión del modelo para la clase de venta baja (0) es del 97 %, lo que indica que el 97 % de las predicciones de ventas bajas fueron correctas.

El recall es del 98 %, lo que significa que el modelo identificó correctamente el 98 % de las transacciones que realmente eran ventas bajas.

El F1-score, que balancea la precisión y el recall, es de 0.97, lo que refleja un buen rendimiento general para esta clase.

Además, el support de esta clase es de 1119, lo que indica que esta cantidad son las transacciones reales de venta baja en el conjunto de datos.

■ Clase 1 (Ventas altas)

La precisión del modelo para la clase de venta alta (1) indica que el 94 % de las predicciones de ventas altas fueron correctas. Mientras que El recall identificó correctamente el 91 % de las transacciones que realmente eran ventas altas. Asimismo, El F1-score, que balancea la precisión y el recall, es de 0.92, muestra un buen desempeño, aunque ligeramente inferior al de la clase 0.

Además, el support de esta clase es de 381 indicando el número de transacciones reales de venta alta en el conjunto de datos.

2.6.1.1 Curva ROC: La Figura 7 muestra que el modelo distingue perfectamente entre ventas altas (1) y ventas bajas (0), sin confusión entre clases. El área bajo la curva (AUC) es igual a 1, lo que indica un rendimiento excelente del modelo en la tarea de clasificación.

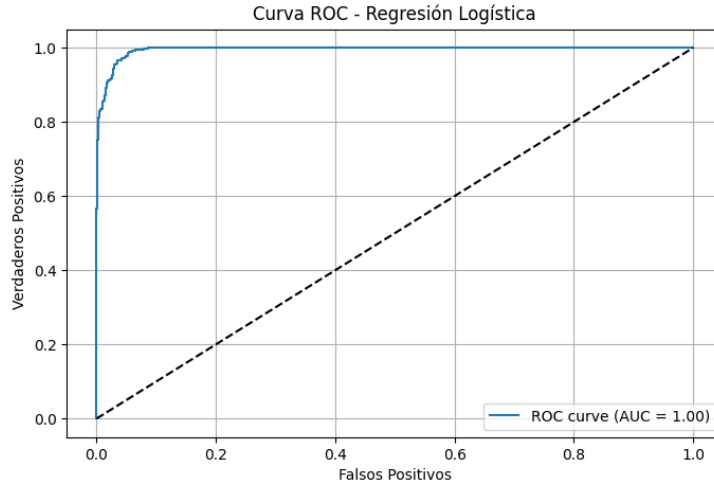


Figura 7: Curva ROC

2.6.1.2 Matriz de Confusión: La matriz de confusión indica que el modelo clasificó correctamente 1095 transacciones como ventas bajas y 348 como ventas altas. Sin embargo, cometió 24 errores al clasificar ventas bajas como altas (falsos positivos) y 33 al clasificar ventas altas como bajas (falsos negativos).

A partir de estos valores, se calculan las siguientes métricas:

- **Precisión para la clase 0 (venta baja):**

$$\text{Precisión}_0 = \frac{1095}{1095 + 24} \approx 0,9786$$

- **Recall para la clase 0 (venta baja):**

$$\text{Recall}_0 = \frac{1095}{1095 + 33} \approx 0,9707$$

- **Precisión para la clase 1 (venta alta):**

$$\text{Precisión}_1 = \frac{348}{348 + 33} \approx 0,9134$$

- Recall para la clase 1 (venta alta):

$$\text{Recall}_1 = \frac{348}{348 + 24} \approx 0,9355$$

- Exactitud general del modelo:

$$\text{Exactitud} = \frac{1095 + 348}{1500} \approx 0,96$$

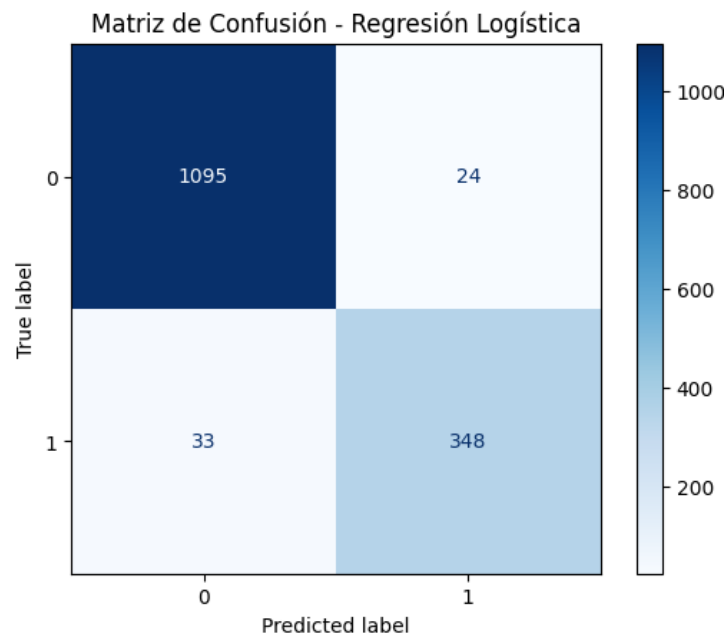
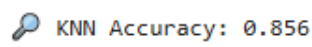


Figura 8: Matriz de Confusión - Regresión Logística

Estos resultados reflejan un rendimiento muy sólido del modelo, con un comportamiento ligeramente más favorable al clasificar ventas bajas.

2.6.2. KNN



	precision	recall	f1-score	support
0	0.89	0.92	0.90	1119
1	0.73	0.68	0.71	381
accuracy			0.86	1500
macro avg	0.81	0.80	0.81	1500
weighted avg	0.85	0.86	0.85	1500

Figura 9: Modelo KNN

■ Clase 0 – Venta baja:

- **Precisión:** Cuando el modelo predice una venta baja, acierta el 89 % de las veces.
- **Recall:** De todas las ventas bajas reales, el 92 % fueron correctamente identificadas.

■ Clase 1 – Venta alta:

- **Precisión:** Cuando el modelo predice una venta alta, acierta el 73 %.
- **Recall:** Solo se logra identificar el 68 % de las ventas altas reales.

2.6.3. Árbol de Decisión

El modelo de Árbol de Decisión logra clasificar correctamente el 95 % de las transacciones, lo que refleja un desempeño sólido. Presenta un buen equilibrio entre precisión y exhaustividad, según lo evidenciado por el F1-score.

En cuanto a la interpretación del árbol, se observa que la variable `unit_price` ocupa la raíz, lo que indica su alto poder predictivo. La variable `quantity` aparece en las ramas del primer nivel, también con gran importancia, mientras que `website_visits` se encuentra en niveles más profundos, contribuyendo a refinar la clasificación final.

🌳 Árbol de Decisión Accuracy: 0.95				
	precision	recall	f1-score	support
0	0.97	0.96	0.97	1119
1	0.89	0.91	0.90	381
accuracy			0.95	1500
macro avg	0.93	0.94	0.93	1500
weighted avg	0.95	0.95	0.95	1500

Figura 10: Árbol de Decisión

2.7. Validación cruzada

```

KNN: Exactitud promedio (cross-val): 0.860 ± 0.009
Regresión Logística: Exactitud promedio (cross-val): 0.959 ± 0.004
Árbol de Decisión: Exactitud promedio (cross-val): 0.957 ± 0.006

```

Figura 11: Resultado de validación cruzada

Figura 11 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 (\pm **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** \pm **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 (\pm **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.

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

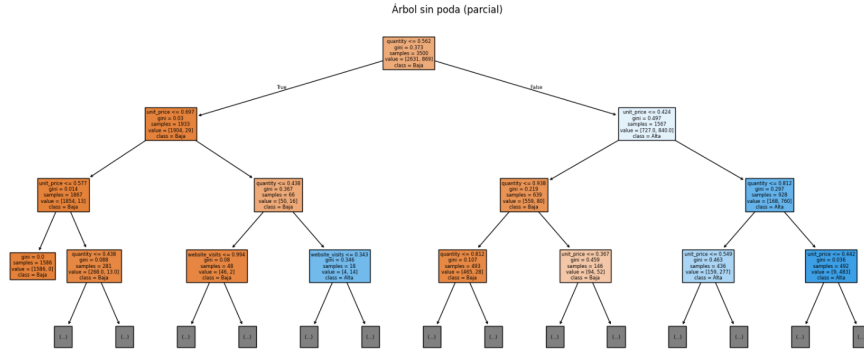


Figura 12: Árbol de Decisión Sin Poda

Como se observa en la Figura 12, 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.

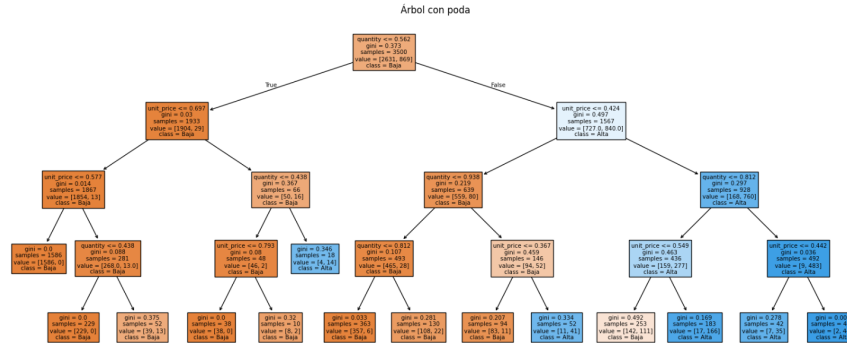


Figura 13: Árbol de Decisión Con Poda

El árbol de decisión podado, mostrado en la Figura 13, 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.

2.9. Conclusión

Los resultados muestran que la regresión logística fue el modelo con mejor desempeño global, alcanzando una exactitud superior al 96%. Por 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.

A. Anexos

A.1. Matriz de Confusión - Árbol de Decisión

A continuación se muestra la matriz de confusión para el modelo de Árbol de Decisión, tanto para la versión con poda como sin poda:

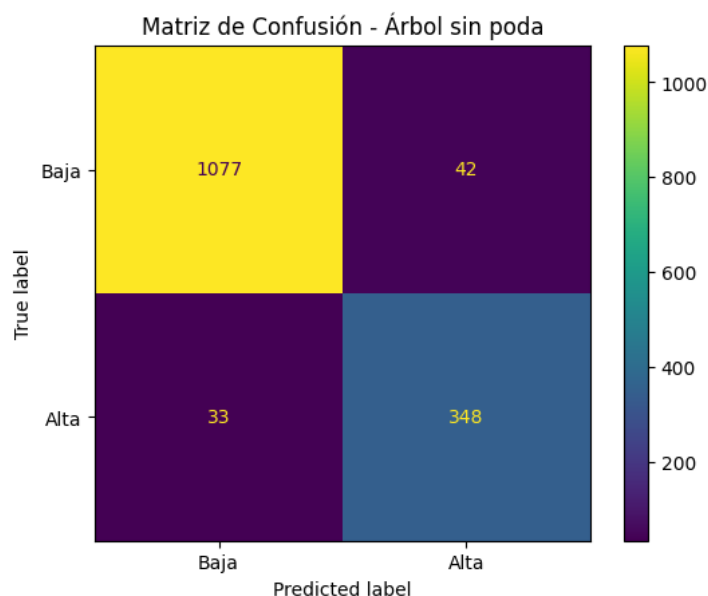


Figura 14: Matriz de Confusión - Árbol de Decisión sin poda

La matriz de confusión del árbol sin podar refleja una mejora en la detección de ventas altas. El modelo alcanza una exactitud general del 95,0 %, con alta precisión (89,22 %) y mejor sensibilidad (91,34 %) para ventas altas comparado con el modelo podado. Clasifica correctamente 348 ventas altas y solo se equivoca en 33, lo que indica una mayor capacidad del modelo para identificar correctamente las transacciones de alto valor. Esta mejora sugiere que el árbol sin podar, aunque más complejo, logra un mejor equilibrio entre precisión y sensibilidad, especialmente valioso en contextos donde detectar ventas altas es prioritario.

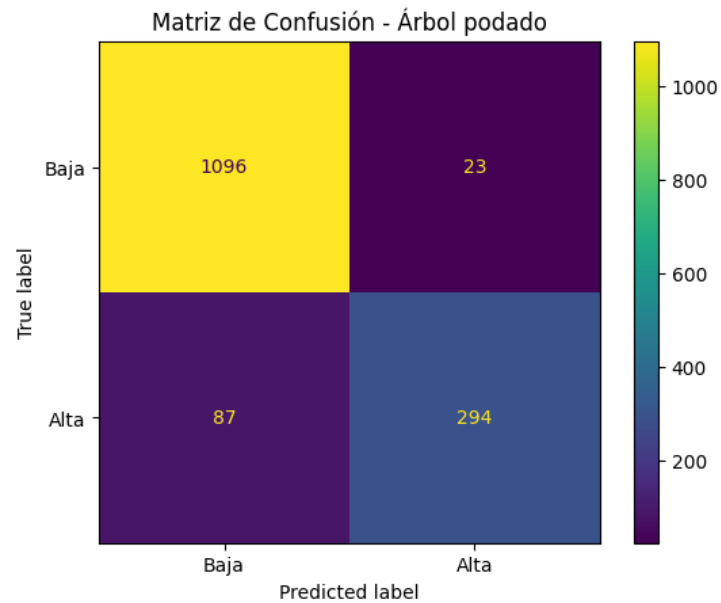


Figura 15: Matriz de Confusión - Árbol de Decisión podado

La matriz de confusión del árbol podado muestra una alta precisión general (92,67 %), con buen desempeño en ventas bajas, pero una sensibilidad moderada (77,17 %) para detectar ventas altas. El modelo tiende a ser conservador, lo que reduce falsos positivos, pero deja pasar algunas ventas altas. Para mejorar esta detección, se recomienda ajustar el umbral de clasificación o aplicar técnicas de balanceo de clases.

A.2. Curva ROC- Árbol de Decisión

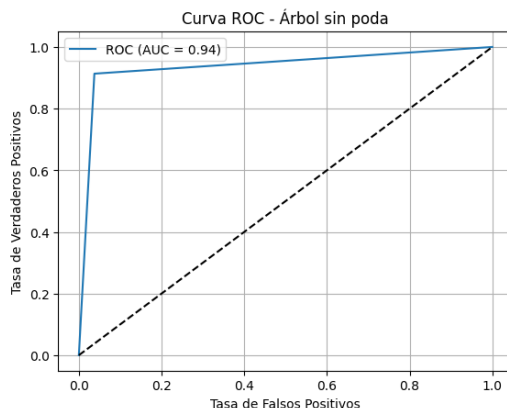


Figura 16: Curva ROC - Árbol de Decisión sin poda

Las curvas ROC comparan el desempeño de los modelos de árbol de decisión sin podar y podado en distintos umbrales de clasificación. El árbol sin podar alcanza un AUC de 0,94, lo que indica una buena capacidad de discriminación entre ventas altas y bajas, aunque con una curva más abrupta, lo que sugiere posible sobreajuste. Por otro lado, el árbol podado presenta un AUC superior de 0,97 y una curva más suave y progresiva, lo que refleja un mejor comportamiento promedio en todos los umbrales y mayor capacidad de generalización.

Aunque el modelo sin podar es más sensible al detectar ventas altas directamente, el modelo podado ofrece un mejor rendimiento global y mayor robustez frente a variaciones en el umbral de decisión.

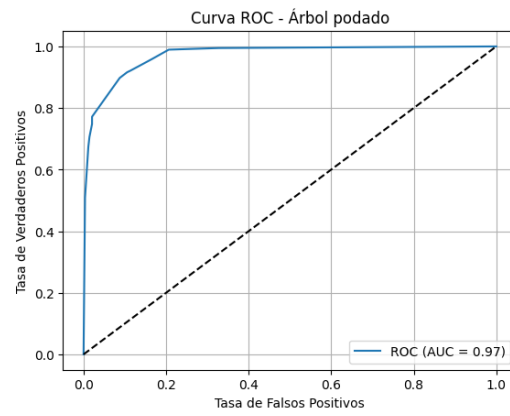


Figura 17: Curva ROC - Árbol de Decisión podado