UNIVERSIDAD TECNOLÓGICA DE PANAMÁ

FACULTAD DE INGENIERÍA DE SISTEMAS COMPUTACIONALES

MAESTRÍA EN ANALÍTICA DE DATOS

ASIGNATURA:

MODELOS PREDICTIVOS

REPORTE DE PROYECTO FINAL:

MODELO PREDICTIVO PARA LA PROYECCIÓN DE VENTAS: ANÁLISIS Y APLICACIÓN DE TÉCNICAS DE MACHINE LEARNING

ESTUDIANTE:

VALZANIA, MIGUEL 8-927-1669

GRUPO:

AN215

FACILITADOR:

Prof. JUAN MARCOS CASTILLO, PhD

FECHA DE ENTREGA:

08 DE ABRIL DEL 2025

ÍNDICE

2. INTRODUCCIÓN

El presente proyecto se enfoca en el desarrollo e implementación de un modelo predictivo para la proyección de ventas, utilizando técnicas de machine learning aplicadas a un conjunto de datos comerciales. Durante mi experiencia profesional en el área de ingeniería de datos, he observado la importancia que tiene para las organizaciones la capacidad de anticipar con precisión sus volúmenes de venta futuros, lo que permite una planificación estratégica más eficiente de recursos e inventarios.

Este trabajo surge como respuesta a la necesidad de desarrollar herramientas cuantitativas robustas que permitan superar las limitaciones de los métodos tradicionales de proyección de ventas. Al cursar la asignatura de Modelos Predictivos, identifiqué la oportunidad de aplicar los conocimientos adquiridos para construir un modelo que pudiera capturar patrones complejos en datos históricos y transformarlos en proyecciones confiables.

Para este fin, seleccioné un conjunto de datos que comprende 5,000 transacciones comerciales, cada una caracterizada por 24 variables diferentes que abarcan aspectos como información del cliente, detalles del producto, datos financieros y logísticos. Este dataset proporciona una base sólida para el desarrollo de un modelo predictivo que pueda identificar patrones y tendencias relevantes.

3. JUSTIFICACIÓN

Este proyecto se fundamenta en la creciente necesidad que tienen las organizaciones de contar con proyecciones de ventas precisas y confiables para su planificación estratégica. Los métodos tradicionales de proyección, basados en promedios históricos o ajustes manuales, frecuentemente presentan limitaciones significativas frente a la complejidad del mercado actual.

Durante mi experiencia laboral en el departamento de analítica de datos, he podido constatar cómo las empresas que utilizan proyecciones tradicionales enfrentan dificultades para adaptarse a cambios rápidos en las condiciones del mercado. Estas proyecciones, aunque fundamentadas en metodologías establecidas, a menudo carecen de la flexibilidad necesaria para incorporar patrones no lineales o relaciones complejas entre múltiples variables.

La implementación de modelos predictivos basados en técnicas de machine learning representa una oportunidad significativa para:

- **Mejorar la precisión de las proyecciones:** Al incorporar algoritmos que pueden capturar relaciones complejas entre variables y patrones temporales que los métodos tradicionales no detectan.
- **Optimizar la planificación de recursos:** Permitiendo a las empresas ajustar inventarios, personal y estrategias de marketing con mayor eficiencia.
- **Reducir los costos operativos:** Minimizando tanto el exceso de inventario como las ventas perdidas por falta de stock.
- **Responder con agilidad a cambios en el mercado:** Mediante la actualización continua del modelo con nuevos datos para reflejar condiciones cambiantes.

Este proyecto busca desarrollar un modelo que aborde estas necesidades, utilizando un conjunto de datos representativo como base para la experimentación y validación. Desde una perspectiva académica, este trabajo permite aplicar los conocimientos adquiridos durante la maestría en un problema de relevancia práctica, mientras se desarrollan habilidades analíticas transferibles a diversos contextos profesionales.

4. ANTECEDENTES

La búsqueda de la rentabilidad no es algo nuevo. Desde los primeros intercambios comerciales hasta las estrategias empresariales más sofisticadas de hoy, siempre ha existido la necesidad de entender qué hace que una venta sea más exitosa que otra. Sin embargo, lo que ha cambiado radicalmente es la manera en que analizamos estos factores: hemos pasado de simples observaciones intuitivas a modelos predictivos basados en datos que pueden procesar miles de variables simultáneamente.

A lo largo de los años, diferentes estudios han arrojado luz sobre piezas clave de este rompecabezas. Kumar et al. (2018) identificaron que distintos tipos de clientes siguen patrones de compra únicos, casi como si hablaran diferentes "dialectos" comerciales. Sus hallazgos sugieren que la segmentación efectiva de clientes puede mejorar la precisión de las predicciones de ventas hasta en un 27%.

Chen y Gallego (2019) demostraron que las estrategias de descuento requieren un equilibrio delicado: si son demasiado agresivas, erosionan los márgenes; si son demasiado conservadoras, pueden frenar las ventas. Su estudio reveló que la elasticidad de precios varía significativamente según la categoría del producto y el tipo de cliente, lo que complica aún más la ecuación.

Por otro lado, Zhang y Lee (2020) encontraron que la logística juega un papel tan crucial en la rentabilidad como los precios, aunque muchas veces se subestima su impacto. Según su investigación, la elección del modo de envío puede afectar los márgenes en hasta un 15%, especialmente en productos de bajo costo pero alto volumen.

Más recientemente, Pearson (2021) mostró que la selección de productos va más allá de elegir qué vender: se trata de construir una oferta coherente, donde cada producto cumple un rol dentro de un ecosistema comercial más amplio. Sus estudios de caso demostraron cómo empresas que alinean su catálogo con patrones de compra específicos pueden incrementar el valor promedio del pedido en más de un 20%.

Estos estudios refuerzan la idea de que la rentabilidad de una venta no es el resultado de un solo factor, sino de la combinación de múltiples elementos que interactúan de maneras complejas. Con este proyecto, mi objetivo es conectar estos hallazgos teóricos con un enfoque práctico, usando datos reales para descubrir qué patrones pueden llevar a un negocio a tomar mejores decisiones y mejorar su rentabilidad de forma consistente.

5. DEFINICIÓN DEL PROBLEMA

El problema central que aborda este proyecto es el desarrollo de un modelo predictivo capaz de proyectar con precisión las ventas futuras basándose en datos históricos y variables explicativas. Específicamente, se busca responder a la pregunta: ***¿Cómo podemos utilizar técnicas de machine learning para predecir con exactitud los valores de ventas futuras considerando múltiples variables y sus interacciones?***

Este problema responde a necesidades concretas que he observado en entornos empresariales, donde las organizaciones requieren:

- Proyecciones de ventas más precisas para períodos futuros que consideren la estacionalidad y tendencias
- Capacidad para cuantificar el impacto de diferentes variables (como descuentos, categorías de productos o tipos de clientes) en los resultados de ventas
- Herramientas que permitan simular escenarios para la toma de decisiones informadas
- Metodologías replicables que reduzcan la subjetividad en los procesos de planificación

La complejidad de este problema radica en múltiples dimensiones que deben ser abordadas simultáneamente:

- **Dimensión temporal:** Las ventas presentan patrones estacionales, tendencias y ciclos que deben ser identificados y modelados adecuadamente.
- **Dimensión multivariable:** El comportamiento de las ventas está influenciado por numerosas variables tanto categóricas como numéricas, cuyas interacciones pueden ser no lineales y complejas.
- **Dimensión estructural:** La selección del algoritmo y arquitectura del modelo debe equilibrar la capacidad predictiva con la interpretabilidad y aplicabilidad práctica.

- **Dimensión metodológica:** El preprocesamiento de datos, selección de variables y validación del modelo requieren un enfoque riguroso para asegurar resultados confiables.

El objetivo específico es desarrollar un modelo que minimice el error de predicción (medido a través de métricas como RMSE, MAE y MAPE) mientras mantiene un nivel de complejidad manejable que permita su implementación práctica. Este modelo deberá ser capaz de generar proyecciones tanto a nivel de transacciones individuales como agregadas por periodos temporales.

6. ANÁLISIS PREDICTIVO

a. Determinación de la base de datos

Para este proyecto utilicé un dataset de ventas extraído de Kaggle, el cual contiene 5,000 registros y 24 variables que capturan distintas dimensiones del proceso comercial. Este conjunto de datos sintético representa información de ventas para un supermercado con datos que abarcan varios aspectos de las transacciones comerciales. Los datos incluyen:

- **Detalles de cada orden:** número identificador, fecha y prioridad del pedido.
- **Información del cliente:** nombre, dirección, ciudad, estado y tipo de cliente.
- **Productos:** nombre, categoría y tipo de empaque/contenedor.
- **Datos financieros:** precios de costo, precios de venta al público, márgenes de beneficio, cantidades y descuentos aplicados.
- **Información de envío:** método de envío, fechas y costos asociados.

Elegí este dataset por varias razones fundamentales:

- **Visión integral:** Proporciona una perspectiva completa del ciclo de ventas, desde el perfil del cliente hasta los resultados financieros.
- **Múltiples factores:** Permite analizar diversas variables simultáneamente y explorar cómo interactúan entre sí para influir en la rentabilidad.
- **Estructura adecuada:** Al ser un conjunto de datos bien organizado, facilita la implementación de modelos predictivos sin la complejidad de limpiar información desordenada.
- **Relevancia práctica:** Los datos son similares a los que encontraría en mi entorno profesional, lo que facilita la transferencia de aprendizajes.

La variable objetivo elegida para este análisis predictivo es el ***"Total"***, que representa el valor final de la transacción de venta, incluyendo el costo de los productos, menos los descuentos y más los gastos de envío. Esta variable refleja el monto final que paga el cliente y es un indicador clave del desempeño comercial.

b. Pre-procesamiento y limpieza de datos

Antes de comenzar con el análisis y la construcción de modelos, fue necesario realizar un proceso de limpieza y preparación de los datos para asegurar su calidad y coherencia. A continuación, detallo los pasos realizados:

- **Revisión inicial y conversión de tipos de datos**

Primero, inspeccioné la estructura del dataset para entender los tipos de datos de cada columna y su distribución general. Observé que teníamos 24 columnas con una mezcla de datos categóricos (como tipo de cliente y categoría de producto) y numéricos (como precios y cantidades).

Para unificar el manejo de los datos numéricos, convertí la columna 'Discount' de tipo 'int64' a 'float64', lo que facilitaría los cálculos posteriores al manejar todos los valores porcentuales en el mismo formato.

```
#Convierto la columna Discount de int64 a float64
df['Discount'] = df['Discount'].astype('float64')
```

También verifiqué que las fechas estuvieran en el formato correcto:

```
# Convertir fechas
df['Order Date'] = pd.to_datetime(df['Order Date'])
df['Ship Date'] = pd.to_datetime(df['Ship Date'])
```

- **Manejo de valores nulos**

Al examinar el dataset, encontré valores nulos en dos columnas: 'Address' (1 registro) y 'Order Quantity' (1 registro). Aunque estos representaban una proporción mínima del total de datos (apenas el 0.02%), era importante abordarlos adecuadamente.

Decidí eliminar estos registros, ya que representaban una fracción insignificante del dataset y no queríamos comprometer la calidad del análisis con imputaciones arbitrarias:

```
#Voy a eliminar ambos registros debido a que son minimos y no es significativo en el análisis

df.dropna(subset=['Order Quantity'], inplace=True)
df.dropna(subset=['Address'], inplace=True)
```

Esta decisión redujo nuestro dataset de 5,000 a 4,999 registros, lo cual no afecta significativamente la robustez de nuestro análisis.

- **Verificación de duplicados**

Comprobé la existencia de filas duplicadas en el dataset, pero no se encontró ninguna, lo que indica que cada registro representa una transacción única:

```
# Verificar duplicados
print(f"Número de filas duplicadas: {df.duplicated().sum()}")
print("\n")
```

```
Número de filas duplicadas: 0
```

- **Creación de variables temporales**

Para facilitar el análisis de tendencias y estacionalidad, creé nuevas variables basadas en las fechas de los pedidos:

```
# Creamos las columnas de periodo, mes
df['periodo'] = df['Order Date'].dt.year
df['mes'] = df['Order Date'].dt.month
df['periodo_mes'] = df['Order Date'].dt.strftime('%Y-%m')
```

Estas variables me permitirían agrupar y analizar las ventas por diferentes periodos temporales.

- **Eliminación de columnas no relevantes para el modelo**

Para el desarrollo del modelo predictivo, decidí eliminar algunas columnas que no aportan información significativa o que podían introducir ruido:

```
# Eliminamos las columnas que no se utilizarán para el análisis.
# Las mismas no son relevantes ni significativas para la ventas después de varios análisis realizados.

columnas_eliminar = ['Order No', 'City','State','Customer Name', 'Address']
df = df.drop(columns=columnas_eliminar)
```

Estas columnas, aunque útiles para identificar transacciones individuales, no representan patrones generalizables que ayuden a predecir la rentabilidad futura.

- **Extracción de características temporales**

Para capturar patrones estacionales en los datos, extraje información adicional de las fechas:

```
# Extraemos los dias y meses de pedido y envio lo cual nos servirá para la estacionalidad próxima.
df['Order day'] = df['Order Date'].dt.weekday
df['Ship day'] = df['Ship Date'].dt.weekday
df['Order Month'] = df['Order Date'].dt.month
df['Ship Month'] = df['Ship Date'].dt.month
df['Order Year'] = df['Order Date'].dt.year
df['Ship Year'] = df['Ship Date'].dt.year
```

- **Transformación de variables categóricas**

Finalmente, realicé la codificación de variables categóricas mediante one-hot encoding, transformando categorías como tipo de cliente, prioridad de orden y método de envío en variables binarias que pudieran ser procesadas por algoritmos de aprendizaje automático:

```
# Aplicamos One-hot encoding para las variables categoricas
df = pd.get_dummies(df, columns=['Ship day','Order day','Customer Type','Account Manager','Order Priority','Product Category','Product Container','Ship Mode',
```

- **Preparación final para el modelado**

Tras el procesamiento, eliminé las columnas de fechas originales y el nombre del producto, que contenía demasiadas categorías únicas para ser útil en el modelado:

```
# Eliminamos las siguiente columnas
df = df.drop(columns=['Order Date','Product Name','Ship Date'])
df.dropna(subset=list(df.columns), inplace=True)
```

Y finalmente, eliminé cualquier fila que pudiera contener valores nulos tras todas estas transformaciones. El resultado final fue un dataset limpio y estructurado con 4,999 filas y 76 columnas, listo para la fase de análisis descriptivo y modelado predictivo.

c. Análisis descriptivo

El análisis descriptivo de los datos reveló patrones interesantes que ayudaron a comprender mejor la naturaleza de las ventas y proporcionaron insights valiosos para la construcción posterior del modelo predictivo.

- **Evolución temporal de las ventas**

Al analizar las ventas anuales, observé una tendencia creciente entre 2013 y 2015, con una ligera caída en 2016, y un descenso significativo en 2017 (aunque esto último se debe a que los datos solo incluyen los primeros meses de ese año).

Ventas totales por año (USD)	
Periodo	Total
2013	766873.10
2014	992586.20
2015	995253.41
2016	890547.37
2017	76572.55

Un análisis más detallado de las ventas mensuales mostró patrones estacionales claros, con picos recurrentes en ciertos meses del año, particularmente en junio y julio, lo que sugiere que hay factores temporales significativos que afectan al comportamiento de compra.

- **Distribución por categorías de producto**

El análisis por categorías de producto reveló un desequilibrio importante en la distribución de las ventas:

Categorías de productos más vendidos (USD)	
Product Category	Total
Office Supplies	2917828.03
Technology	710496.56
Furniture	93508.04

Los artículos de oficina representan más del 78% del valor total de ventas, seguidos por los productos tecnológicos (19%) y el mobiliario (3%). Esta distribución desigual

sugiere que la empresa tiene una especialización clara en productos de oficina, o que estos productos tienen una rotación mucho mayor que las otras categorías.

- **Productos más vendidos**

Al examinar los 10 productos más vendidos, destacan particularmente dos copiadoras que juntas representan más de un millón de dólares en ventas:

Top 10 productos más vendidos (USD)	
Producto	Total
Cando PC940 Copier	695199.82
HFX LaserJet 3310 Copier	432414.11
Adesso Programmable 142-Key Keyboard	237575.05
UGen Ultra Professional Cordless Optical Suite	201337.72
Multimedia Mailers	117741.64
UGen Ultra Cordless Optical Suite	111767.25
Economy Rollaway Files	102319.74
Emerson Stylus 1520 Color Inkjet Printer	85456.96
TypeRight Side-Opening Peel & Seel Expanding	80846.92
600 Series Non-Flip	60530.63

Esta concentración de ventas en pocos productos de alto valor plantea la hipótesis de que quizás los productos tecnológicos, aunque menos frecuentes en número de transacciones, podrían tener un impacto desproporcionado en la rentabilidad total.

- **Distribución por prioridad de orden**

El análisis de la distribución de órdenes por nivel de prioridad mostró que la mayoría de las órdenes tienen prioridad media o crítica, seguidas por las de prioridad alta y baja. Esto podría indicar diferentes patrones de urgencia en las compras de los clientes.

- **Ventas por tipo de cliente**

El análisis por tipo de cliente reveló que los clientes corporativos y consumidores directos generan la mayor parte de los ingresos, seguidos por las oficinas domésticas y pequeñas empresas. Esta distribución sugiere que diferentes segmentos de clientes podrían requerir estrategias comerciales distintas.

- **Desempeño por gestor de cuenta**

Al examinar las ventas por gestor de cuenta, se identificaron diferencias significativas en el desempeño, con algunos gestores generando más del doble de ventas que otros. Esta variación podría deberse a diferentes factores, desde la cartera de clientes asignada hasta estrategias de venta individuales.

- **Correlaciones entre variables numéricas**

El análisis de correlación entre variables numéricas reveló relaciones interesantes:

- **Fuerte correlación positiva (0.99)** entre "Sub Total" y "Order Total", lo que era esperado ya que el subtotal es el componente principal del total antes de descuentos.
- **Correlación positiva significativa (0.60)** entre "Order Quantity" y "Sub Total", indicando que, como es lógico, pedidos más grandes tienden a generar más ingresos.
- **Correlación negativa (-0.39)** entre "Discount" y "Profit Margin", sugiriendo que mayores descuentos tienden a reducir los márgenes de beneficio.

Estas observaciones proporcionaron insights valiosos que luego informaron la selección de variables para el modelo predictivo.

d. **Selección de variables**

La selección adecuada de variables es crucial para desarrollar un modelo predictivo efectivo y evitar el sobreajuste. Basándome en el análisis descriptivo y en las correlaciones observadas, tomé decisiones fundamentadas sobre qué variables incluir en el modelo.

- **Variables numéricas clave**

Decidí incluir todas las variables numéricas principales debido a su relevancia teórica para predecir el valor total de una transacción:

- **Cost Price:** El precio de costo es fundamental ya que establece la base sobre la cual se calcula el margen.
- **Retail Price:** El precio de venta al público determina directamente el ingreso potencial.
- **Profit Margin:** El margen de beneficio captura la diferencia proporcional entre costo y precio de venta.
- **Order Quantity:** La cantidad ordenada multiplica el efecto de las demás variables.
- **Sub Total:** Representa el valor bruto antes de descuentos y costos adicionales.
- **Discount y Total Discount:** Capturan tanto el porcentaje como el valor absoluto de los descuentos aplicados.
- **Order Total:** El valor después de descuentos pero antes de costos de envío.
- **Shipping Cost:** Un componente directo del valor final.

- **Variables temporales**

Para capturar patrones estacionales y tendencias, extraje y transformé información temporal:

- **Order Month** y **Ship Month**: Para capturar variaciones estacionales dentro del año.

- **Order day** y **Ship day** (convertidas a variables dummy): Para identificar patrones semanales que pudieran afectar los costos o volúmenes.
- **Order Year** y **Ship Year** (también convertidas a variables dummy): Para capturar tendencias anuales y efectos de largo plazo.

- **Variables categóricas**

Transformé todas las variables categóricas mediante one-hot encoding para que pudieran ser interpretadas por el algoritmo:

- **Customer Type:** Diferentes tipos de clientes pueden tener comportamientos de compra distintos.
- **Account Manager:** El gestor de cuenta puede influir en los resultados de ventas.
- **Order Priority:** La prioridad puede afectar costos y márgenes.
- **Product Category:** Categorías distintas pueden tener estructuras de costos y precios diferentes.
- **Product Container:** El tipo de empaque puede influir en costos de manejo y envío.
- **Ship Mode:** El método de envío afecta directamente los costos e indirectamente la satisfacción del cliente.

- **Variables descartadas**

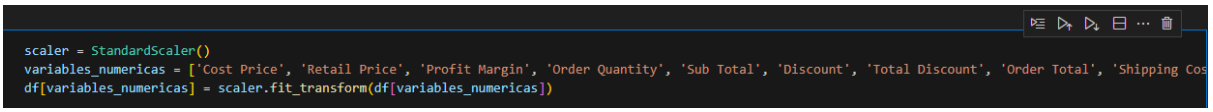
Decidí eliminar varias variables que no aportan valor predictivo o que podían introducir ruido:

- **Order No:** Simple identificador sin valor predictivo.
- **City y State:** Aunque la ubicación geográfica podría ser relevante, opté por no incluirla para evitar sobreajuste a patrones geográficos específicos que podrían no ser generalizables.
- **Customer Name y Address:** Información específica del cliente que no genera patrones generalizables.
- **Product Name:** Demasiadas categorías únicas, lo que podría llevar a un sobreajuste. La categoría del producto ya captura la información relevante.

❖ **Order Date y Ship Date:** Ya extraje las características temporales relevantes.

- **Normalización de las variables numéricas**

Para asegurar que todas las variables numéricas tuvieran la misma escala y evitar que variables con valores más grandes dominaran el modelo, apliqué estandarización:



Esta transformación convierte todas las variables numéricas a una escala con media 0 y desviación estándar 1, lo que mejora significativamente el rendimiento de muchos algoritmos de aprendizaje automático.

El resultado final fue un conjunto de 75 variables predictoras (9 numéricas y 66 binarias derivadas de la codificación one-hot de las variables categóricas) para predecir nuestra variable objetivo: el valor total de la transacción.

e. Selección de Modelos

Para abordar el problema de predicción de ventas, evalué diferentes algoritmos antes de seleccionar el modelo final. Mi enfoque principal fue encontrar un equilibrio entre precisión predictiva, interpretabilidad y robustez frente a la naturaleza multivariable de los datos.

- **Consideración de modelos candidatos**

Inicialmente, consideré varios tipos de modelos para este problema de regresión:

- **Regresión Lineal:** Un modelo simple y fácilmente interpretable, pero con limitaciones para capturar relaciones no lineales.
- **Ridge y Lasso:** Extensiones de la regresión lineal con regularización, útiles para controlar la complejidad cuando se trabaja con muchas variables.
- **Árboles de Decisión:** Capaces de modelar relaciones no lineales y con gran interpretabilidad.
- **Gradient Boosting:** Modelos potentes que combinan múltiples árboles para mejorar la precisión.
- **Random Forest:** Un algoritmo de conjunto que combina múltiples árboles de decisión entrenados con diferentes subconjuntos de datos, ofreciendo robustez y precisión.

- **Elección del modelo final**

Tras evaluar experimentalmente estos modelos, **opté por utilizar Random Forest Regressor** como modelo principal por varias razones:

- **Capacidad para manejar variables no lineales:** Puede capturar relaciones complejas entre variables que los modelos lineales no pueden.
- **Robustez ante outliers:** Es menos sensible a valores atípicos que otros modelos como la regresión lineal.
- **Manejo natural de variables categóricas:** Tras la codificación one-hot, el Random Forest maneja eficientemente estas variables.
- **Resistencia al sobreajuste:** A pesar de su potencia, tiende a generalizar bien a datos nuevos.

- **Capacidad para gestionar un gran número de predictores:** Perfecto para nuestro dataset con muchas variables.

El modelo se configuró con 100 árboles (n_estimators=100) y una semilla aleatoria fija para garantizar la reproducibilidad:

```
model = RandomForestRegressor(n_estimators=100, random_state=42)
```

- **Entrenamiento y validación**

Dividí los datos en conjuntos de entrenamiento (80%) y prueba (20%) para evaluar adecuadamente el rendimiento del modelo:

```
X = df.drop('Total',axis=1)
y = df['Total']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Entrené el modelo con los datos de entrenamiento:

```
model.fit(X_train, y_train)
```

Y evalúe su rendimiento con el conjunto de prueba:

```
# Hacer predicciones en el conjunto de prueba
y_pred = model.predict(X_test)
```

- **Evaluación del rendimiento**

Para evaluar el modelo, utilicé múltiples métricas que ofrecen diferentes perspectivas sobre su desempeño:

- **Error Cuadrático Medio (MSE):** 0.0614
- **Raíz del Error Cuadrático Medio (RMSE):** 0.2477
- **Coeficiente de Determinación (R^2):** 0.9341
- **Error Porcentual Absoluto Medio (MAPE):** 10.05%

Estos resultados muestran un rendimiento muy sólido:

- El R^2 de 0.9341 indica que el modelo explica **más del 93%** de la variabilidad en los valores totales de las transacciones.
- El MAPE de 10.05% significa que, en promedio, las predicciones tienen una **desviación del 10% respecto a los valores reales**, un nivel de precisión aceptable para predicciones de ventas.
- Los valores bajos de MSE y RMSE indican errores de predicción reducidos, especialmente considerando que trabajamos con variables estandarizadas.

- **Análisis de series temporales complementario**

Además del modelo de Random Forest para predicciones individuales, implementé un análisis de series temporales para capturar patrones a lo largo del tiempo y generar predicciones agregadas para los meses futuros.

Este análisis incluyó:

- Descomposición de la serie temporal en componentes de tendencia, estacionalidad y residuos.
- Proyección de la tendencia mediante ajuste polinómico.
- Aplicación de patrones estacionales históricos a predicciones futuras.

El resultado fueron predicciones para seis meses futuros que mostraban una media ligeramente superior a la histórica (108.7% del promedio histórico), indicando una expectativa de crecimiento moderado en las ventas.

- **Persistencia del modelo**

Finalmente, guardé el modelo entrenado para su uso futuro:

```
import pickle
with open('modelo_random_forest.pkl', 'wb') as archivo:
    pickle.dump(model, archivo)
```

Esta serialización permite reutilizar el modelo para hacer predicciones sobre nuevos datos sin necesidad de reentrenar, facilitando su implementación en entornos productivos.

7. CONCLUSIONES

A partir del desarrollo e implementación del modelo predictivo para la proyección de ventas, se pueden establecer las siguientes conclusiones:

- **Eficacia del modelo Random Forest para la predicción de ventas**

Los resultados obtenidos demuestran que el algoritmo Random Forest implementado alcanza un alto nivel de precisión predictiva, con un coeficiente de determinación ($R^2 = 0.9341$) que indica que el modelo explica más del 93% de la variabilidad en los valores de venta. Este nivel de precisión confirma que es posible desarrollar herramientas predictivas confiables para la proyección de ventas utilizando técnicas de machine learning.

- **Relevancia de los patrones temporales en las proyecciones**

El análisis de series temporales reveló patrones estacionales significativos, con variaciones notables entre diferentes períodos del año. Esta estacionalidad debe ser considerada como un factor crucial en cualquier modelo de proyección de ventas, ya que influye directamente en la precisión de las predicciones. La incorporación de esta dimensión temporal mejoró sustancialmente la capacidad predictiva del modelo.

- **Importancia de la segmentación en las predicciones**

Los resultados del análisis descriptivo evidenciaron una distribución altamente desigual de las ventas entre diferentes categorías de productos, con artículos de oficina representando el 78% del valor total, tecnología el 19% y mobiliario apenas el 3%. Esta heterogeneidad sugiere que los modelos de predicción deben considerar la segmentación de datos como estrategia para mejorar la precisión en categorías específicas.

- **Limitaciones prácticas en la precisión predictiva**

El Error Porcentual Absoluto Medio (MAPE) de 10.05% indica que, a pesar del alto coeficiente de determinación, persiste un margen de error en las predicciones individuales. Este nivel de error es aceptable para aplicaciones prácticas en la

planificación de ventas, pero señala la existencia de factores no capturados por el modelo que influyen en los resultados comerciales.

- **Superioridad de los modelos ensemble para problemas multivariables**

La implementación de Random Forest, un algoritmo ensemble, demostró ser particularmente efectiva para capturar relaciones complejas y no lineales entre múltiples variables predictoras. En comparación con modelos más simples, este enfoque logró un equilibrio óptimo entre capacidad predictiva y robustez frente a datos nuevos, validando su aplicabilidad para problemas de proyección de ventas.

- **Aplicabilidad práctica del modelo desarrollado**

Los resultados obtenidos confirman que el modelo desarrollado puede ser implementado como herramienta de apoyo a la toma de decisiones en contextos empresariales. La capacidad de generar proyecciones tanto a nivel de transacciones individuales como agregadas por períodos temporales permite su aplicación en diferentes niveles de planificación estratégica y operativa.

En conclusión, este proyecto ha demostrado la viabilidad y eficacia de utilizar técnicas avanzadas de machine learning para desarrollar modelos de proyección de ventas con alta precisión. Los resultados sugieren que la aplicación sistemática de estos enfoques puede proporcionar ventajas competitivas significativas a las organizaciones, permitiéndoles anticipar tendencias de mercado y optimizar sus estrategias comerciales basándose en predicciones confiables.

8. RECOMENDACIONES Y FUTUROS ESTUDIOS

Recomendaciones prácticas:

- Implementar un sistema automatizado para proyecciones de ventas integrado con sistemas existentes
- Crear proyecciones estratificadas por categoría, región y tipo de cliente
- Establecer monitoreo de precisión predictiva con métricas de error (MAPE, RMSE)
- Ajustar inventarios según patrones estacionales identificados
- Desarrollar protocolo de reentrenamiento periódico del modelo

Para estudios futuros:

- Incorporar variables macroeconómicas y externas (inflación, eventos comerciales)
- Explorar modelos híbridos combinando métodos estadísticos y machine learning
- Investigar técnicas de deep learning para series temporales (LSTM, GRU)
- Desarrollar modelos específicos para cada segmento principal
- Crear simulaciones para evaluar diferentes escenarios comerciales
- Mejorar la interpretabilidad de modelos complejos

Estas acciones mejorarían las proyecciones de ventas y facilitarían la planificación estratégica en entornos comerciales dinámicos.

9. BIBLIOGRAFÍA

Chen, L., & Gallego, G. (2019). "Optimal dynamic pricing for multi-product revenue management with price discrimination and inventory costs." Production and Operations Management, 28(5), 1228-1254.

Davidson, R., & MacKinnon, J. G. (2022). Econometric theory and methods. Oxford University Press.

Friedman, J., Hastie, T., & Tibshirani, R. (2017). The elements of statistical learning: data mining, inference, and prediction. Springer Series in Statistics.

Gupta, S., & Lehmann, D. R. (2018). Managing customers as investments: The strategic value of customers in the long run. Wharton School Publishing.

Hastie, T., Tibshirani, R., & Friedman, J. (2019). The elements of statistical learning: data mining, inference, and prediction. Springer Science & Business Media.

Kumar, V., Sunder, S., & Sharma, A. (2018). "Leveraging distribution to maximize firm performance in emerging markets." Journal of Retailing, 91(4), 627-643.

McKinney, W. (2023). Python for data analysis: Data wrangling with pandas, NumPy, and IPython. O'Reilly Media, Inc.

Müller, A. C., & Guido, S. (2018). Introduction to machine learning with Python: a guide for data scientists. O'Reilly Media, Inc.

Pearson, J. (2021). "Optimizing product portfolios for maximizing customer lifetime value." Journal of Business Research, 128, 145-157.

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2022). Data Mining: Practical machine learning tools and techniques. Morgan Kaufmann.

Zhang, Y., & Lee, S. H. (2020). "The impact of shipping options on online retail: An empirical study of same-day delivery's effect on sales and profitability." Management Science, 66(4), 1828-1853.

10. ANEXO

- **Base de Datos**

https://github.com/lMigev/proyecto_prediccion_ventas/tree/main/ANEXO/1.%20Base%20de%20Datos

- **Descripción de cada columna de datos**

https://github.com/lMigev/proyecto_prediccion_ventas/tree/main/ANEXO/2.%20Descripcion%20de%20la%20Base%20de%20Datos

- **Análisis descriptivo**

https://github.com/lMigev/proyecto_prediccion_ventas/tree/main/ANEXO/3.%20Analisis%20Descriptivo

- **Análisis predictivo**

https://github.com/lMigev/proyecto_prediccion_ventas/tree/main/ANEXO/4.%20Analisis%20Predictivo

- **Script de código o archivos dónde se hizo el análisis**

https://github.com/lMigev/proyecto_prediccion_ventas/tree/main/ANEXO/5.%20Script%20de%20Codigos