**📊 Análisis de Tendencias de Compra**

**Proyecto de Ciencia de Datos**

**Introducción**

Este proyecto tiene como objetivo analizar el comportamiento de compra de los consumidores utilizando un dataset de transacciones. Buscamos descubrir patrones clave que puedan ayudar a mejorar estrategias de ventas, marketing y atención al cliente en el sector minorista.

A través de la exploración de datos, limpieza, análisis y visualización, extraeremos información valiosa para la toma de decisiones estratégicas.

**Objetivos del Proyecto**

✅ Identificar qué categorías de productos tienen mayor demanda.  
✅ Analizar las diferencias en el comportamiento de compra según el género.  
✅ Evaluar el impacto de las estaciones del año en los montos de compra.  
✅ Determinar los métodos de pago más utilizados.  
✅ Extraer insights para mejorar estrategias de venta y marketing.

**Preguntas Clave**

✩ ¿Cuál es la categoría de productos más comprada?  
✩ ¿Existe diferencia en el comportamiento de compra entre géneros?  
✩ ¿Durante qué temporada del año se gasta más dinero?  
✩ ¿Cuál es el método de pago preferido por los clientes?  
✩ ¿En qué rango de precios se concentran la mayoría de las compras?

**Público Objetivo**

📌 **Empresas de Retail y E-commerce**: Para optimizar estrategias de ventas y promociones.  
📌 **Departamentos de Marketing**: Para diseñar campañas dirigidas según el comportamiento del consumidor.  
📌 **Analistas de Datos**: Para identificar tendencias y mejorar la toma de decisiones basada en datos.  
📌 **Equipos Financieros**: Para mejorar la gestión de ingresos y previsiones de ventas.

**Metodología y Herramientas Utilizadas**

🖥️ **Lenguaje de Programación:** Python  
📂 **Bibliotecas Utilizadas:** Pandas, Seaborn, Matplotlib  
🛠️ **Procesos Realizados:**

* Carga y exploración de datos.
* Limpieza y transformación de datos.
* Análisis exploratorio con visualización de datos.
* Generación de conclusiones basadas en hallazgos.

**Conclusiones Principales**

📊 **Categoría más comprada:** 'Clothing' es la categoría con mayor número de compras, indicando una alta demanda en ese segmento.  
👥 **Diferencias de compra por género:** 'Male' es el género que realiza más compras, lo que sugiere estrategias de marketing específicas.  
📆 **Temporada de mayor gasto:** 'Fall' es la estación del año con el gasto promedio más alto, clave para campañas promocionales.  
💳 **Método de pago más utilizado:** 'Credit Card' es la forma de pago preferida por los consumidores, lo que destaca su importancia en plataformas de venta.  
💰 **Rango de montos de compra:** La mayoría de las compras están entre **$20.00 a $100.00 USD**, lo que sugiere el rango óptimo para promociones y descuentos.

**🔍 Análisis Avanzado:**

## 🧠 Enfoque del Análisis: Supervisado y No Supervisado

En este trabajo se propone realizar **dos tipos de análisis de datos complementarios**, utilizando técnicas de Machine Learning:

### 1. 📊 Análisis No Supervisado

También se aplicará un análisis no supervisado, específicamente **clustering (agrupamiento)**, para **identificar patrones naturales o grupos de clientes** con características similares, sin necesidad de una variable objetivo.

* **Objetivo:** explorar la estructura subyacente de los datos y segmentar a los clientes en grupos homogéneos.
* **Algoritmo utilizado:** K-Means.
* **Utilidad:** descubrir perfiles de usuarios y mejorar la personalización de estrategias comerciales o de fidelización.

### 2. 🔍 Análisis Supervisado

Se aplicarán modelos de aprendizaje supervisado con el objetivo de **predecir si un cliente utilizará un código promocional**, basándonos en sus características (edad, método de pago, temporada, etc.).

* **Objetivo:** construir un modelo predictivo que permita anticipar comportamientos de los clientes.
* **Algoritmos utilizados:** Árbol de Decisión
* **Utilidad:** identificar segmentos con mayor propensión a usar promociones, lo que puede guiar campañas de marketing más efectivas.

**1)Segmentación de Clientes**

Con el objetivo de obtener una comprensión más profunda del comportamiento de los clientes, se aplicó un modelo de clustering no supervisado (K-Means) sobre los datos de compra. Este enfoque permitió identificar grupos de clientes con características y patrones de consumo similares, lo cual aporta valor para diseñar estrategias de marketing personalizadas.

**🧪 Metodología**

Se seleccionaron las siguientes variables para construir los perfiles de cliente:

Género

Temporada de compra

Método de pago

Categoría del producto

Monto de compra

Las variables categóricas fueron transformadas con One-Hot Encoding y todos los datos fueron estandarizados antes del entrenamiento. Se eligió un valor de k = 4 para segmentar a los clientes en cuatro grupos distintos.

**📊 Resultados del Modelo**

| Cluster | Género Predominante | Temporada Más Común | Método de Pago Principal | Gasto Promedio (USD) |
| --- | --- | --- | --- | --- |
| 0 | Female | Fall | Credit Card | 61.18 |
| 1 | Male | Winter | Cash | 59.93 |
| 2 | Female | Spring | Venmo | 58.94 |
| 3 | Female | Summer | Bank Transfer | 59.87 |

**🧠 Conclusiones Estratégicas por Cluster**

Cluster 0: Mujeres que compran principalmente en otoño y prefieren pagar con tarjeta de crédito. Ideal para promociones de temporada con descuentos bancarios.

Cluster 1: Hombres que compran en invierno y prefieren pagar en efectivo. Potenciales compradores tradicionales o de tiendas físicas.

Cluster 2: Mujeres que compran en primavera y pagan con Venmo. Perfil joven o digital, ideal para campañas por redes sociales.

Cluster 3: Mujeres que compran en verano y usan transferencias bancarias. Potencialmente más planificadas; pueden responder a ofertas anticipadas o exclusivas.

**2) Análisis con Árbol de Decisión**

En este proyecto se abordó la tarea de clasificar si un cliente usará un código promocional o no, a partir de variables demográficas y de comportamiento de compra. El proceso incluyó las siguientes etapas:

1. **Preparación de datos:**
   * Se seleccionaron variables predictoras relevantes y la variable objetivo, la cual fue codificada binariamente (1 = usa código promocional, 0 = no usa).
   * Las variables categóricas fueron codificadas mediante one-hot encoding para ser compatibles con el modelo.
2. **Entrenamiento inicial y evaluación:**
   * Se entrenó un árbol de decisión con profundidad máxima limitada para evitar sobreajuste.
   * Se evaluó el modelo usando un split de entrenamiento y prueba, obteniendo una accuracy cercana al 74%, con un buen desempeño en identificar clientes que usaron código, pero cierta dificultad en discriminar quienes no lo usaron.
3. **Validación cruzada:**
   * Se implementó validación cruzada 5-fold para obtener una evaluación más estable y robusta del modelo.
   * Se observó que el rendimiento promedio fue similar (aprox. 74-75%), pero con alta variabilidad entre folds, indicando cierta inestabilidad.
4. **Ajuste de hiperparámetros:**
   * Se aplicó Grid Search para explorar combinaciones de parámetros del árbol (profundidad, criterio, tamaño mínimo de splits y hojas).
   * Los mejores parámetros encontrados fueron coincidentes con la configuración inicial, sin mejora significativa en la métrica de accuracy.
5. **Conclusiones:**
   * El modelo de árbol de decisión simple es capaz de identificar bien los clientes que usan código promocional, pero presenta un número considerable de falsos positivos.
   * La inestabilidad observada sugiere que el modelo es sensible a la partición de los datos.
   * Para mejorar el desempeño, se recomienda explorar modelos más complejos como Random Forest, aplicar técnicas de balanceo de clases o realizar ingeniería adicional de variables.

### 🔍 Conclusiones clave:

### 🏦 Método de pago y preferencia influyen en el comportamiento:

* Los clientes que **no utilizan PayPal** y **prefieren medios como Venmo o tarjetas de crédito/débito** presentan una **mayor probabilidad de usar un código promocional**.
* Este comportamiento puede asociarse a un perfil más **familiarizado con herramientas digitales** y **beneficios online**, como promociones y cupones.

### 👶 La edad es un factor relevante:

* El modelo detectó que los **usuarios menores a 35 años** son más propensos a utilizar códigos promocionales.
* Esto puede deberse a una **mayor sensibilidad al precio**, hábito de buscar descuentos o mayor interacción con canales promocionales digitales.

### 🎯 Segmentación útil para marketing:

* Gracias a la interpretación del árbol de decisión, es posible **identificar segmentos ideales para campañas promocionales**, como:
  + Jóvenes adultos (menos de 35 años),
  + Que no utilizan PayPal,
  + Y que tienen preferencia por medios digitales de pago.

### ✅ Valor del modelo:

* Más allá del rendimiento (accuracy ~76%), este modelo permite **interpretar fácilmente las reglas de decisión**, lo que lo convierte en una herramienta útil para áreas de marketing, fidelización o planificación de campañas.