SISTEMA DE CLASIFICACION

Satisfacción y Segmentación de Clientes en E-commerce

INFORME PROYECTO FINAL

BOOTCAMP DE INTELIGENCIA ARTIFICIAL, MACHINE LEARNING Y ANALITICA AVANZADA

Grupo # 4

Emilio Palacín G.
Oscar Espinoza
Javier Ballén
José Navarro

Profesor:
Ing. Carlos Rafhael Satizabal Sánchez

Febrero 06 de 2026

Planteamiento

CONTEXTO

- El sector e-commerce recibe grandes volúmenes de reseñas de clientes.
- La mayoría de este feedback es texto no estructurado.
- Analizar manualmente esta información no es viable.

PROBLEMA

- La insatisfacción del cliente no se detecta a tiempo.
- Esto genera pérdida de clientes y decisiones tardías.

NECESIDAD

- Transformar reseñas en información accionable mediante analítica avanzada

Objetivo General

Desarrollar un modelo de clasificación y un sistema de análisis de datos que permita:

- Predecir la satisfacción del cliente
- Segmentar el mercado
- Apoyar la toma de decisiones en e-commerce
- Usando técnicas de Machine Learning y análisis de sentimiento

Objetivos Específicos

- Realizar análisis exploratorio de datos (EDA) e ingeniería de características.
- Entrenar y validar modelos de clasificación para predecir recomendación del cliente.
- Aplicar técnicas de clustering para segmentar clientes.
- Visualizar resultados en un dashboard interactivo en Power BI.

Alcance

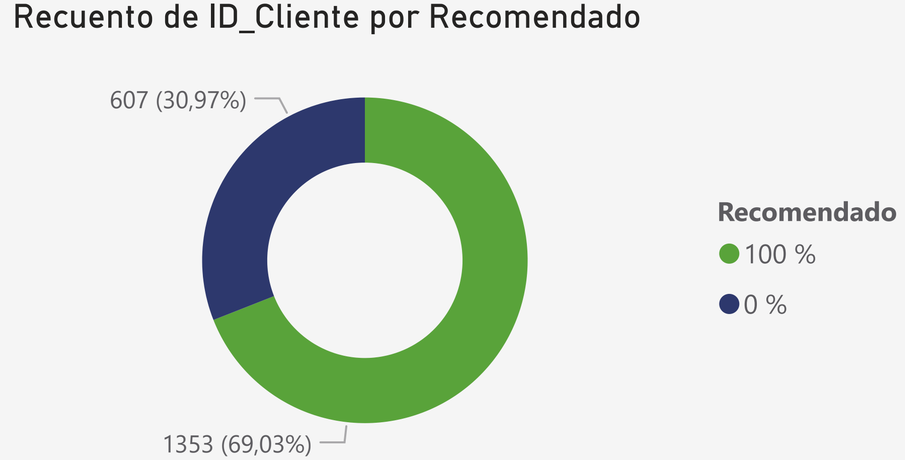
- Dataset con 1.960 registros del sector retail.
- Procesamiento y modelado en Python.
- Modelos supervisados y no supervisados.
- Visualización de indicadores clave.

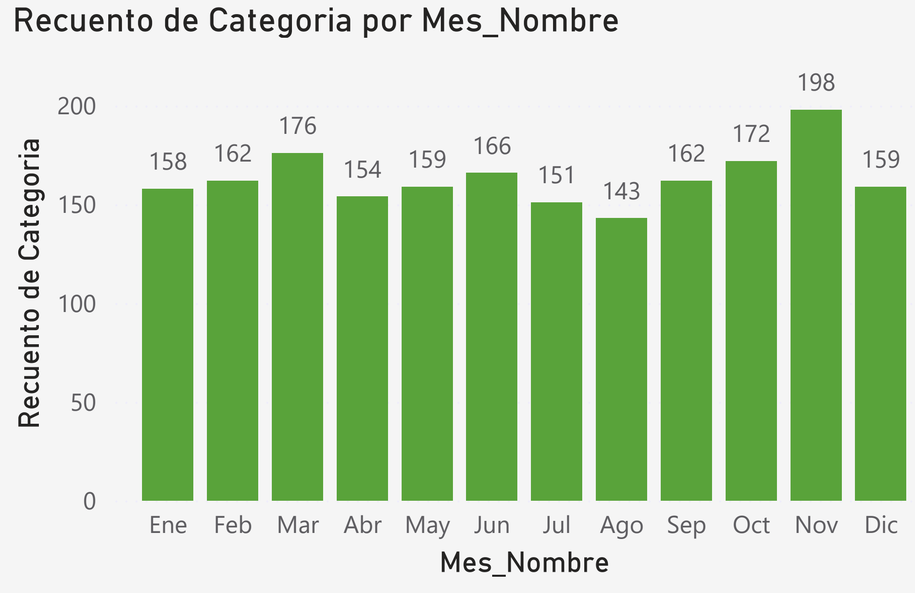
```
--- INFO INICIAL ---
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2000 entries, 0 to 1999
Data columns (total 10 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   ID_Cliente     2000 non-null   int64  
 1   Edad           1980 non-null   float64
 2   Categoria      2000 non-null   object 
 3   Precio         2000 non-null   float64
 4   Calificacion   2000 non-null   int64  
 5   Texto_Resena   1960 non-null   object 
 6   Fecha_Compra   2000 non-null   object 
 7   Recomendado    2000 non-null   int64  
 8   Departamento   2000 non-null   object 
 9   Tipo_Envio     2000 non-null   object 
dtypes: float64(2), int64(3), object(5)
memory usage: 156.4+ KB
None
<bound method NDFrame.head of       ID_Cliente  Edad   Categoria  Precio  Calificacion  \
0           1000   NaN   Chaquetas   39.76             5   
1           1001  36.0     Vestidos  131.93             4   
2           1002  19.0   Chaquetas   43.67             2   
3           1003  29.0  Pantalones   15.95             1   
4           1004  59.0   Chaquetas   64.46             4   
...          ...   ...         ...     ...           ...   
1995        2995  37.0   Chaquetas  143.35             4   
1996        2996  30.0       Blusas   29.04             3   
1997        2997  41.0   Chaquetas   20.21             2   
1998        2998  65.0  Accesorios   24.65             1   
1999        2999  21.0  Pantalones   74.82             2   

                                  Texto_Resena Fecha_Compra  Recomendado  \
0                                          NaN   2023-04-13            1   
1     Totalmente recomendado, me queda perfecto.   2023-05-30            1   
2                 Mala calidad, costuras sueltas.   2023-06-10            0   
3        El envío tardó demasiado y llegó dañado.   2023-02-28            0   
4          Muy cómodo y elegante para la oficina.   2023-02-20            1   
...                                           ...          ...          ...   
1995  Me encanta este producto, la tela es suave.   2023-02-28            1   
1996            Llegó a tiempo, calidad aceptable.   2023-05-26            0   
1997       El envío tardó demasiado y llegó dañado.   2023-01-14            0   
1998        Talla incorrecta, es demasiado pequeño.   2023-08-05            0   
1999    No lo recomiendo, muy caro para lo que es.   2023-02-17            0   

     Departamento Tipo_Envio  
0      Ropa Mujer   Estándar  
1         Deportes    Express  
2     Ropa Hombre   Estándar  
3     Ropa Hombre   Estándar  
4           Niños   Estándar  
...           ...         ...  
1995     Deportes    Express  
1996   Ropa Mujer   Estándar  
1997  Ropa Hombre   Estándar  
1998   Ropa Mujer   Estándar  
1999        Niños   Estándar  
```

Metodología

Metodología utilizada CRISP-DM (adaptada)

- Comprensión del negocio
- Comprensión de los datos
- Preparación de los datos
- Modelado
- Evaluación
- Visualización y análisis de resultados





Tratamiento de Datos (Data Cleaning)

Tratamiento de nulos en el DataFrame

- Objetivo:

- Imputar los valores faltantes de la columna 'Edad' usando la mediana de esa columna
- Eliminar filas que no tengan texto en la columna 'Texto_Resena' (importante para tareas de NLP).

- Efectos:

- Modifica la columna 'Edad' rellenando NaN con la mediana actual de la columna.
- Elimina filas del DataFrame que tengan NaN en 'Texto_Resena'.

- Consideraciones:

- Se asume que `df` es un pandas.DataFrame y contiene las columnas `Edad` y `Texto_Resena`.
- La imputación con la mediana preserva la escala central de la variable numérica.
- La eliminación de filas es crucial antes de convertir texto a vectores (p. ej. CountVectorizer).
- El uso de `inplace=True` en `dropna` modifica `df` en sitio.

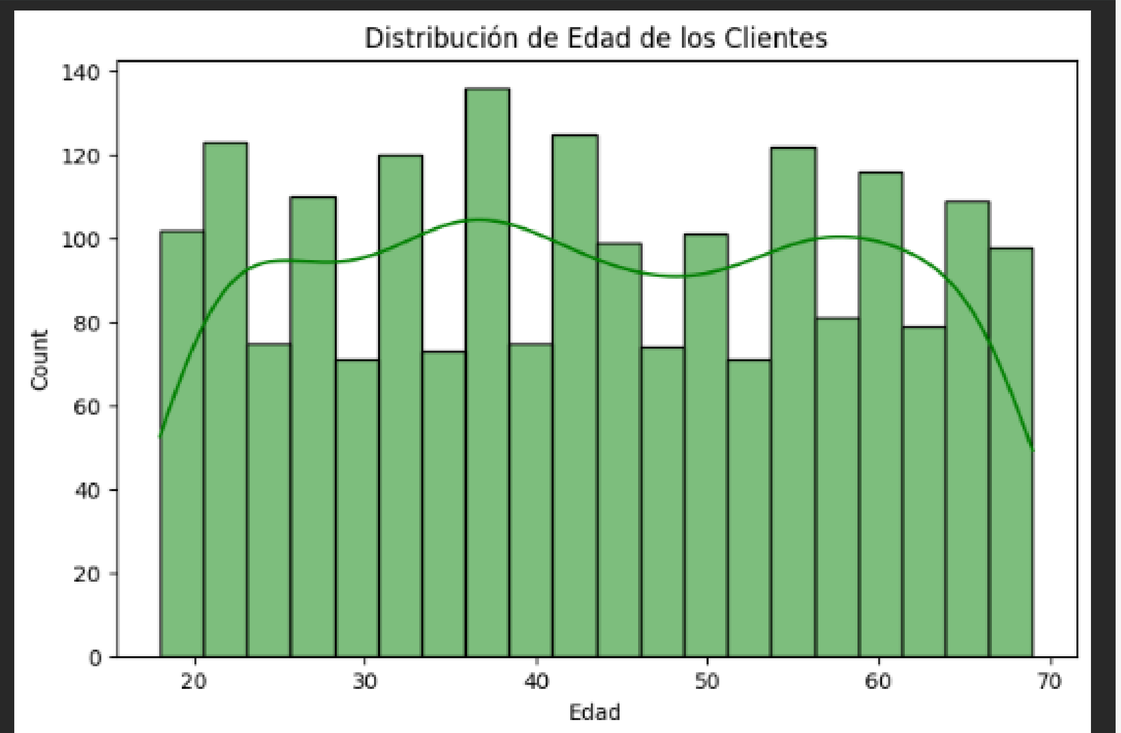
```python
import pandas as pd

df = pd.read_csv('/content/dataset_ecommerce_moda.csv')
df['Edad'] = df['Edad'].fillna(df['Edad'].median())
df.dropna(subset=['Texto_Resena'], inplace=True)
print(df.head)
```

```
<bound method NDFrame.head of       ID_Cliente  Edad   Categoria  Precio  Calific
1           1001  36.0    Vestidos  131.93            4
2           1002  19.0   Chaquetas   43.67            2
3           1003  29.0  Pantalones   15.95            1
4           1004  59.0   Chaquetas   64.46            4
5           1005  61.0   Chaquetas   78.01            5
...          ...   ...         ...     ...          ...
1995        2995  37.0   Chaquetas  143.35            4
1996        2996  30.0      Blusas   29.04            3
1997        2997  41.0   Chaquetas   20.21            2
1998        2998  65.0  Accesorios   24.65            1
1999        2999  21.0  Pantalones   74.82            2

                                  Texto_Resena Fecha_Compra  Recomendado  \
1     Totalmente recomendado, me queda perfecto.   2023-05-30            1
2              Mala calidad, costuras sueltas.   2023-06-10            0
3       El envío tardó demasiado y llegó dañado.   2023-02-28            0
4         Muy cómodo y elegante para la oficina.   2023-02-20            1
5     Superó mis expectativas, volveré a comprar.   2023-11-25            1
...                                         ...          ...          ...
1995  Me encanta este producto, la tela es suave.   2023-02-28            1
1996          Llegó a tiempo, calidad aceptable.   2023-05-26            0
1997    El envío tardó demasiado y llegó dañado.   2023-01-14            0
1998     Talla incorrecta, es demasiado pequeño.   2023-08-05            0
1999   No lo recomiendo, muy caro para lo que es.   2023-02-17            0

     Departamento Tipo_Envio
1       Deportes    Express
2    Ropa Hombre   Estándar
3    Ropa Hombre   Estándar
4          Niños   Estándar
5          Niños   Estándar
...          ...        ...
1995    Deportes    Express
1996  Ropa Mujer   Estándar
1997 Ropa Hombre   Estándar
1998  Ropa Mujer   Estándar
1999       Niños   Estándar

[1960 rows x 10 columns]>
```

Análisis Exploratorio de Datos (EDA)

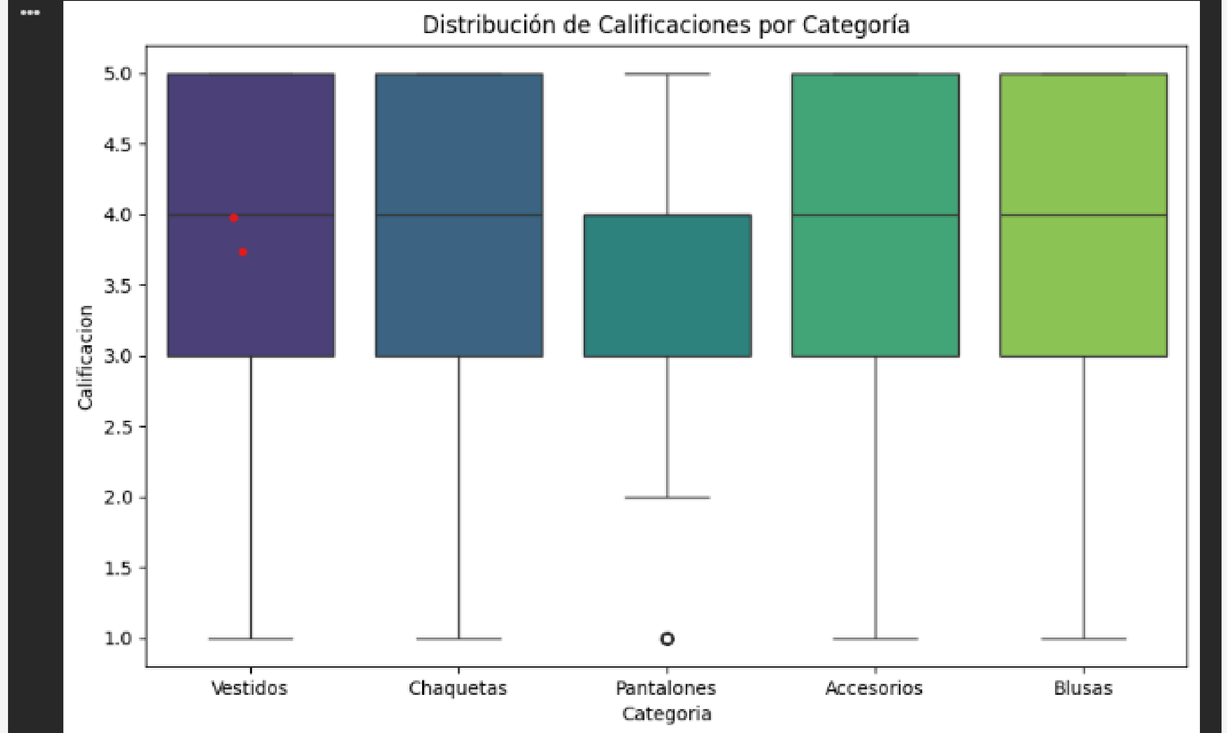
Visualización de la densidad demográfica y los picos de edad que definen el mercado objetivo.

```python
# Gráfico 1: Distribución de Edades
plt.figure(figsize=(8, 5))
sns.histplot(df['Edad'], bins=20, kde=True, color='green')
plt.title('Distribución de Edad de los Clientes')
plt.savefig('grafico_edad.png') # Guarda la imagen
plt.show()
```



Identifica la variabilidad en la satisfacción (calificaciones) y detecta valores atípicos

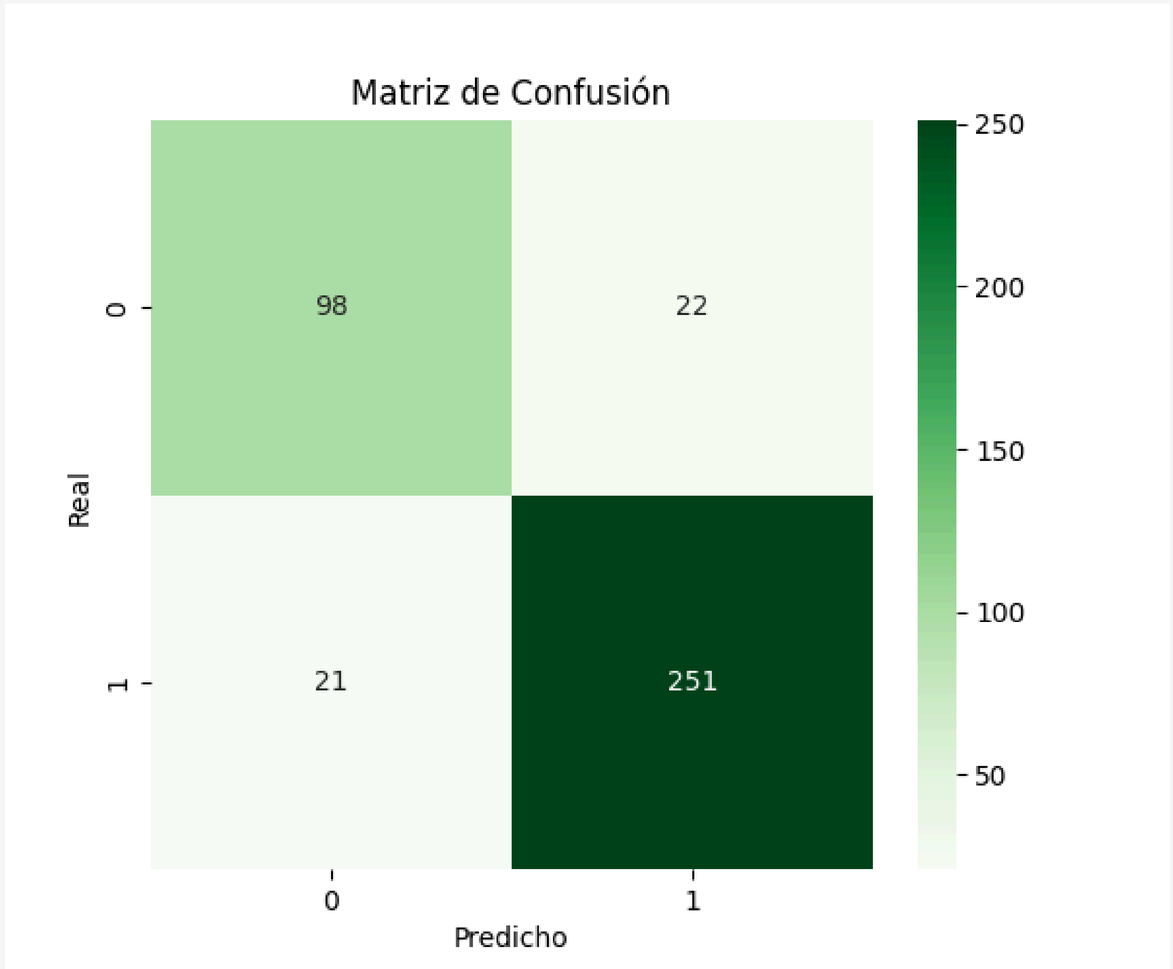
```python
# Gráfico 2: Calificaciones por Categoría
plt.figure(figsize=(10, 6))
sns.boxplot(x='Categoria', y='Calificacion', hue='Categoria', data=df, palette='viridis', legend=False)
plt.title('Distribución de Calificaciones por Categoría')
plt.savefig('grafico_boxplot.png')
plt.show()
```



Validación de la precisión del modelo **Random Forest**.
Muestra qué tan bien el sistema predice si un cliente recomendará un producto (Verdaderos Positivos vs. Falsos Negativos).

```python
y_pred = rf.predict(X_test)
print("\n--- REPORTE DE CLASIFICACIÓN (MODELO SUPERVISADO) ---")
print(classification_report(y_test, y_pred))

"""
Visualización de la Matriz de Confusión.

- Objetivo:
    Mostrar gráficamente el rendimiento del modelo de clasificación comparando
    las etiquetas reales (y_test) con las predicciones (y_pred).

    Linea 22.  `sns.heatmap(...)`: Genera un mapa de calor con la matriz de confusi
        - `confusion_matrix(y_test, y_pred)`: Calcula la matriz (filas=real, column
        - `annot=True`: Muestra los valores numéricos en cada celda.
        - `fmt='d'`: Formato entero para los números.
        - `cmap='Blues'`: Paleta de colores azul.
- Interpretación de la matriz:
    - Diagonal principal: Predicciones correctas (Verdaderos Positivos y Verdaderos
    - Fuera de la diagonal: Errores del modelo (Falsos Positivos y Falsos Negativos
"""
plt.figure(figsize=(6,5))
sns.heatmap(confusion_matrix(y_test, y_pred), annot=True, fmt='d', cmap='Greens')
plt.title('Matriz de Confusión')
plt.ylabel('Real')
plt.xlabel('Predicho')
plt.savefig('grafico_matriz.png')
plt.show()
```

```
...
--- REPORTE DE CLASIFICACIÓN (MODELO SUPERVISADO) ---
              precision    recall  f1-score   support

           0       0.82      0.82      0.82       120
           1       0.92      0.92      0.92       272

    accuracy                           0.89       392
   macro avg       0.87      0.87      0.87       392
weighted avg       0.89      0.89      0.89       392
```

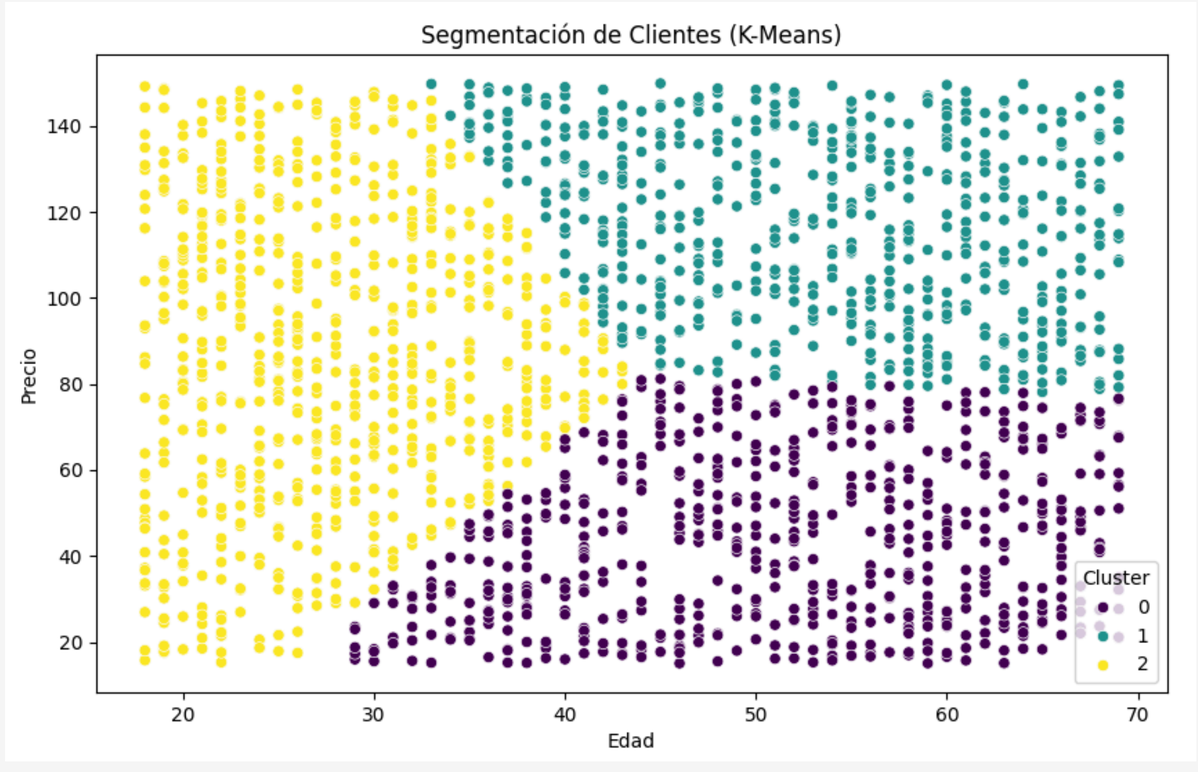


Visualización de la segmentación lograda con **K-Means**.
Es fundamental para que el área de Marketing entienda cómo se agrupan los clientes por su comportamiento de gasto y edad.

```python
# Clustering: Agrupar clientes por Edad y Precio (Gasto)
X_cluster = df[['Edad', 'Precio']]
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X_cluster)
```

```python
# K-Means con 3 clústeres
kmeans = KMeans(n_clusters=3, random_state=42, n_init=10)
df['Cluster'] = kmeans.fit_predict(X_scaled)

```

```python
"""
Visualización de Clústeres mediante gráfico de dispersión.

- Objetivo:
    Mostrar gráficamente la segmentación de clientes realizada por el algoritmo K-Means,
    representando cada cliente según su Edad y Precio (gasto).

- Interpretación:
    - Cada punto representa un cliente.
    - Los colores indican a qué clúster pertenece cada cliente (0, 1 o 2).
    - Permite identificar patrones de segmentación basados en edad y comportamiento de gasto.
"""

plt.figure(figsize=(10, 6))
sns.scatterplot(x='Edad', y='Precio', hue='Cluster', data=df, palette='viridis')
plt.title('Segmentación de Clientes (K-Means)')
plt.savefig('grafico_cluster.png')
plt.show()
```



SISTEMA DE CLASIFICACIÓN: Satisfacción y Segmentación de Clientes de E-Commerce

Total Clientes	Nivel Promedio Satisfacion NPS (Proy)	Precio Promedio	Satisfacción Global
1960	69,03 %	7530	3,59

Meses

1	4	7	10
2	5	8	11
3	6	9	12

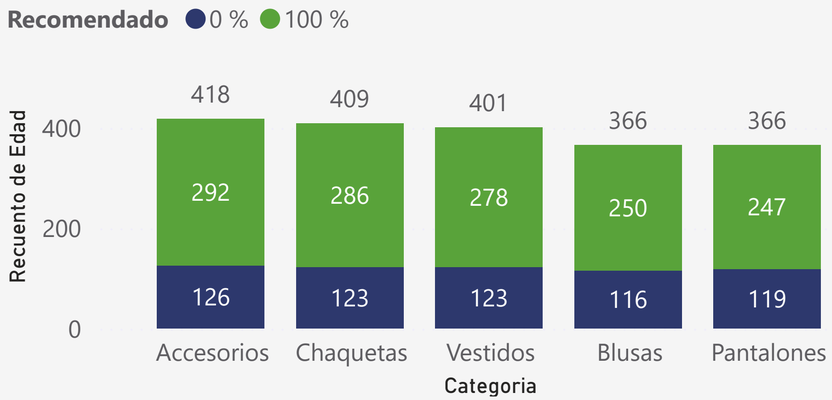
Categoría

Accesorios	Pantalones
Blusas	Vestidos
Chaquetas	

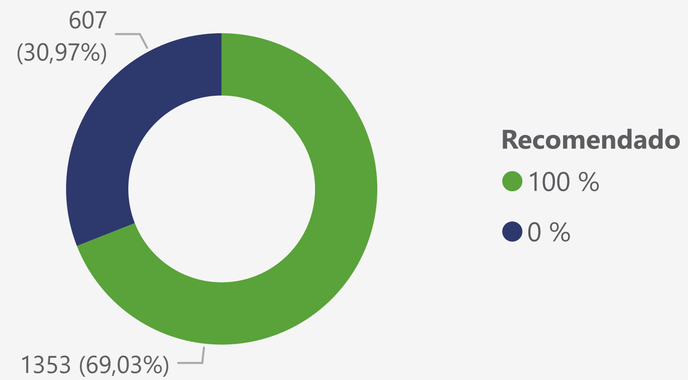
Departamento

Deportes	Ropa Hombre
Niños	Ropa Mujer

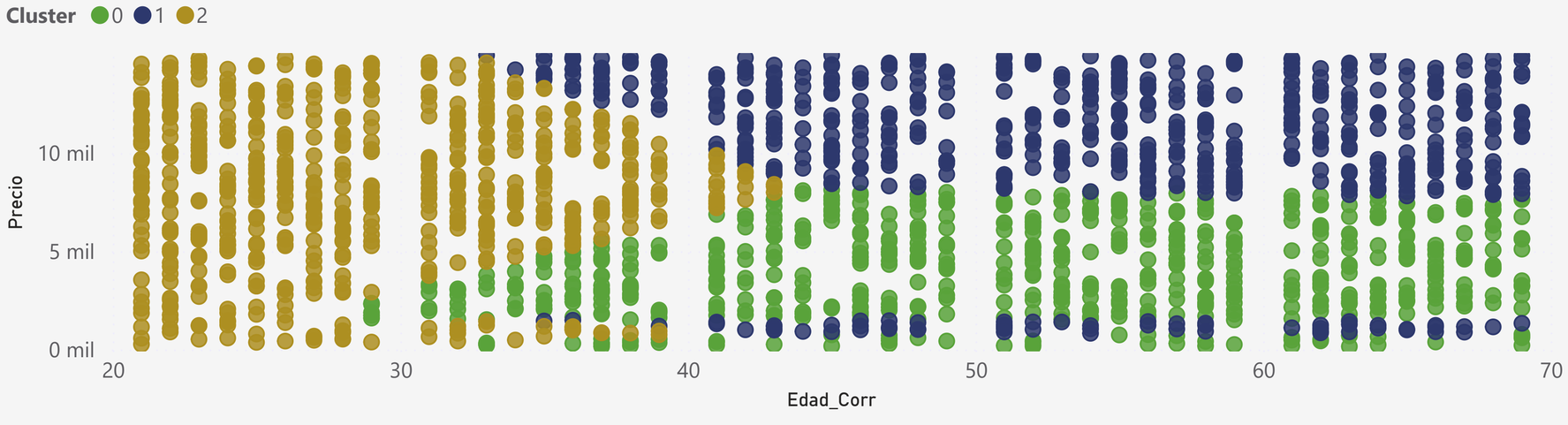
Distribución de Recomendación por Categoría de Producto



Participación de Clientes por Estatus de Recomendación



Segmentación de Clientes por Cluster: Edad vs. Nivel de Gasto



Conclusiones y Recomendaciones

Conclusiones

- **Salud de Marca**: Se registra un NPS proyectado del 69,03%, lo que posiciona a la mayoría de la base como promotores.
- **Segmento de Valor**: El Cluster 1 (Azul) agrupa a los clientes de mayor edad (40-70 años) con los tickets de compra más altos, siendo el segmento más rentable.
- **Volumen Operativo**: La categoría de Accesorios lidera en cantidad de transacciones, representando el mayor flujo de datos del sistema.
- **Nivel de Satisfacción**: La calificación global de 3,59 indica un margen de mejora operativa, especialmente en el 30,97% de clientes que no recomiendan el servicio.

Recomendaciones

- **Enfoque en Rentabilidad**: Priorizar campañas de fidelización para el Cluster 1, dado su historial de compras de alto valor.
- **Acción Correctiva**: Investigar los factores de insatisfacción en el 31% de detractores para elevar la Satisfacción Global por encima de 4,0.
- **Estrategia de Crecimiento**: Implementar tácticas de *cross-selling* en la categoría de Pantalones, que actualmente muestra el volumen más bajo de clientes.
- **Estabilización de Precios**: Desarrollar incentivos para que el Cluster 0 (Verde) incremente su gasto por encima del promedio actual de 7.530.

Gracias!