Informe: Análisis de Segmentación de Clientes y Predicción de Gasto

1. Introducción

Este informe detalla el proceso de análisis realizado para segmentar clientes y predecir su comportamiento de gasto. Se utilizaron dos enfoques principales:

- **Análisis No Supervisado (K-Means):** Para identificar grupos de clientes con características similares.
- **Modelado Supervisado (Red Neuronal Multicapa - MLP):** Para predecir si un cliente tendrá un gasto alto o bajo.

2. Preprocesamiento de Datos

- Manejo de valores faltantes.

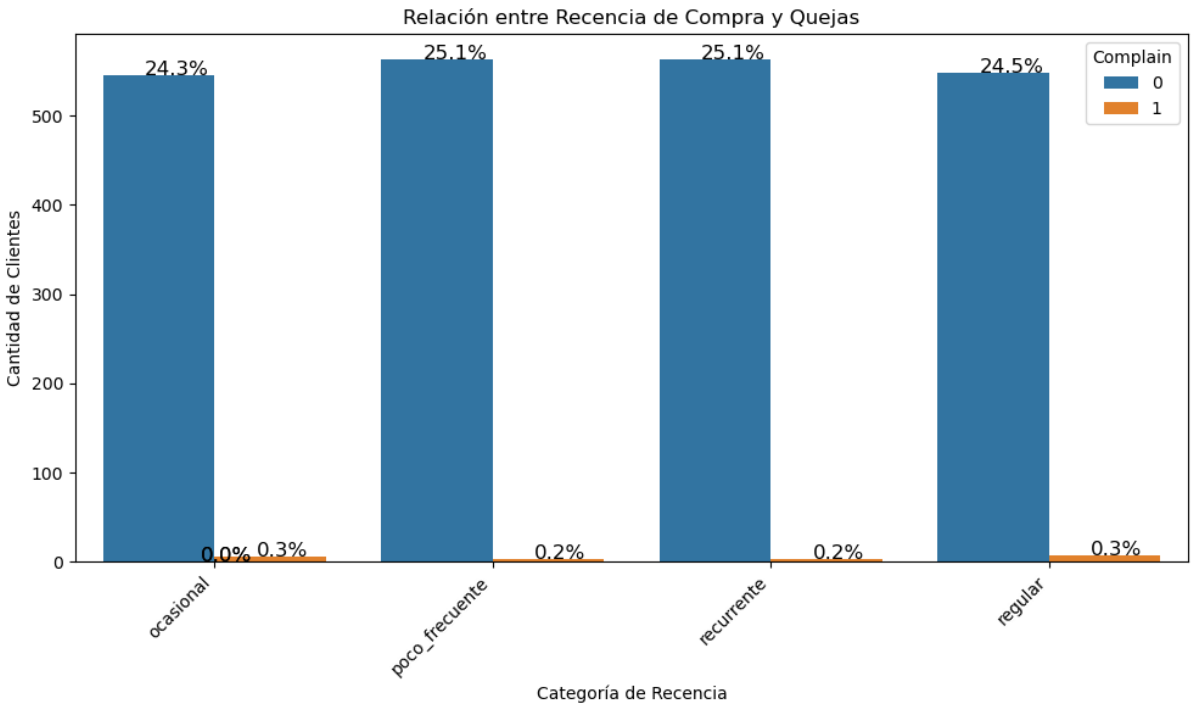
  Encontramos 24 valores nulos en la columna income que resolvi eliminar por reprentar poco mas del 1% de las filas

- Transformación de variables.

  Creacion la columna Recency_Category
  Función de categorización: Definimos una función categorize_recency que toma como entrada el número de días desde la última compra y devuelve una categoría basada en ese número:

  - 'Recent' para compras realizadas en los últimos 30 días.
  - 'Intermediate' para compras realizadas entre 31 y 180 días.
  - 'Old' para compras realizadas hace más de 180 días.

  Recent_Category con complain buscando relaciones con las quejas de los consumidores



Creación de las columnas

Number_Campaigns_Accepted:

- Esta columna probablemente almacena el número total de campañas que un cliente ha aceptado.
- Es un valor numérico que indica cuántas veces un cliente ha respondido positivamente a las campañas de marketing.
- Puede ser útil para segmentar a los clientes según su nivel de participación en las campañas.

Accepted:

- Esta columna podría ser un indicador binario (0 o 1) que muestra si un cliente ha aceptado una campaña específica.
- Un valor de 1 podría indicar que el cliente aceptó la campaña, mientras que un valor de 0 indicaría que no la aceptó.
- Es útil para análisis específicos de campañas individuales y para calcular tasas de aceptación.

Estas columnas son esenciales para entender el comportamiento de los clientes frente a las campañas de marketing y para tomar decisiones basadas en datos sobre futuras estrategias de marketing.

Claro, aquí tienes un resumen de las nuevas variables creadas y las variables modificadas (renombradas) en el código proporcionado:

Nuevas Variables Creadas

1. **Spending**: Total del gasto sumando varias categorías de productos.
2. **Children**: Cantidad total de hijos sumando niños y adolescentes.
3. **Has_child**: Indicador de si tiene hijos o no.
4. **Age**: Edad calculada a partir del año de nacimiento.

Variables Modificadas (Renombradas)

1. **Marital_Status**: Renombrado para diferenciar si están solos o en pareja.
2. **MntWines** a **Wines**:
3. **MntFruits** a **Fruits**:
4. **MntMeatProducts** a **Meat**:
5. **MntFishProducts** a **Fish**:
6. **MntSweetProducts** a **Sweets**:
7. **MntGoldProds** a **Gold**:
8. **Education**: Renombrado para agrupar niveles educativos.
9. **NumWebPurchases** a **Web**:

10. **NumCatalogPurchases** a **Catalog**:
11. **NumStorePurchases** a **Store**:

- Eliminamos columnas innecesarias.
- Eliminamos outliers basados en desviaciones estándar.
- Elimina filas con valores nulos.
- Creamos una copia del DataFrame limpio.

3. Análisis No Supervisado (K-Means)

- **Características Utilizadas:** Income, Age, Spending, NumDealsPurchases, NumWebVisitsMonth.
- **Número de Clusters (k):** 4.
- **Descripción de los Clusters:**
    - **Cluster 0:** Ingresos muy altos, edad media, gasto alto, pocas compras con descuento, pocas visitas web. Clientes premium.

      Grupo 0: "Conocedores adinerados"

      - Ingresos más altos ($76,500)

      - Gastadores de alto valor ($1,334 promedio)

      - Grandes consumidores de vino

      - Principalmente parejas sin hijos

      - Visitas web menos frecuentes
    - **Cluster 1:** Ingresos bajos, jóvenes, gasto muy bajo, pocas compras con descuento, muchas visitas web. Clientes jóvenes explorando o con bajo presupuesto.

      Grupo 1: "Jóvenes conscientes del presupuesto"

      - Grupo más joven (37 años)

      - Ingresos más bajos ($32,200)

      - Gasto mínimo ($114 promedio)

      - Tiene hijos

      - Visitantes frecuentes en línea

    - **Cluster 2:** Ingresos medios, edad media, gasto medio, muchas compras con descuento, muchas visitas web. Cazadores de ofertas.

Grupo 2: "Cazadores de ofertas"**

- Ingresos medios ($53,600)

- Mayor número de ofertas compradas

- Gasto moderado a alto

- Visitantes web más frecuentes

- Tiene hijos

- **Cluster 3:** Ingresos medios, mayores, gasto bajo, pocas compras con descuento, visitas web moderadas. Clientes mayores con hábitos de consumo más tradicionales.

  Grupo 3: "Gastadores moderados mayores"

  - Grupo de mayor edad (56 años)

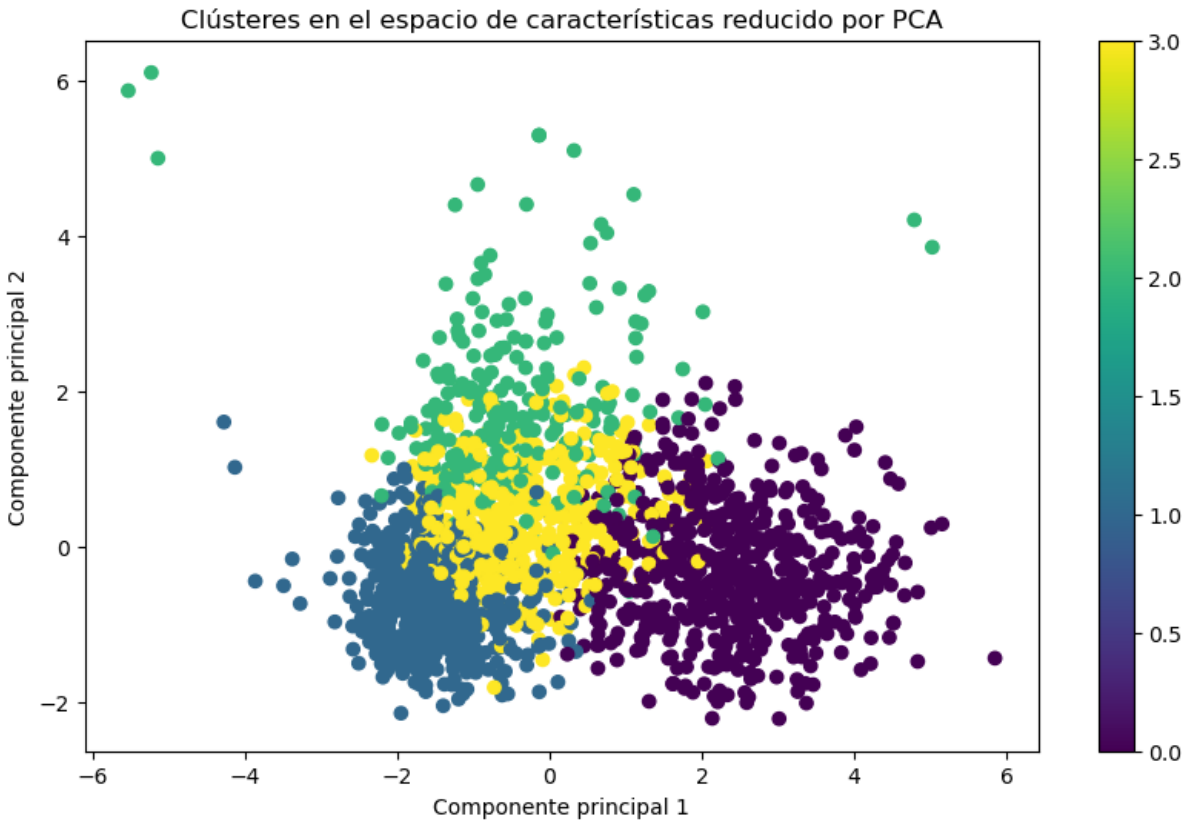
  - Ingresos y gastos moderados

  - Tiene hijos

  - Interacción web promedio

- **Puntuación de Silueta:** 0.28. Este valor sugiere cierta superposición entre los clusters.

- **Visualización:**



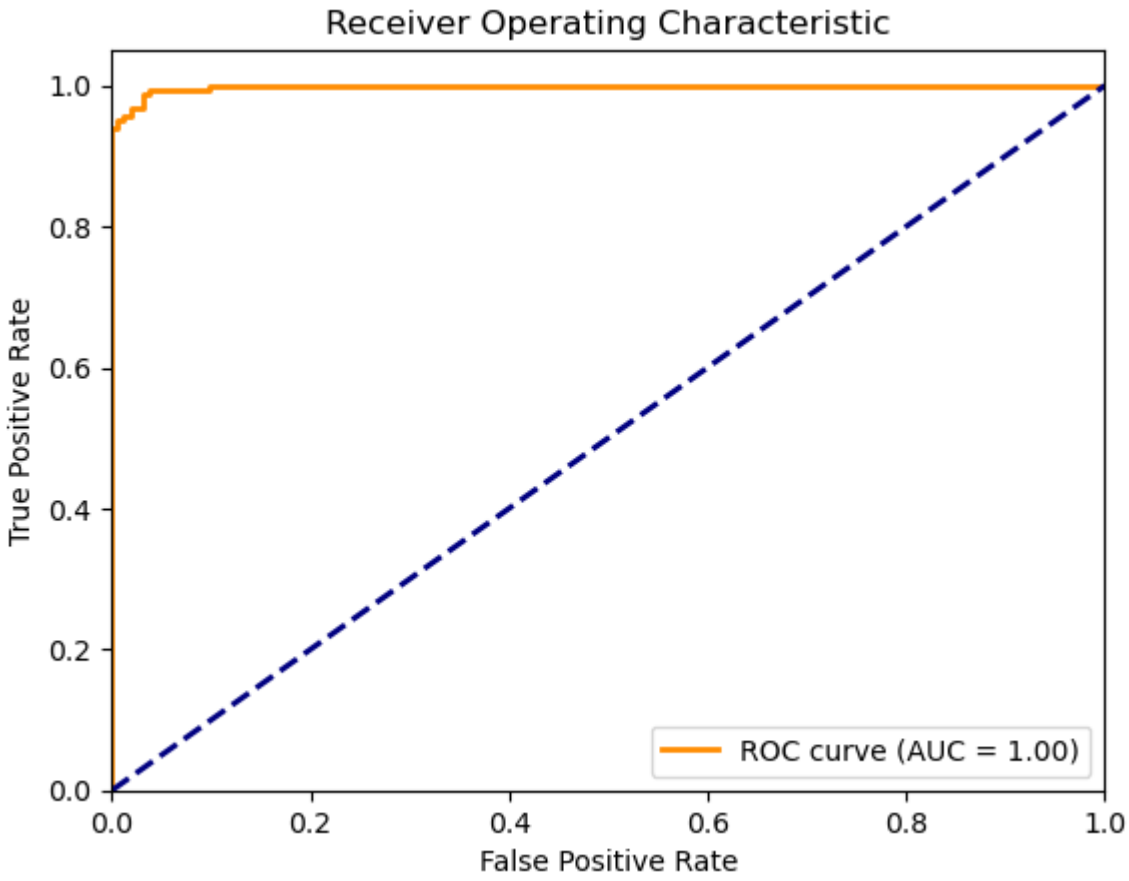
- **Número de componentes para el 90% de la varianza (PCA):** 4.
- **Conclusión del K-Means:** Se identificaron 4 segmentos de clientes con características diferenciadas. Sin embargo, la baja puntuación de silueta indica la necesidad de explorar otras configuraciones o métodos de clustering.

4. Modelado Supervisado (MLP)

- **Características Utilizadas:** Income, Age, NumDealsPurchases, NumWebVisitsMonth, Wines, Meat, Fish.
- **Arquitectura del Modelo:**
    - Capa de entrada con 7.
    - Capa oculta 1: 64 neuronas, activación ReLU, regularización L2 (0.01), Dropout (0.2).
    - Capa oculta 2: 32 neuronas, activación ReLU, regularización L2 (0.01), Dropout (0.2).
    - Capa de salida: 1 neurona, activación Sigmoide.
- **Optimizador:** Adam con learning rate de 0.002.
- **Función de Pérdida:** Binary Crossentropy.
- **Métricas:** Accuracy, Precision, Recall, F1-score, AUC.
- **Regularización:** L2, Dropout y Early Stopping (patience=3, restore_best_weights=True).
- **Batch Size:** 32.

- **Resultados en Validación (Epoch 27 - Mejor Epoch):**
    - Validation Loss: 0.1050
    - Validation Accuracy: 0.9785
    - Validation Precision: 0.9641
    - Validation Recall: 0.9938
    - Validation F1-score: 0.9787
    - Validation AUC: 0.9982
- **Resultados en Test:**
    - Test Accuracy: 97.78%
    - Classification Report (Test):
        - Precision: 0.98 (clase 0), 0.98 (clase 1)
        - Recall: 0.98 (clase 0), 0.98 (clase 1)
        - F1-score: 0.98 (clase 0), 0.98 (clase 1)

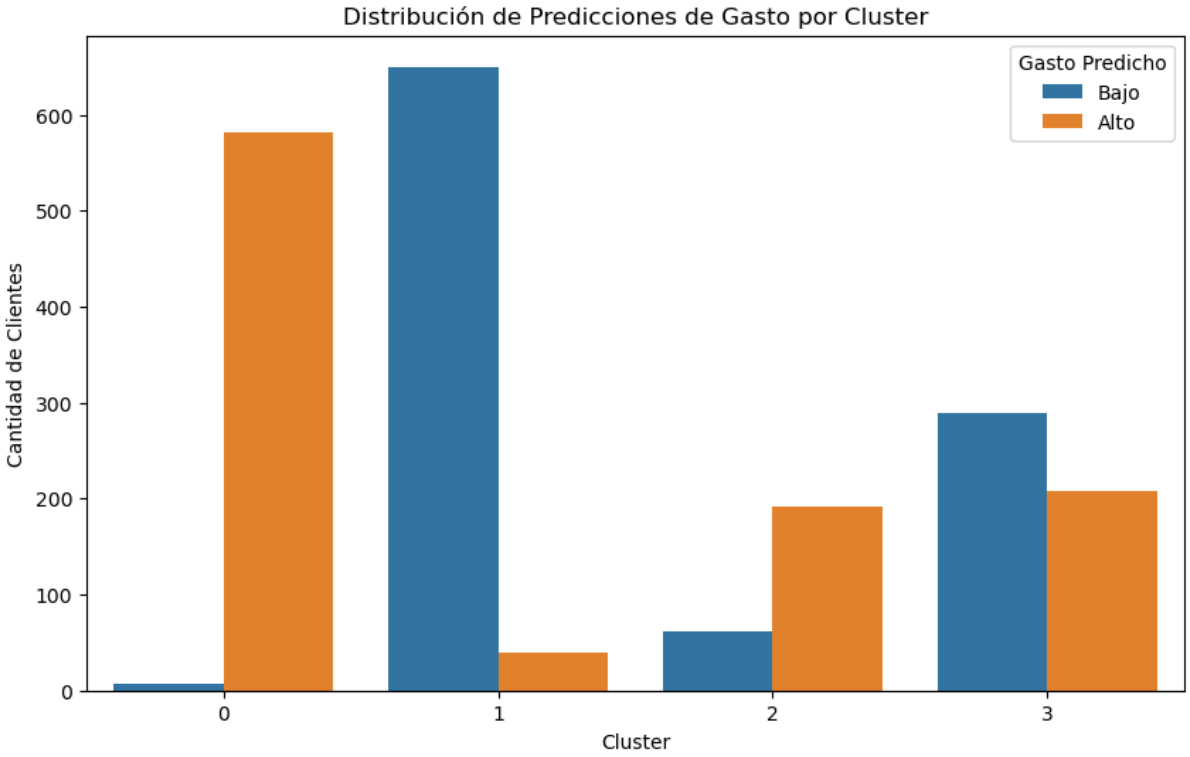
- **Curva ROC:**



5.

Relación entre K-Means y MLP

- **Distribución de Predicciones por Cluster:**



- **Análisis:**
    - Cluster 0: Mayoría predicha como "Alto Gasto".
    - Cluster 1: Mayoría predicha como "Bajo Gasto".
    - Cluster 2: Mayor proporción predicha como "Alto Gasto".
    - Cluster 3: Mayor proporción predicha como "Bajo Gasto", pero con una proporción significativa de "Alto Gasto".
- **Conclusión:** Existe una correlación entre los clusters identificados por K-Means y las predicciones del MLP. Sin embargo, los clusters 2 y 3 muestran una mayor mezcla, sugiriendo la necesidad de información adicional para mejorar la predicción en estos grupos.

6. Conclusiones Finales

- El modelo MLP muestra un buen rendimiento en la predicción del comportamiento de gasto, con métricas similares en los conjuntos de validación y test, lo que indica una buena generalización y ausencia de sobreajuste significativo.
- El análisis de K-Means identificó 4 segmentos de clientes con características distintas, que se correlacionan con las predicciones del MLP.
- Se recomienda explorar la incorporación de la pertenencia al cluster como una característica en el MLP y la adición de otras variables para mejorar la precisión de las predicciones, especialmente en los clusters 2 y 3.

7. Trabajo Futuro

- Incorporar la pertenencia al cluster como característica en el MLP.

- Explorar otras variables predictivas.
- Ajustar los hiperparámetros del MLP.
- Probar diferentes métodos de clustering o configuraciones del K-Means.
- Analizar la estabilidad de los clusters a lo largo del tiempo.