Informe de hallazgos en el proyecto

SEGMENTACIÒN DE CLIENTES

Lisy Velasco

ÍNDICE

1.INTRODUCCIÓN

En este proyecto, se trabajó con datos demográficos de clientes proporcionados por una empresa de telecomunicaciones, con el objetivo de desarrollar un modelo de clasificación que prediga la categoría de servicio que más se ajusta a cada cliente.

El EDA se centró en identificar relaciones entre las variables, detectar datos atípicos y explorar la distribución de las clases objetivo. Este análisis no solo sirvió como base para tomar decisiones informadas sobre la preparación de los datos y la ingeniería de características, sino que también destacó la importancia de trabajar con datos limpios y bien estructurados para obtener modelos efectivos.

En este informe, se detalla cada paso del EDA, incluyendo visualizaciones clave y hallazgos relevantes. Además, se analizan los desafíos encontrados, como posibles problemas de desequilibrio en las clases, y cómo estos factores influenciaron las decisiones posteriores en el proceso de modelado.

2.DATOS RECOLECTADOS

En este proyecto se utilizó el conjunto de datos **TeleCust1000.csv**, proporcionado por la empresa de telecomunicaciones. Este dataset incluye información demográfica y patrones de uso de servicios de 1,000 clientes, segmentados en cuatro categorías principales: *Basic Service, E-Service, Plus Service y Total Service.*

El dataset cuenta con un total de 11 columnas: 10 características predictoras y 1 variable objetivo (custcat), que clasifica a cada cliente en una de las cuatro categorías mencionadas. Las variables son una combinación de datos numéricos, categóricos y ordinales.

La estructura del dataset es la siguiente:

- Número de registros (filas): 1,000
- Número de variables (columnas): 11
- Tipos de datos:
  - Categóricos: region, marital, ed, gender, custcat (ordinal o nominal).
  - Continuos: income.
  - Discretos: tenure, age, address, employ, reside.
  - Binarios: marital, gender, retire.

3.DICCIONARIO DE DATOS

region

- Descripción: Indica regiones con valores entre 1 y 3.
- Media: 2.022 (cercano a la categoría 2).
- Desviación estándar: 0.8162 (dispersión moderada).
- Rango: De 1 (mínimo) a 3 (máximo).

tenure

- Descripción: Tiempo (en meses) asociado a un período de permanencia.
- Media: 35.53 meses.
- Rango: De 1 (mínimo) a 72 (máximo).

age

- Descripción: Edad de las personas.
- Media: 41.68 años.
- Rango: De 18 años (mínimo) a 77 años (máximo).

marital

- Descripción: Variable binaria para estado civil (0 = soltero, 1 = casado).
- Media: 0.495 (aproximadamente la mitad están casados).
- Rango: De 0 a 1.
- Mediana: 0.

address

- Descripción: Tiempo (en años) viviendo en la dirección actual.
- Media: 11.55 años.
- Rango: De 0 (recién llegados) a 55 años.

income

- Descripción: Ingresos anuales del cliente, en miles de unidades monetarias.
- Media: 77.54.
- Desviación estándar: 107.04 (alta variabilidad).
- Rango: Máximo de 1668.

ed (educación)

- Descripción: Niveles de educación en escala ordinal (1 a 5).
    a. Sin estudios.
    b. Estudios básicos.
    c. Educación secundaria/bachillerato.
    d. Educación superior.
    e. Posgrados.
- Media: 2.67 (niveles intermedios).
- Rango: De 1 a 5.

DICCIONARIO DE DATOS

employ

- Descripción: Años de empleo o experiencia laboral.
- Media: 10.99 años.
- Máximo: 47 años (experiencia extensa).

retire

- Descripción: Variable binaria (0 = no retirado, 1 = retirado).
- Media: 0.047 (pocas personas están retiradas, ~4.7%).
- Rango: De 0 a 1.

gender

- Descripción: Variable binaria para género (1 = masculino, 0 = femenino).
- Media: 0.517 (proporción equilibrada de géneros).

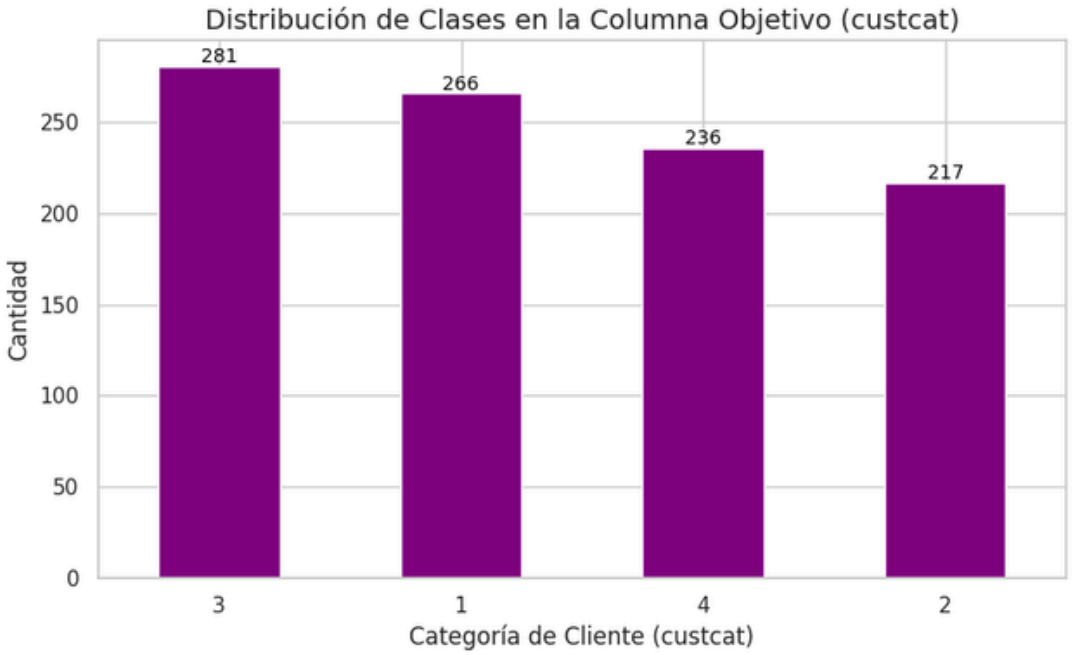
reside

- Descripción: Número de residentes por hogar.
- Media: 2.33 (promedio de 2 a 3 residentes).
- Máximo: 8 residentes.

cuscat

- Descripción: Variable objetivo, con categorías:
  - Basic Service.
  - E-Service.
  - Plus Service.
  - Total Service.

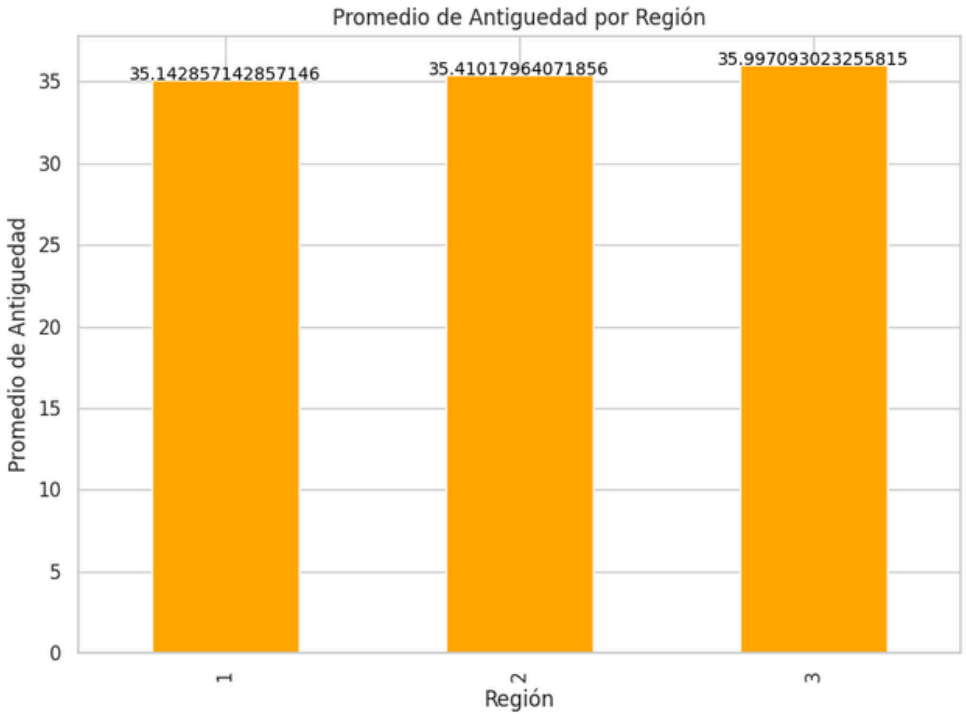
4.ANÀLISIS EXPLORATORIO



Gràfico de la columna Cuscat y la cantidad de clientes

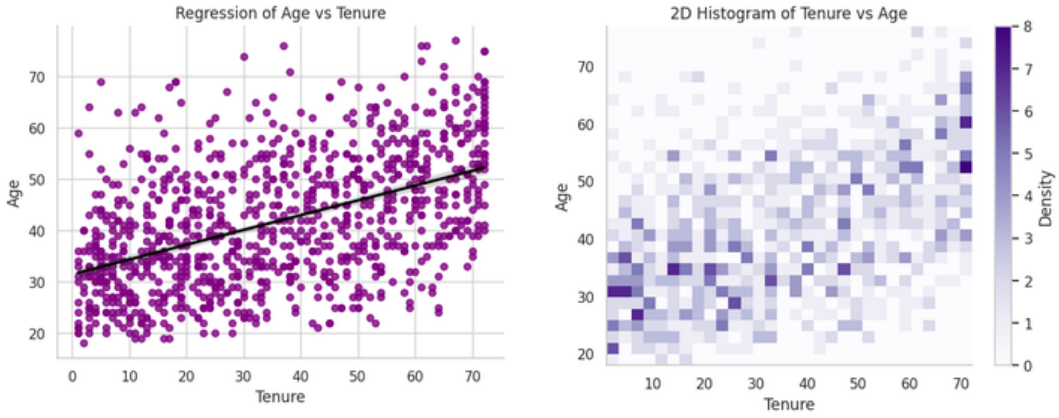
En el gráfico de barras anterior se puede observar que la categoría 2 tiene menos clientes en comparación con las demás, con un valor de 217.En cuanto al balance de clases no son exactamente iguales, pero estan razonablemente balanceadas.

Gráfico de Antigüedad por región

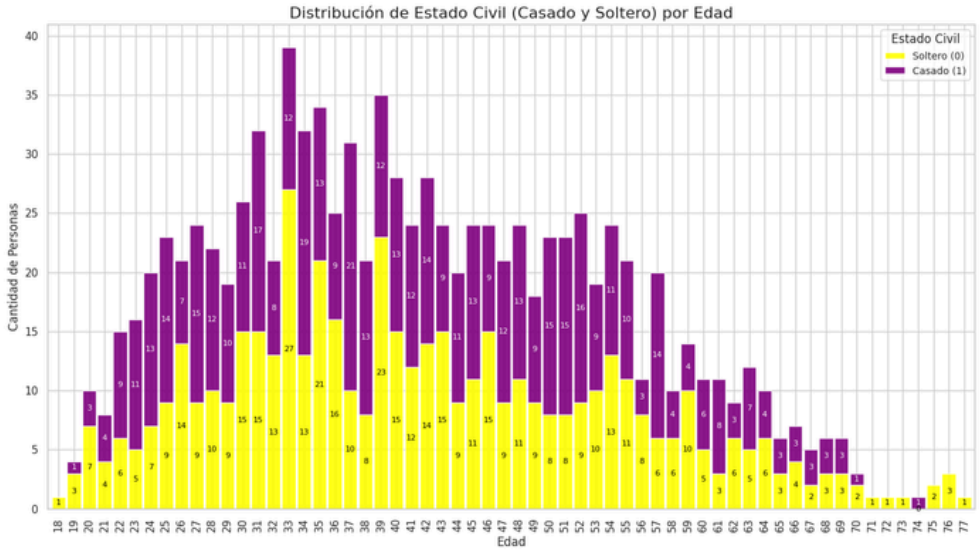


Al comparar los promedios de *tenure* entre las tres regiones, se puede observar que la región 3 tiene el valor más alto de antigüedad en meses, seguida de la región 2 y luego la región 1. Esto sugiere que los clientes de la región 3 tienen, en promedio, una mayor antigüedad o tiempo como clientes de la empresa.

Gráfico de Antigüedad vs Edad

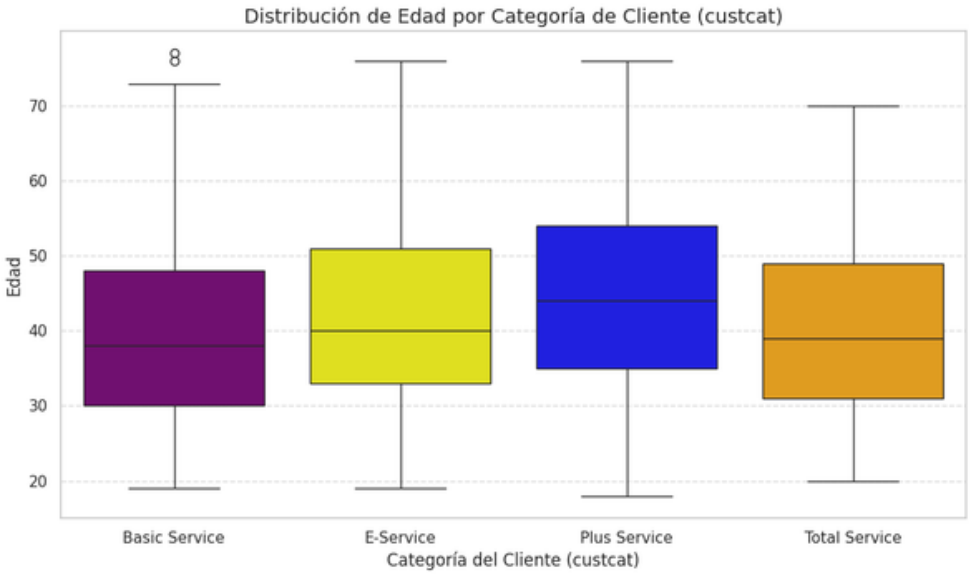


Estos gráficos, tanto de dispersión como el mapa de calor muestran la relación entre la "Antiguedad/tenure" y "Edad/age".
Hay una gran dispersión de puntos, lo que sugiere una relación compleja entre estas variables.
Aún así, parece haber una tendencia general ascendente, indicando que a mayor tenure o antiguedad, hay mayores valores en la edad.
Hay algunos puntos outliers que se desvían significativamente de la tendencia general, lo cual puede deberse a factores adicionales que influyen en la métrica.



Gràfico de Edad vs Estado civil

La altura total de cada barra muestra la cantidad total de personas en esa edad. Los números indican las cantidades específicas de solteros y casados dentro de cada grupo de edad.
Edad Temprana (18-30 años):
Predominan los solteros (amarillo), especialmente en edades más jóvenes.
En edades cercanas a los 30 años, comienza a observarse un aumento en la proporción de personas casadas.
Edad Media (30-50 años):
La proporción de casados aumenta significativamente en esta etapa.
Las barras son más altas, lo que indica que estas edades concentran la mayor cantidad de personas en el dataset.
Edad Avanzada (50 años en adelante):
La cantidad de personas disminuye en general a medida que la edad aumenta.
Se observa una mayor proporción de casados en estas edades.

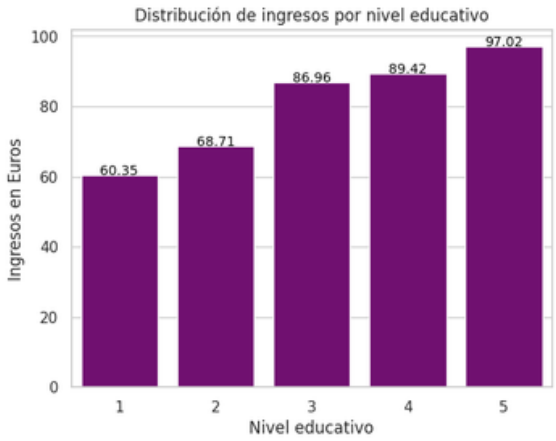


Gràfico de Distribución de Edad por Categoría de Cliente

- **Plus Service** parece atraer a clientes ligeramente mayores.
- **E-Service** tiene a atraer medianamente clientes de mas de 45 años
- **Basic Service** y **Total Service** tienen una población de clientes más variada en términos de edad.

El rango completo de edades muestra que todas las categorías tienen una clientela diversa.

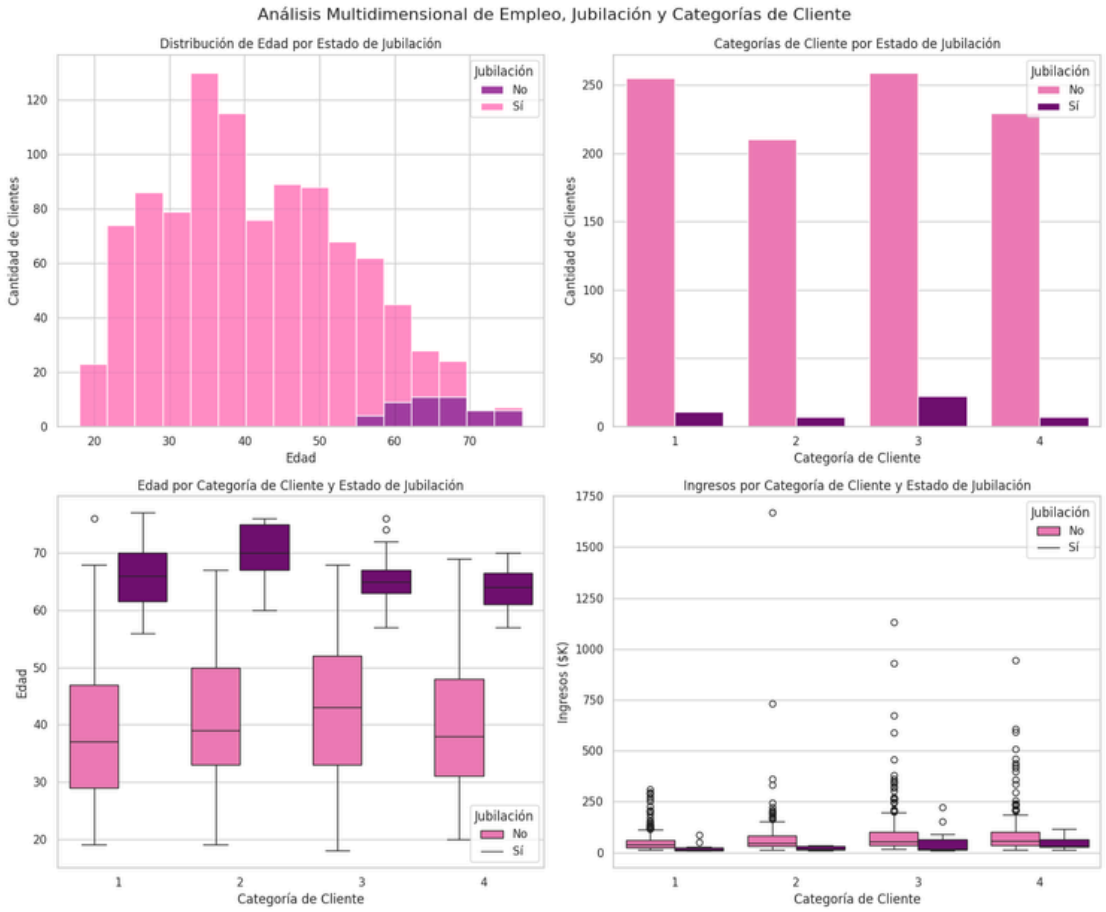
Gráfico de Ingresos y Educación (income-ed)



Hay una clara correlación positiva entre nivel educativo e ingresos
El mayor salto en ingresos se da entre el nivel 2 y 3
La diferencia entre el nivel más bajo y el más alto es de aproximadamente 36,670.

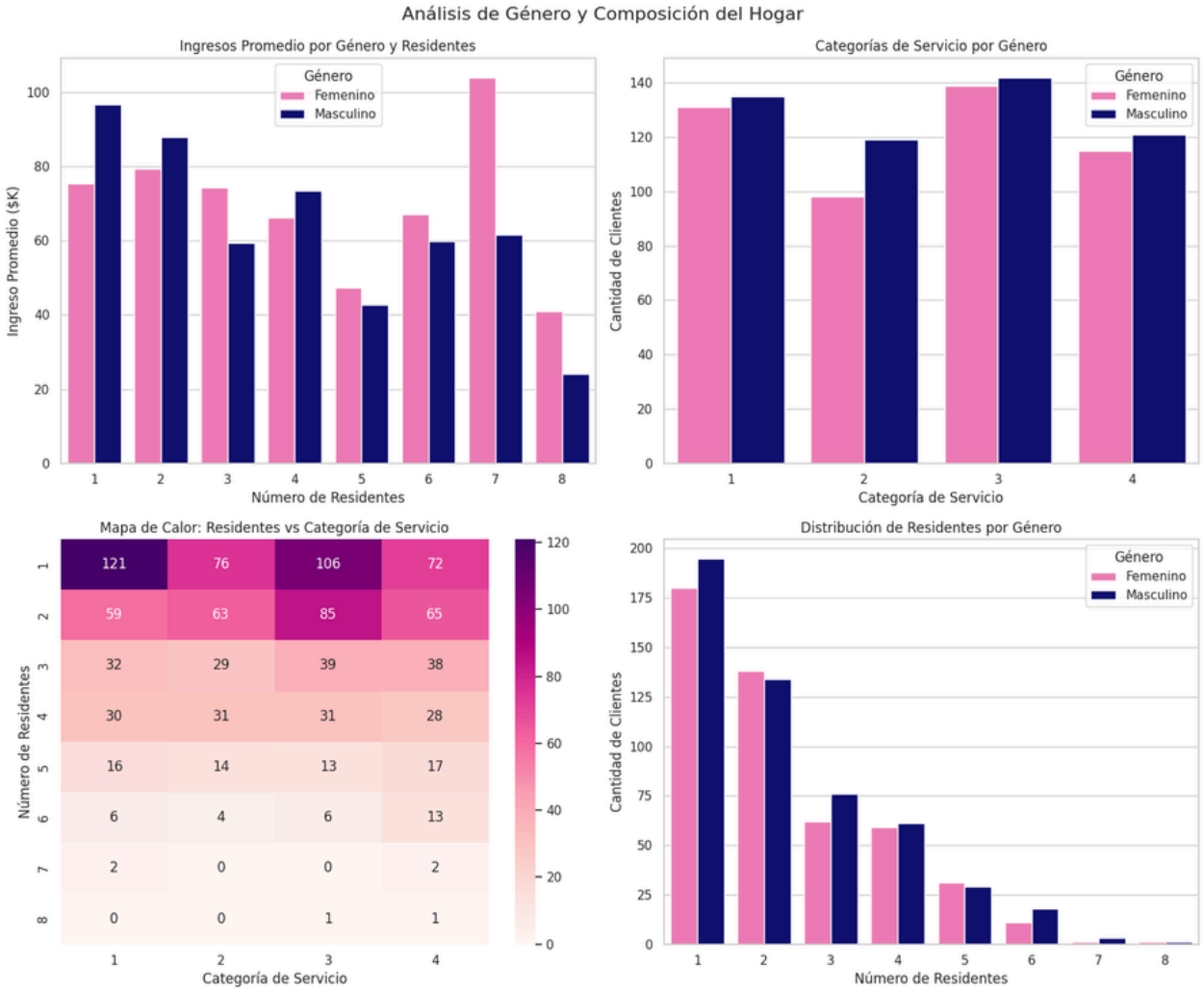
Los incrementos son más pronunciados en los niveles iniciales y se van suavizando en los niveles superiores

Gràfico de Años de Empleo y Estado de Jubilación (employ-retire)



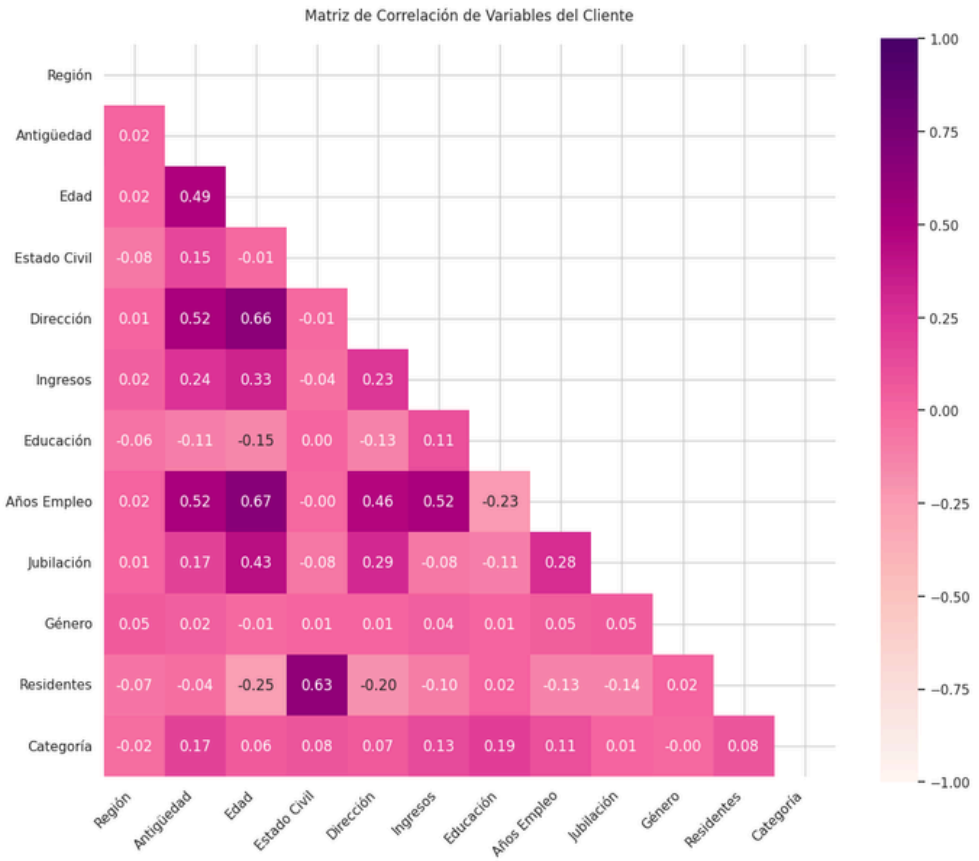
Con base en los datos disponibles, se pueden identificar características importantes donde la mayoría de **los clientes son personas activas que trabajan y tienen entre 30 y 50 años**. Los jubilados, que suelen superar los 60 años, son una minoría, en cuanto a consumo y preferencias **las categorías 1 y 3 son las que más llaman la atención de los clientes en general**, los trabajadores activos están presentes en casi todas las categorías, pero **los jubilados prefieren claramente la categoría 3**. Esto indica que debemos prestar especial atención a lo que esta categoría ofrece para los jubilados. Los clientes tienen ingresos muy diferentes entre sí, sin importar la categoría. **Donde los jubilados, no siempre ganan menos que los activos.** Hay clientes con ingresos especialmente altos, tanto activos como jubilados. Esto sugiere que es importante no asumir que todos los jubilados tienen ingresos bajos. **La categoría 4** tiene mayor variedad de ingresos. Para **la categoría 2** se puede inferir que predominan clientes activos, tiene la menor proporción de clientes jubilados y muestra patrones de ingreso más estables.

Grafico Género y Residentes del Hogar (gender-reside)



En hogares pequeños (1-2 residentes), los hombres tienden a tener ingresos más altos. Sin embargo, esta dinámica cambia en hogares grandes (7 residentes), donde las mujeres destacan con ingresos más altos. La categoría 3 es la favorita tanto para hombres como para mujeres, lo que sugiere que tiene una oferta muy atractiva en general. Aunque hay una leve preferencia masculina en las otras categorías, la categoría 2 se distingue por ser la más balanceada entre géneros.En la matriz de correlacion las variables muestran correlaciones débiles entre sí, indicando que son bastante independientes.

Matriz de correlación



Correlaciones relacionadas con la Edad:
Edad - Años Empleo (0.670): La correlación más fuerte. Natural, ya que a mayor edad, más años de experiencia laboral
Edad - Dirección (0.660): Sugiere que personas mayores tienden a tener direcciones más estables/permanentes
Edad - Antigüedad (0.490): Clientes mayores tienden a tener más tiempo con la empresa
Edad - Jubilación (0.429): Correlación positiva lógica, a mayor edad más probabilidad de jubilación
Edad - Ingresos (0.328): Correlación moderada que sugiere que los ingresos aumentan con la edad

Correlaciones relacionadas con Antigüedad:
Antigüedad - Dirección (0.523): Clientes con más tiempo en la empresa tienden a tener direcciones más estables
Antigüedad - Años Empleo (0.520): Mayor estabilidad laboral se relaciona con mayor tiempo como cliente
Antigüedad - Edad (0.490): Clientes más antiguos tienden a ser de mayor edad

Correlaciones relacionadas con Años de Empleo:

Años Empleo - Ingresos (0.516): Mayor experiencia laboral se relaciona con mayores ingresos
Años Empleo - Antigüedad (0.520): Más años trabajando se relaciona con más tiempo como cliente
Años Empleo - Dirección (0.463): Estabilidad laboral se relaciona con estabilidad residencial

Correlaciones relacionadas con Estado Civil y Residentes:

Estado Civil - Residentes (0.626): Fuerte correlación que indica que las personas casadas tienden a tener más residentes en el hogar.

5.CONCLUSIONES

El análisis exploratorio de datos (EDA) reveló aspectos clave sobre la calidad y el balance de los datos en el conjunto TeleCust1000.csv. Si bien los datos proporcionaron una base sólida para desarrollar modelos predictivos, se identificaron desafíos importantes, como el desbalanceo entre las clases de la variable objetivo (custcat), que requería estrategias específicas para garantizar que el modelo no favoreciera categorías mayoritarias en detrimento de las minoritarias.

Además, el análisis mostró que la calidad de los datos era aceptable en general, con pocos valores faltantes y una representación clara de las características demográficas y de uso. No obstante, el éxito del modelo dependió en gran medida de la adecuada transformación y normalización de los datos, así como de una cuidadosa selección de características relevantes.
En resumen, este proyecto resalta una lección fundamental en ciencia de datos: la calidad y el balance de los datos son pilares críticos para construir modelos efectivos y generalizables. Aunque la tecnología y las técnicas avanzadas de modelado juegan un papel importante, es el tratamiento cuidadoso de los datos lo que determina en última instancia el éxito de cualquier proyecto de aprendizaje automático.

Lisy Velasco