PROYECTO FINAL - CIENCIA DE DATOS

Por: Zahira Chaia

1.

Abstract

2.

Preguntas de interés

3.

Insights

4.

Modelado

5.

Conclusiónes finales

1. Abstract - Problema

Una de las razónes fundamentales por las que las empresas fracasan es la falta de conocimiento sobre los gustos y necesidades de sus clientes. La personalización resultante no solo mejora la experiencia del cliente, sino que también contribuye al éxito sostenible de la empresa en el mercado.

1. Abstract - Solución

Es por esta razón, que el propósito central de este estudio consiste en desarrollar diversos modelos estadisticos que sean de utilidad para una mejorar las ventas y las campañas de Marketing de una empresa especializada en la distribución de una amplia gama de productos alimenticios a través de diversos canales de venta.

1. Abstract - Dataset

Esta base cuenta con 2240 registros, y 29 variables (categoricas y numericas).

A lo largo del dataset, se elimanaron varias variables que no fueron pertinentes.

2. Preguntas de intéres

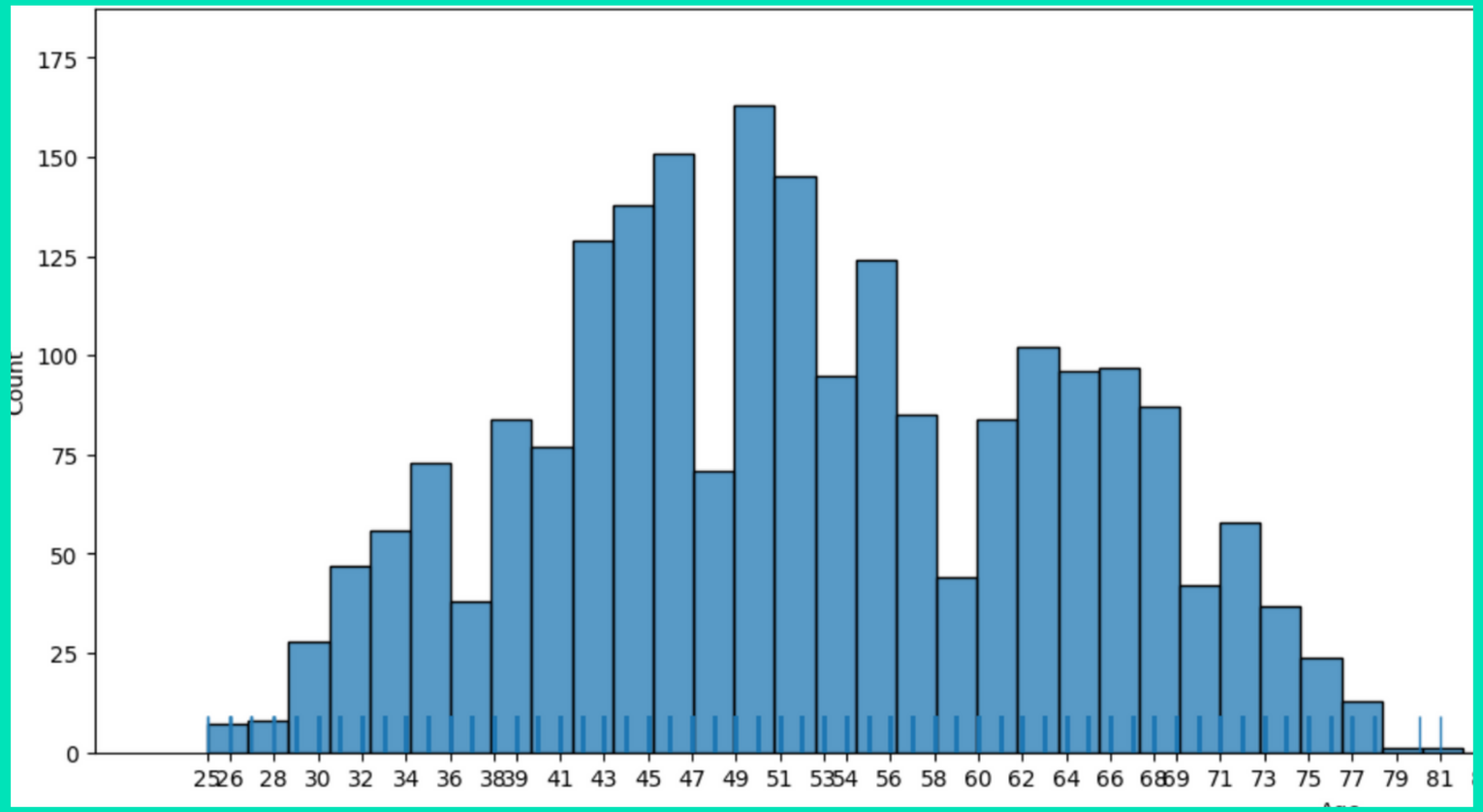
- 1- Predecir la probabilidad de ingreso de los consumidores basándose en sus patrones de consumo.

- 2- Indigar sobre la personalidad del consumidor, haciendo una segmentación de clientes.

3. Insights

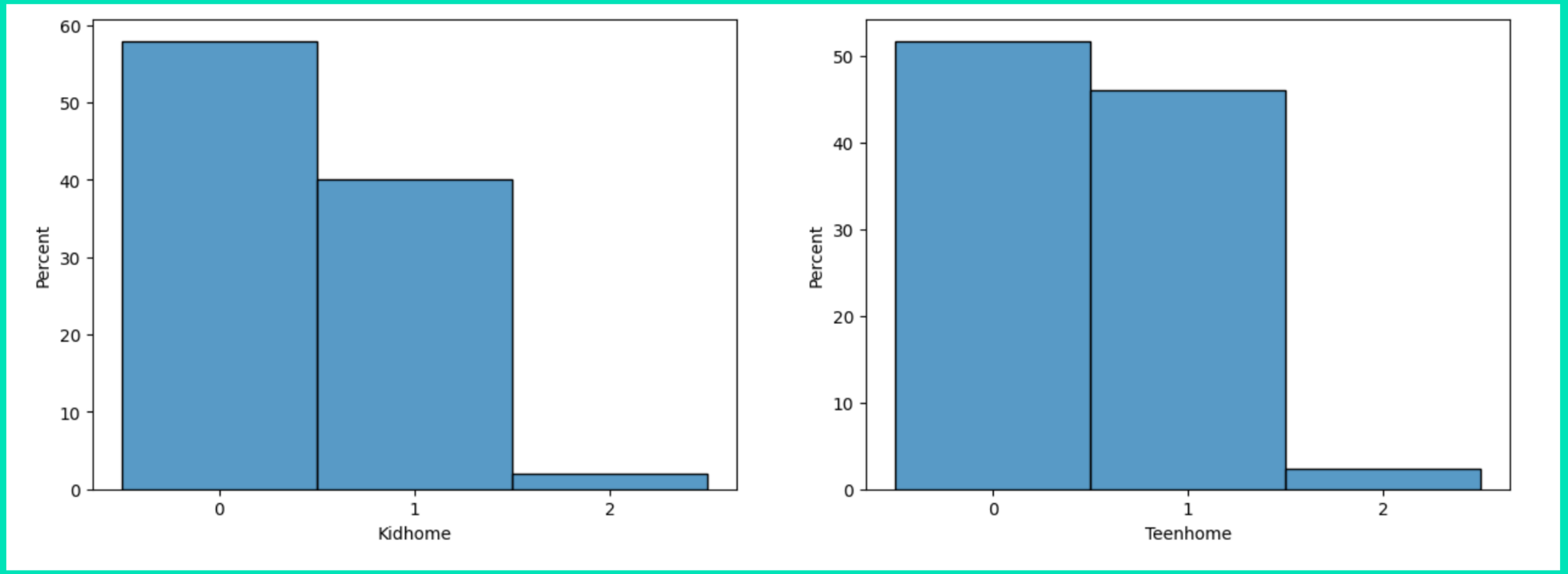
Antes de comenzar a indagar sobre las preguntas de intéres, observaremos algunos insigths sobre los clientes, para luego realizar estrategias de Marketing mas adecuadas de acuerdo a esta informacién.

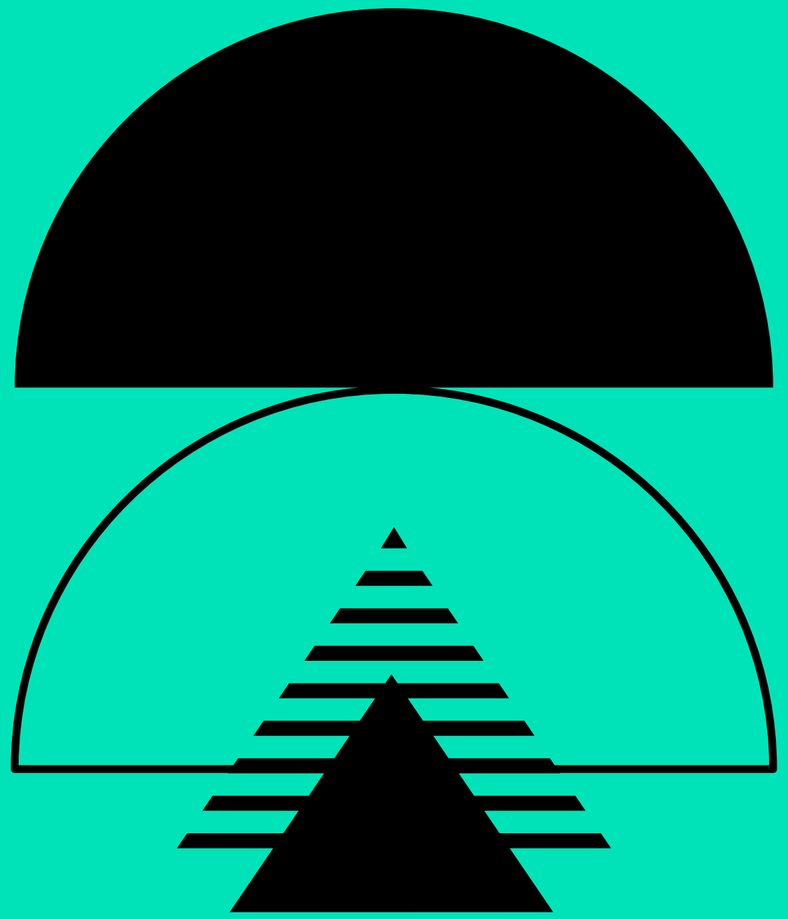
- La edad de los clientes se concentra principalmente entre los 40 y 60 años, los jóvenes (menores de 30 años) son muy pocos.



3. Insights

- La gran mayoría de los clientes tienen 1 o 0 hijos en casa.
- Muy pocos tienen 2 hijos y ninguno tiene hijos mayores de 2 años.

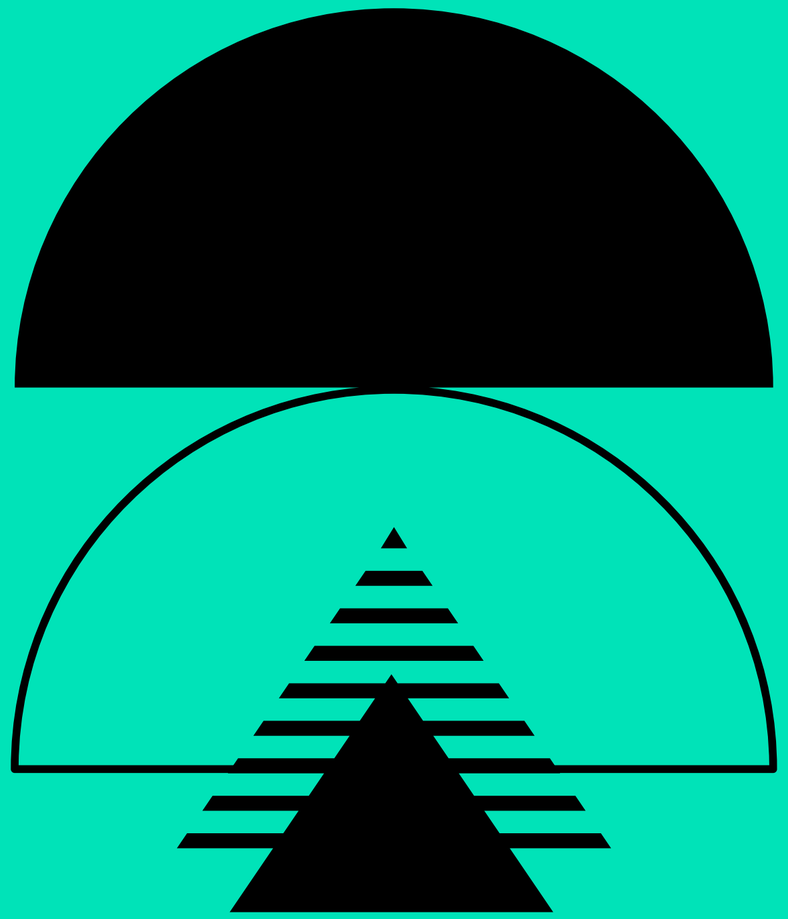




4. Modelado 1

Para responder a la pregunta 1 de investigación de si se puede estimar el ingreso de los consumidores basándose en sus patrones de consumo haremos dos principales modelos y luego compararemos sus métricas.

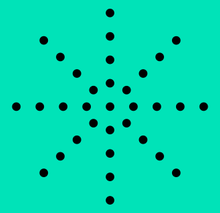
Estos modelos son algoritmos de regresión que se utilizan para predecir valores numéricos continuos, En este tipo de problemas, el objetivo es predecir un valor en función de variables de entrada.



4. Modelado 1

Previo al modelado, se realizó un filtraje de las 12 variables mas importantes relacionadas al consumo, mediante diferentes tecnicas como la matriz de correlación (que nos brinda información sobre cómo las variables están relacionadas entre sí) y el PCA (técnica utlizada para reducir la dimensionalidad de las variables).

4. Modelado 1



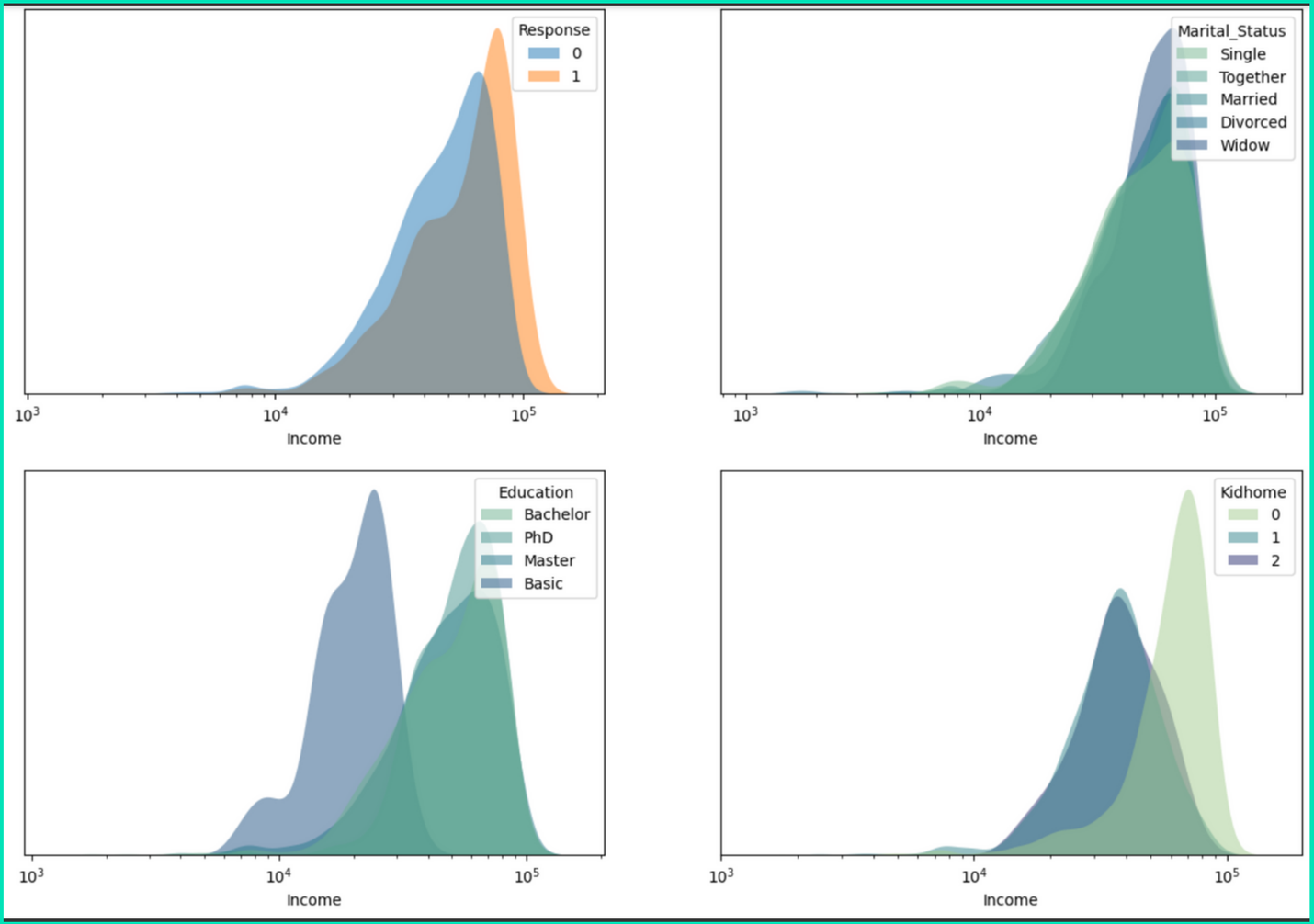
1- Modelo de Regresión lineal

El objetivo principal es predecir el valor de la variable dependiente (en este caso el Income, el ingreso de los consumidores) en función de los valores de las variables independientes (12 variables que quedaron seleccionadas anteriorimente relacionadas al consumo de los clientes).

2- Modelo de Regresión de Árboles de Decisión

Este algoritmo de regresión divide iterativamente el espacio de características en subconjuntos cada vez más pequeños, utilizando decisiones basadas en las características de los datos.

4. Insights que nos dejaron estos modelos



- Los individuos pertenecientes a grupos de altos ingresos muestran una mayor tendencia a aceptar la oferta de la campaña (aunque hay una leve diferencia)
- No se observa una distinción marcada en los ingresos según el estado civil de las personas.
- Se aprecia una disparidad significativa en los ingresos entre los clientes con educación básica en comparación con aquellos con títulos de licenciatura, maestría o doctorado, mientras que no se percibe una discrepancia evidente entre estos últimos grupos.
- Los clientes que no tienen hijos en casa tienden a tener niveles de ingresos superiores.

4. Comparando metricas

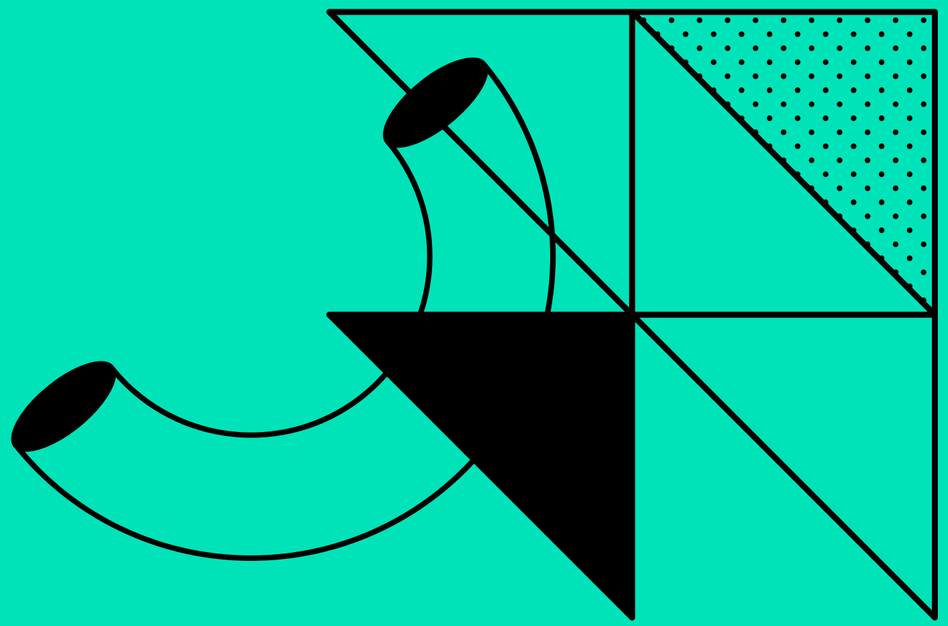
```
Métricas para conjunto de prueba de modelo de regresion:
Mean Absolute Error: 0.2087243249792556
Median Absolute Error: 0.1551735013392887 6
Max Error: 3.4306953452694415
Error Cuadrático Medio: 0.10308423048435741
R2: 0.6024785008106759
```

4. Comparando metricas

```
Métricas para conjunto de prueba (Modelo de regresion):
Mean Absolute Error: 0.2087243249792556
Median Absolute Error: 0.1551735013392887 6
Max Error: 3.4306953452694415
Error Cuadrático Medio: 0.10308423048435741
R2: 0.6024785008106759
```

```
Métricas para conjunto de prueba (Regresión de Árboles de Decisión):
Mean Absolute Error: 0.1705851598726519
Median Absolute Error: 0.1037921829062327 1
Max Error: 1.6262556973602589
Error Cuadrático Medio: 0.07374820810611873
R2: 0.7156063724672297
```

Métricas



01 MAE: el del Modelo 2 es menor que el del Modelo 1, lo que indica que el Modelo 2 tiene una menor discrepancia promedio entre las predicciones y los valores reales en el conjunto de prueba.

02 Median Absolute Error: El Modelo 2 también tiene una mediana del error absoluto más baja, lo que sugiere que el Modelo 2 es más consistente en términos de precisión.

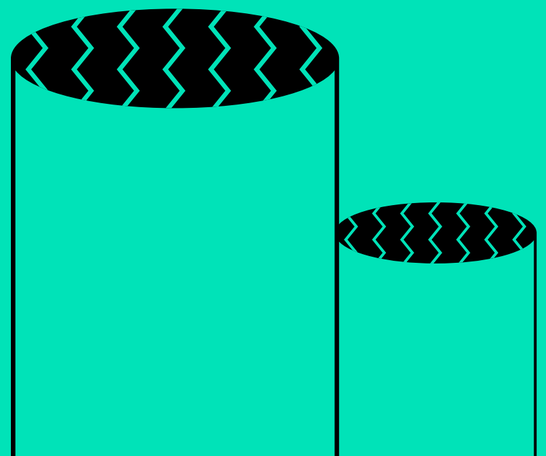
03 Max Error: El Modelo 2 tiene un error máximo significativamente menor que el Modelo 1, lo que indica que el Modelo 2 tiene menos valores atípicos extremos en sus predicciones.

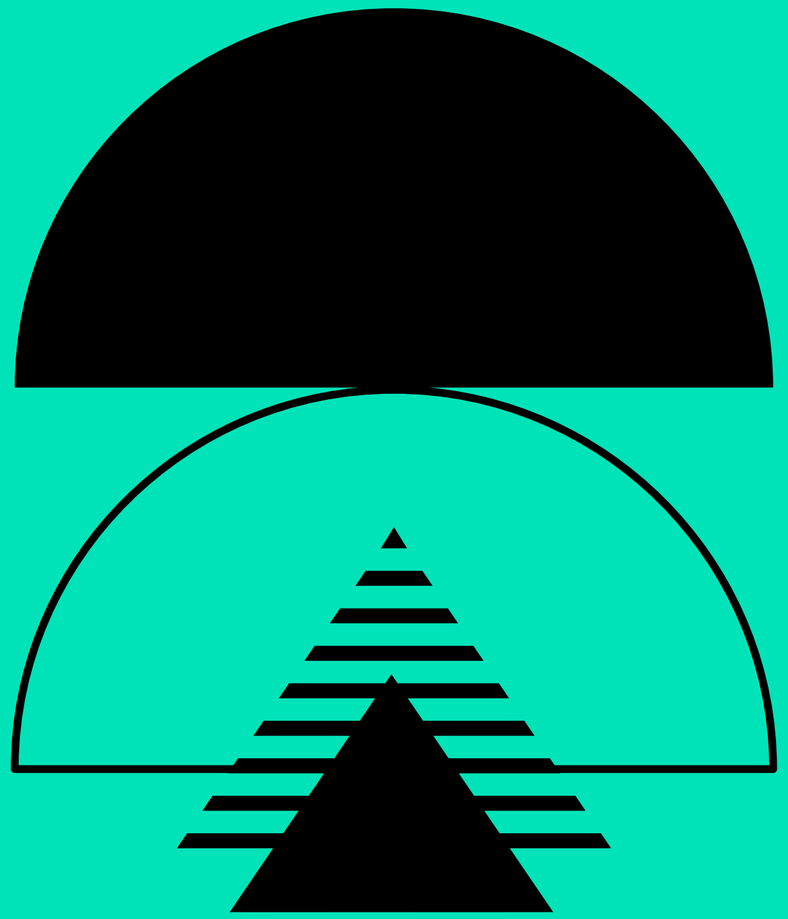
04 Error Cuadrático Medio: el del Modelo 2 es menor que el del Modelo 1, lo que indica que el Modelo 2 tiene una mejor precisión en general al predecir los valores de ingresos.

05 R2: El del Modelo 2 es más alto que el del Modelo 1, lo que sugiere que el Modelo 2 explica una mayor proporción de la variabilidad en los datos de ingresos.

4. Modelo 1 - Conclusiones

Luego de comparar las mismas métricas para ambos modelos, podemos concluir que el Modelo 2, que es la Regresión de Árboles de Decisión, parece ser mejor en términos de precisión y capacidad predictiva en comparación con el Modelo 1, que es la Regresión Lineal. Por lo tanto, podríamos preferir el Modelo 2 para predecir los ingresos de los consumidores.



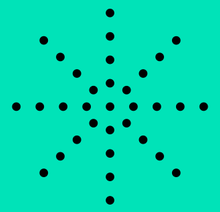


4. Modelado 2

Para indigar sobre la personalidad del consumidor, haciendo una segmentación de clientes, utilizaremos dos modelos para luego comparar sus metricas.

Estos modelos forman parte del aprendizaje no supervisado, en donde se entrena en datos que no tienen etiquetas o valores objetivo asociados. En lugar de predecir una variable de salida específica, el objetivo principal del aprendizaje no supervisado es descubrir patrones o estructuras subyacentes en los datos.

4. Modelado 2



3- Modelo de K-Nearest Neighbors (K-NN)

Este algortimo clasifica las muestras según la clase de sus vecinos más cercanos en el espacio de características

4- Modelo de clustering jerárquico aglomerativo

Es un método que agrupa datos basándose en la distancia entre ellos. Los elementos se organizan en una jerarquía de clústers con forma de árbol.

4. Insights que nos dejaron estos modelos

01

Ambos modelos lograron generar la misma cantidad de clusters: 2. Es decir, estos modelos segmentaron a los clientes en 2 grandes grupos.

02

En relacion al nivel de ingresos de los consumidores, se observa que en el cluster 1, los sujetos tienen ingresos mas altos que el cluster 2.

03

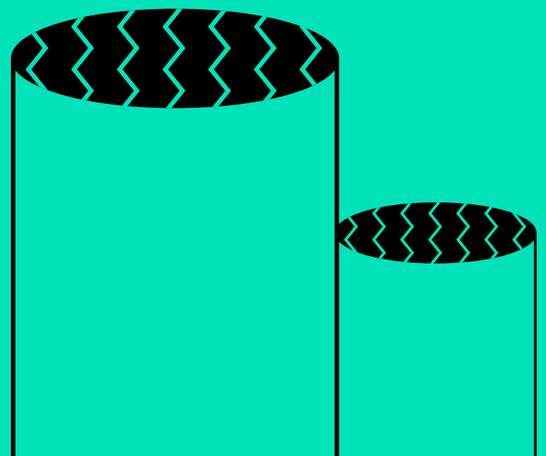
En relación a la cantidad de hios, se puede ser que en el cluster 1 la mayoria no tienen hijos. Y los que si tienen, suelen tener 1. Los del cluster 2, la mayoria tienen 1 hijo. Hay algunos que tienen 2.

04

Se da una gran cantidad de consumo en la mayoria de los prouctos por parte de los sujetos del cluster 1. Estos a la vez, optan por hacerlo en un medio particular: tiendas fisicas. A la vez, se observa que el grupo 2, consume menos. Por ende, se podria volver a afirmar que los de este grupo cuentan con un poder adquisitivo menor.

4. Modelo 2 - Conclusiones

Podemos observar que en ambos modelos, el coeficiente de silueta, una medida utilizada en el análisis de grupos (clustering) para evaluar la calidad de los clusters formados por diferentes algoritmos es muy similar. Esto sugiere que los objetos están relativamente bien emparejados con sus propios clusters y tienen una separación moderada con los clusters vecinos.



5. Conclusiones finales

- Luego de un analisis arduo sobre los datos, se pudo cumplir con el objetivo propuesto, es decir, encontrar modelos estadisticos eficientes para mejorar las ventas y las campañas de Marketing.

- Para la primera pregunta de investigación, que se centra en predecir los ingresos de los consumidores basándonos en sus patrones de consumo, hemos seleccionado el modelo 2 (Regresión de Árboles de Decisión). Esta elección se basa en el hecho de que sus métricas nos proporcionan resultados superiores.

- En cuanto a la segunda pregunta de investigación, que busca explorar la personalidad del consumidor mediante la segmentación de clientes, hemos observado que tanto el modelo de clustering jerárquico aglomerativo como el modelo K-Nearest Neighbors (K-NN) son igualmente efectivos para abordar esta cuestión.

- Además, es crucial considerar los diversos insights que hemos obtenido a lo largo de nuestro trabajo. Estos insights son de suma importancia, ya que proporcionan información sobre cómo se compone nuestra clientela. En otras palabras, nos brindan una visión más clara del problema comercial al que nos enfrentamos.