Preprocesamiento de datos

Info del dataset

El archivo estaba separado por tabs.
El dataset tiene 2240 filas y 29 columnas.

Diccionario de datos del dataset.

index	Columna	Tipo	Descripción	Ejemplo	Grupo
0	ID	int64	Identificador único del cliente	12345	People
1	Year_Birth	int64	Año de nacimiento del cliente	1985	People
2	Education	object	Nivel educativo del cliente	Graduado	People
3	Marital_Status	object	Estado civil del cliente	Casado	People
4	Income	float64	Ingreso anual del hogar del cliente	50000	People
5	Kidhome	int64	Número de niños en el hogar del cliente	1	People
6	Teenhome	int64	Número de adolescentes en el hogar del cliente	0	People
7	Dt_Customer	datetime64	Fecha de registro del cliente en la empresa	2019-01-01	People
8	Recency	int64	Días desde la última compra del cliente	10	People
9	Complain	int64	1 si el cliente se quejó en los últimos 2 años, 0 en caso contrario	0	People
10	MntWines	float64	Monto gastado en vino en los últimos 2 años	300	Products
11	MntFruits	float64	Monto gastado en frutas en los últimos 2 años	50	Products
12	MntMeatProducts	float64	Monto gastado en carne en los últimos 2 años	200	Products
13	MntFishProducts	float64	Monto gastado en pescado en los últimos 2 años	75	Products
14	MntSweetProducts	float64	Monto gastado en dulces en los últimos 2 años	30	Products
15	MntGoldProds	float64	Monto gastado en productos de oro en los últimos 2 años	20	Products
16	NumDealsPurchases	int64	Número de compras realizadas con descuento	5	Promotion
17	AcceptedCmp1	int64	1 si el cliente aceptó la oferta de la primera campaña, 0 en caso contrario	0	Promotion
18	AcceptedCmp2	int64	1 si el cliente aceptó la oferta de la segunda	0	Promo

			campaña, 0 en caso contrario		tion
19	AcceptedCmp3	int64	1 si el cliente aceptó la oferta de la tercera campaña, 0 en caso contrario	0	Promotion
20	AcceptedCmp4	int64	1 si el cliente aceptó la oferta de la cuarta campaña, 0 en caso contrario	0	Promotion
21	AcceptedCmp5	int64	1 si el cliente aceptó la oferta de la quinta campaña, 0 en caso contrario	1	Promotion
22	Response	int64	1 si el cliente aceptó la oferta de la última campaña, 0 en caso contrario	1	Promotion
23	NumWebPurchases	int64	Número de compras realizadas a través del sitio web	3	Place
24	NumCatalogPurchases	int64	Número de compras realizadas usando un catálogo	2	Place
25	NumStorePurchases	int64	Número de compras realizadas directamente en tiendas	4	Place
26	NumWebVisitsMonth	int64	Número de visitas al sitio web en el último mes	7	Place
27	CostContact	float64	Costo de contactar al cliente	1.5	Costs
28	Revenue	float64	Ingresos generados por el cliente	120.75	Revenue

Pre Procesamiento

Nulos

Income: Se aplicó la moda de acuerdo a Education de cada registro.

Sin duplicados.

Eliminación de campos.

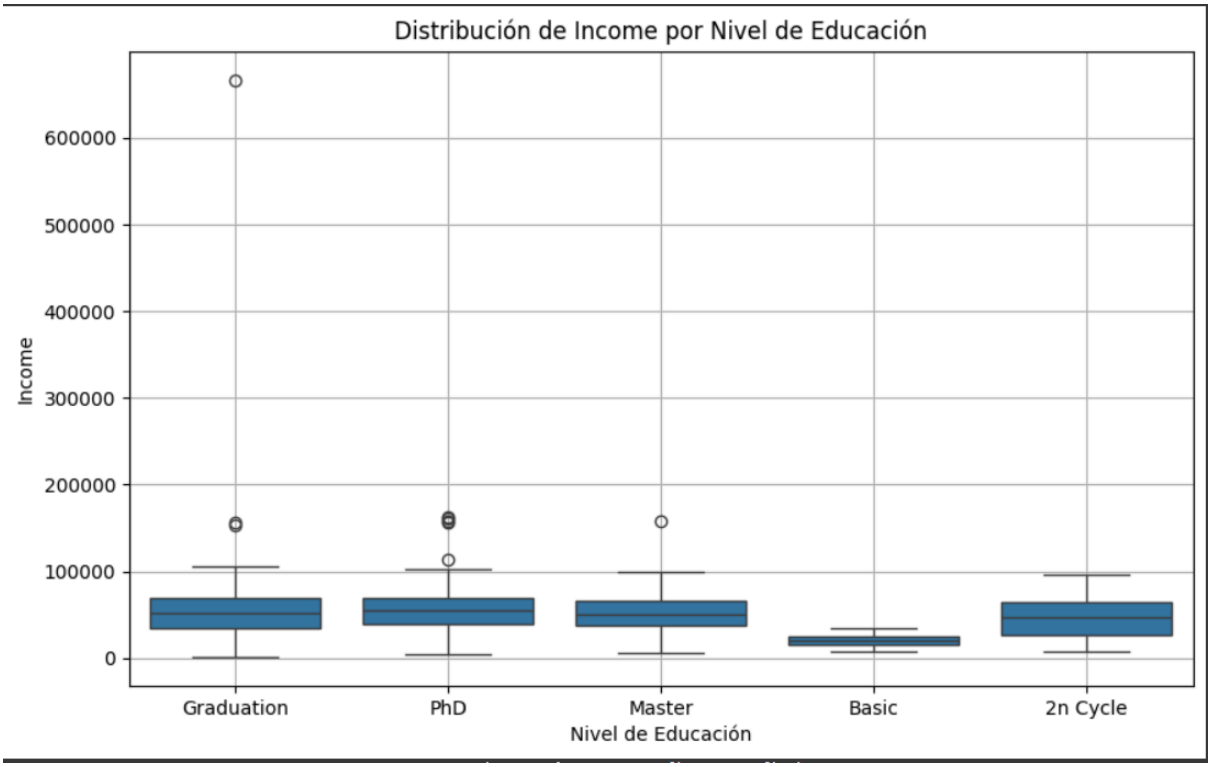
Dt_Customer: Más del 30% de nulos.

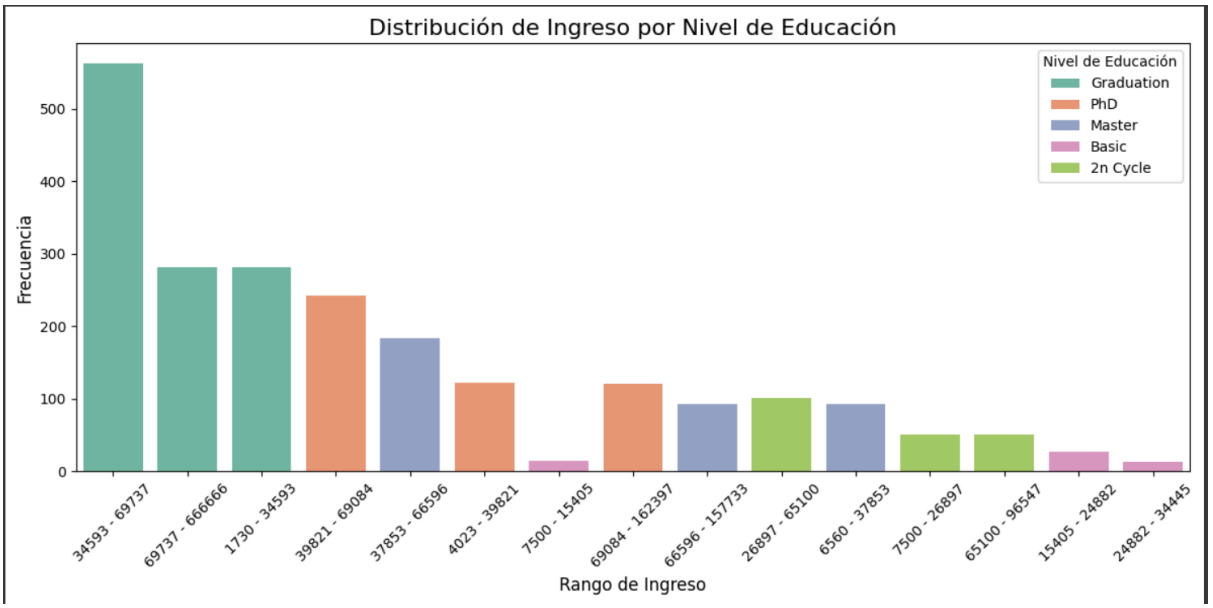
Reclasificar Categorías.

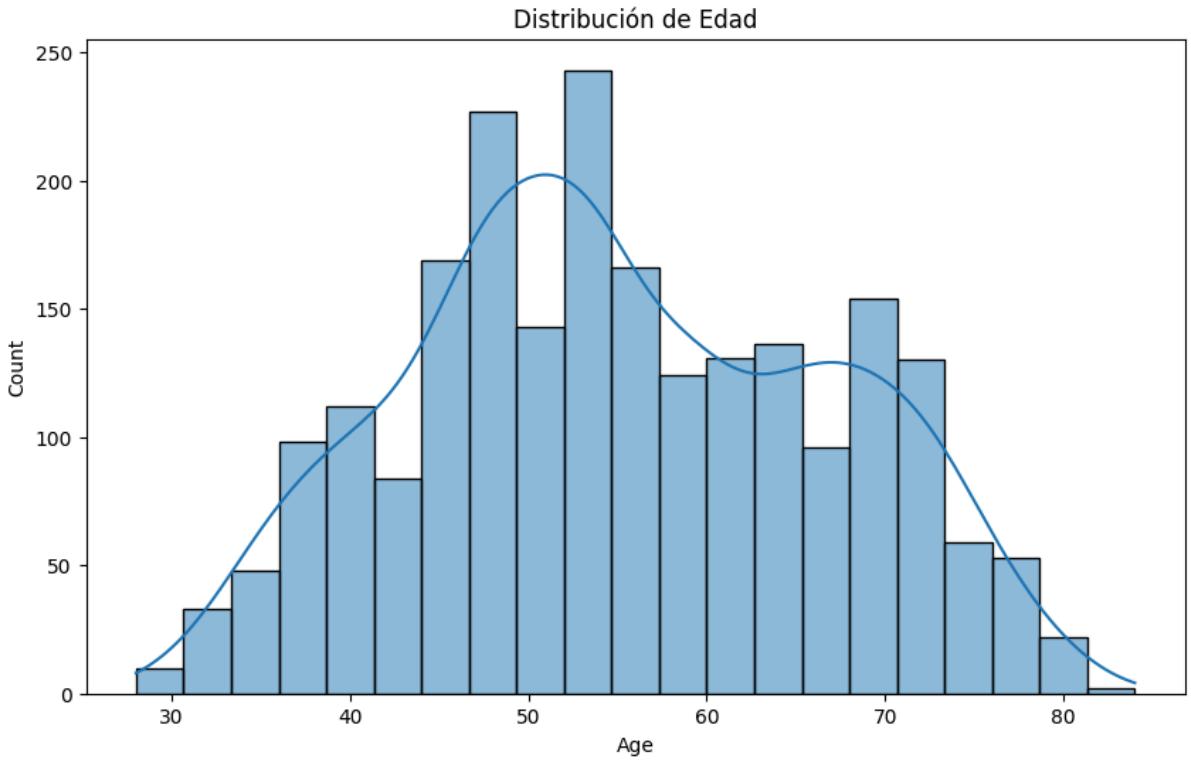
Las categorías Alone, Absurd y YOLO fueron registradas como Single

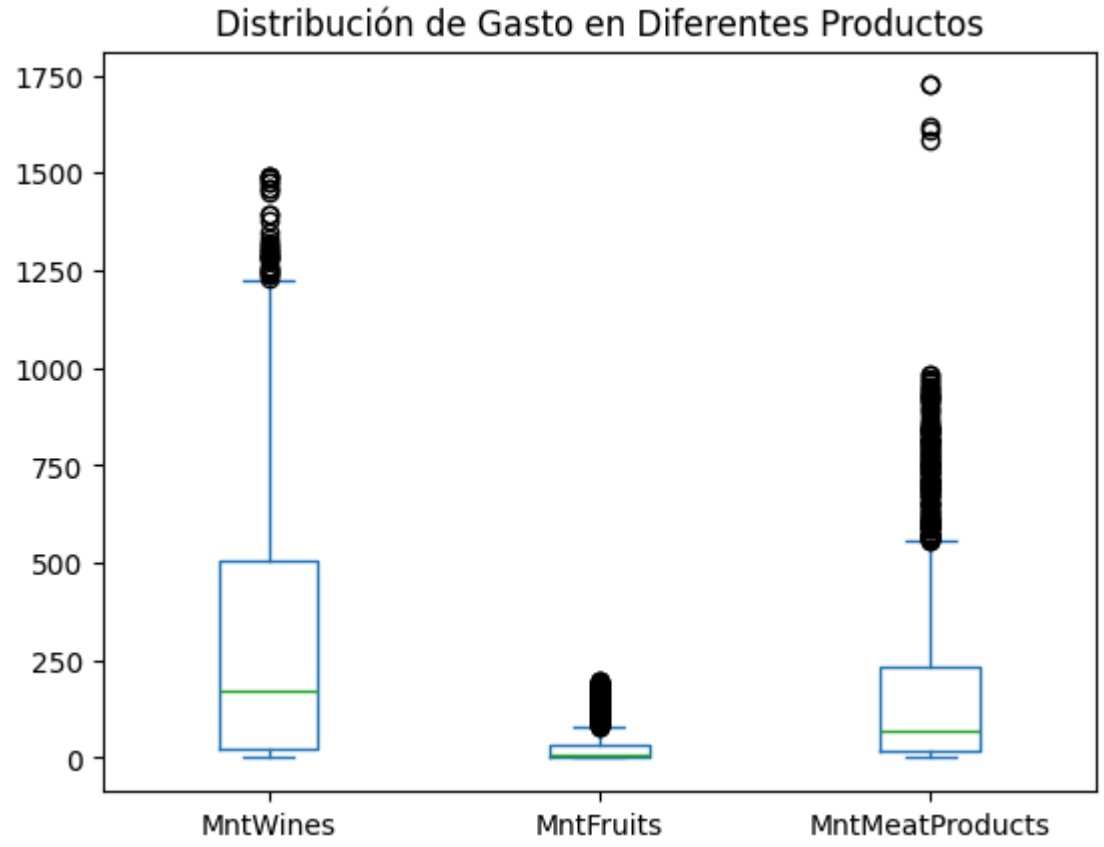
OutLiers

Year_Birth: Todos los que son menores a 1930 le aplique la más frecuente 1977.


Distribución de Income por Nivel de Educación
Income
600000
500000
400000
300000
200000
100000
0
Graduation
PhD
Master
Basic
2n Cycle
Nivel de Educación


Distribución de Ingreso por Nivel de Educación
Nivel de Educación
Graduation
PhD
Master
Basic
2n Cycle
Frecuencia
500
400
300
200
100
0
34593 - 69737
69737 - 666666
1730 - 34593
39821 - 69084
37853 - 66596
4023 - 39821
7500 - 15405
69084 - 162397
66596 - 157733
26897 - 65100
6560 - 37853
7500 - 26897
65100 - 96547
15405 - 24882
24882 - 34445
Rango de Ingreso





No se eliminan registros por outliers. Son valores de la realidad a modelar.

Nuevas Características

Age: Se aplica la resta 2024 del Year_Birth

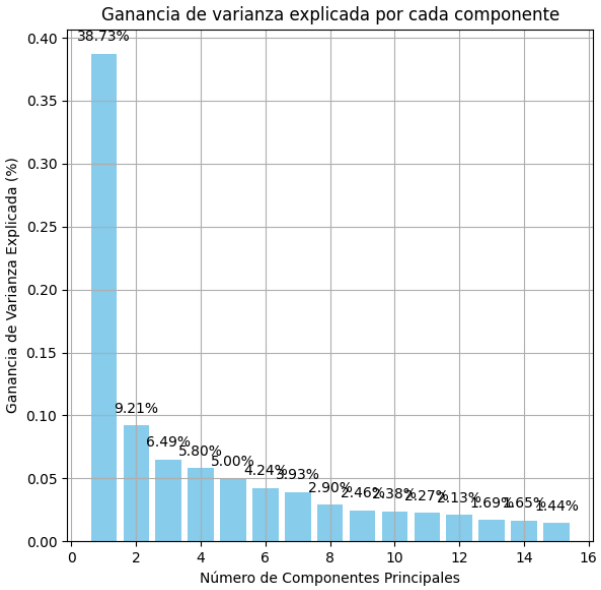
Análisis no supervisado

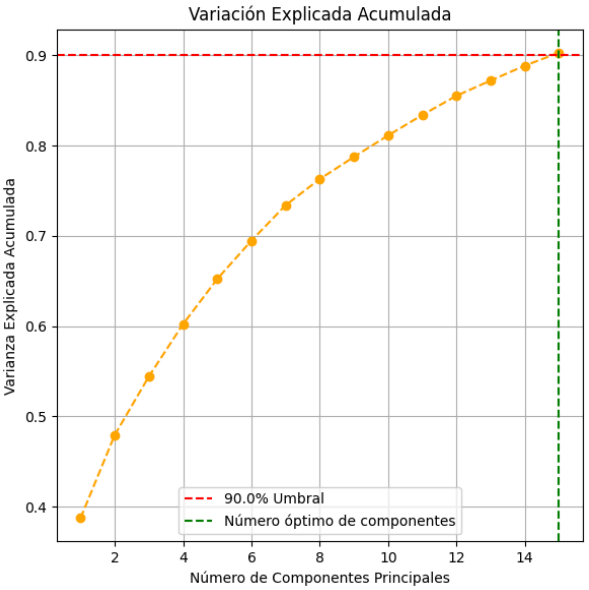
Enfoque no supervisado PCA

Varianza Mínima Explicada: 90%
Escalamiento: Robustscaler. Menos componentes que con Standard Scaler.
Componentes: 15
Variabilidad: 90.31%

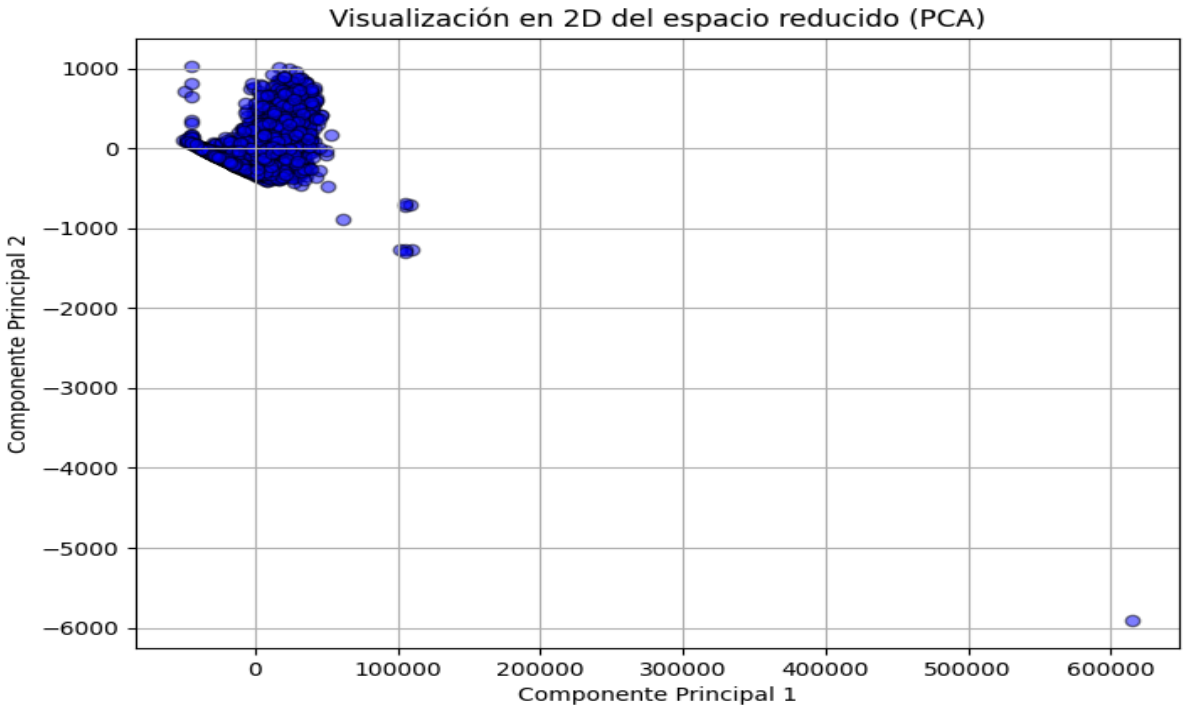
Gráficos.

Standard Scaler requiere 22 componentes para representar el 90.25%

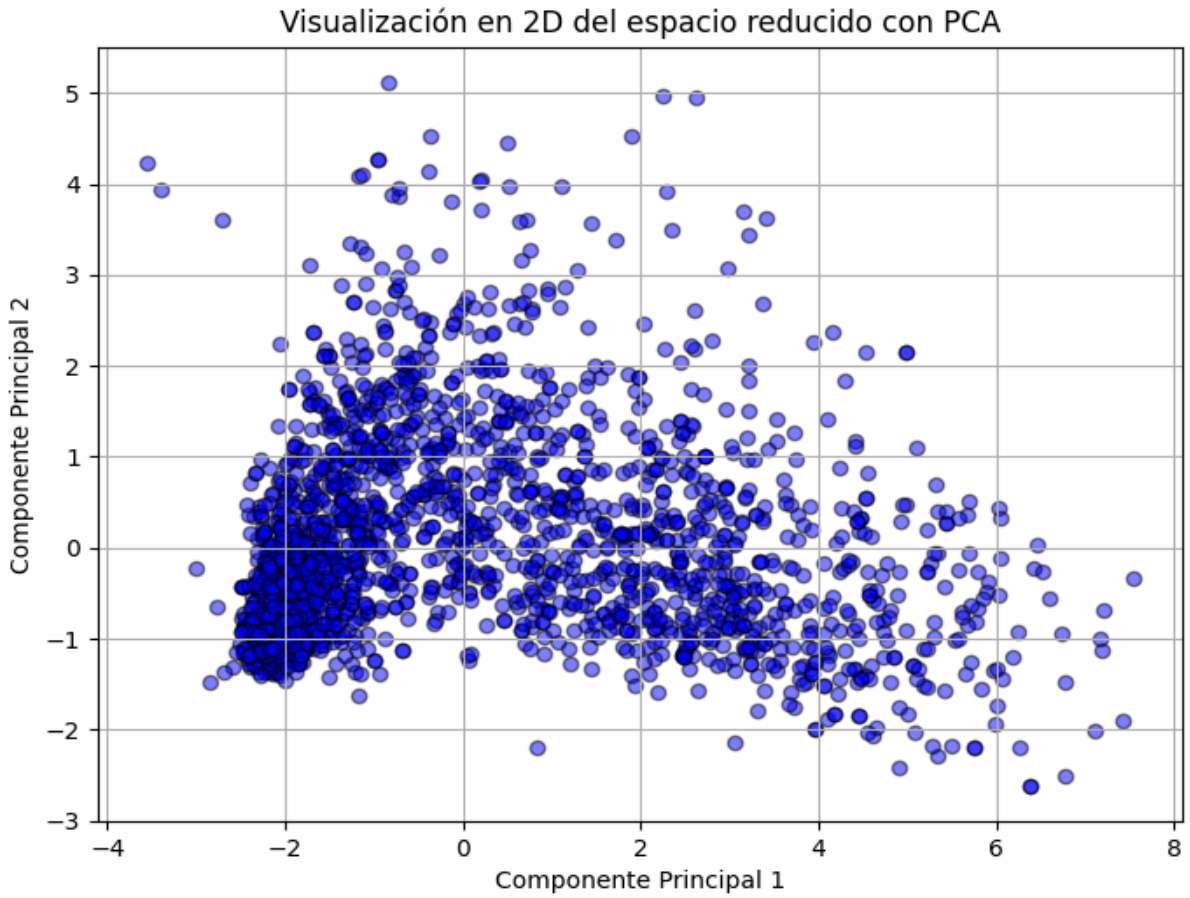


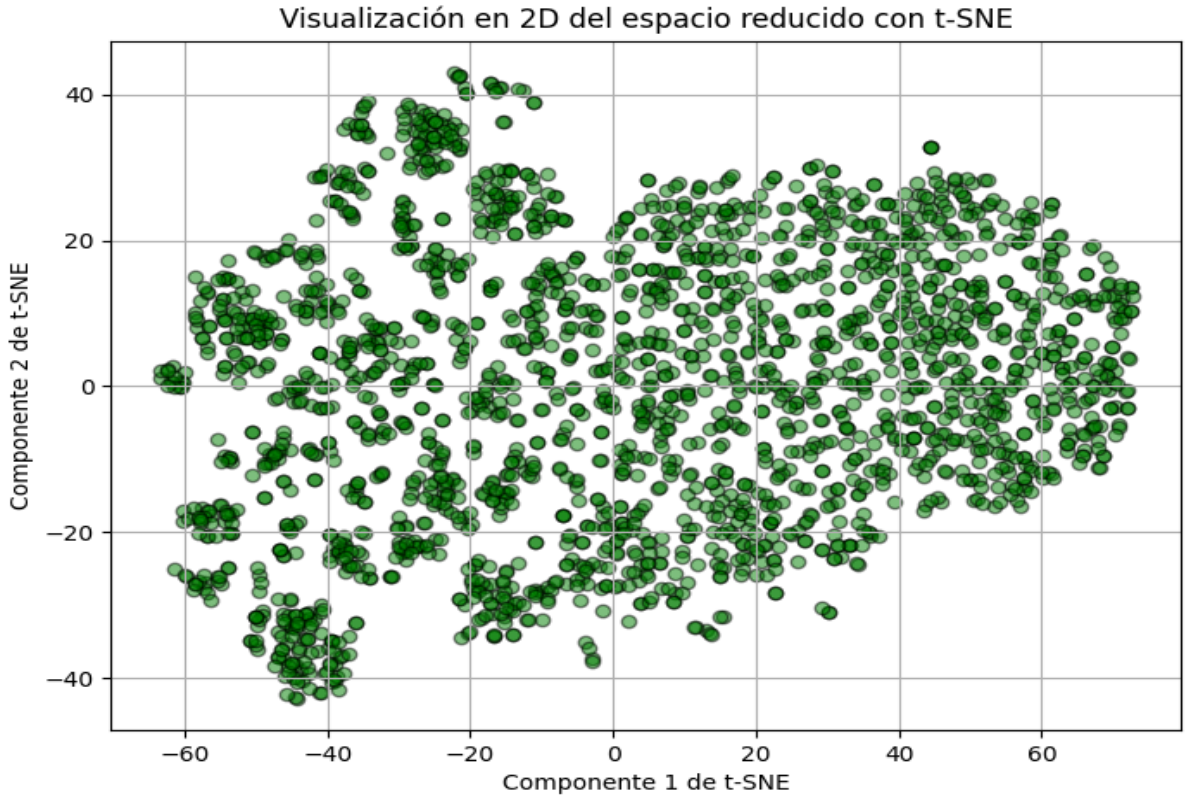


Sin Escalamiento

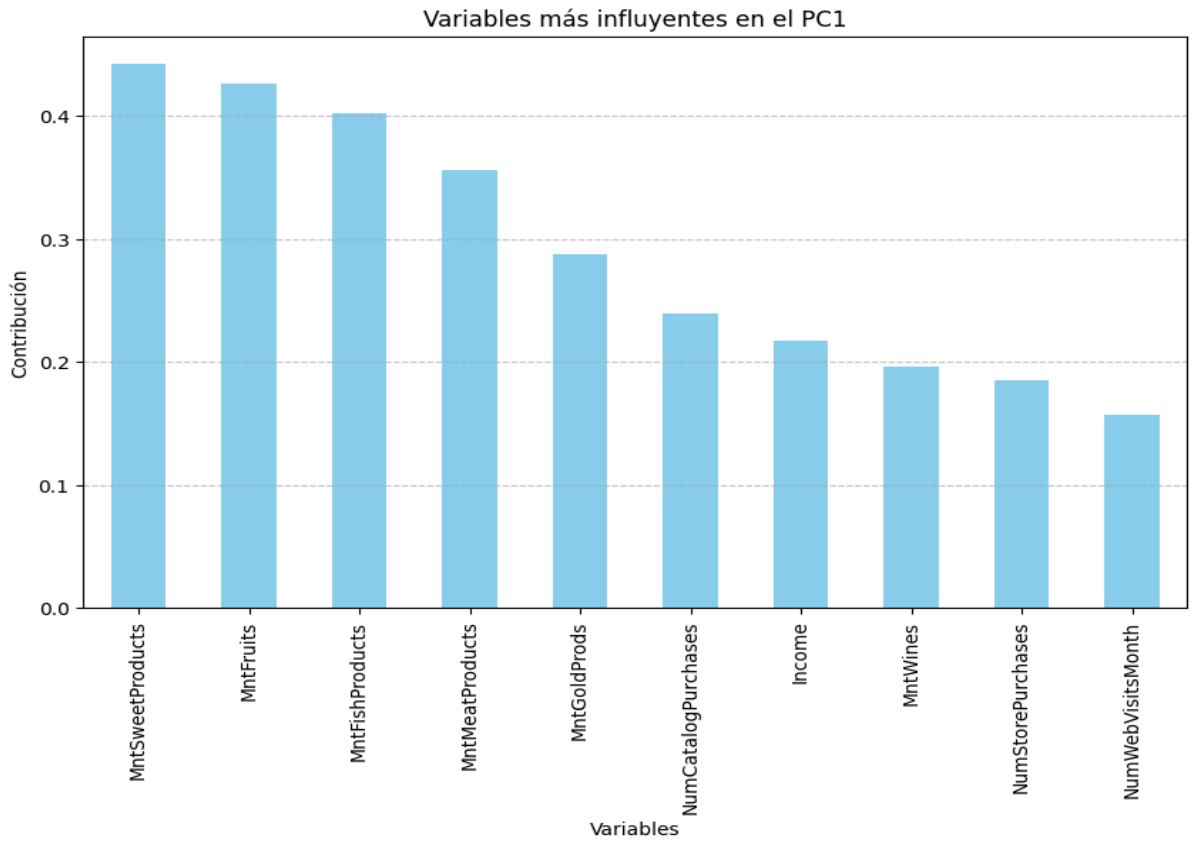


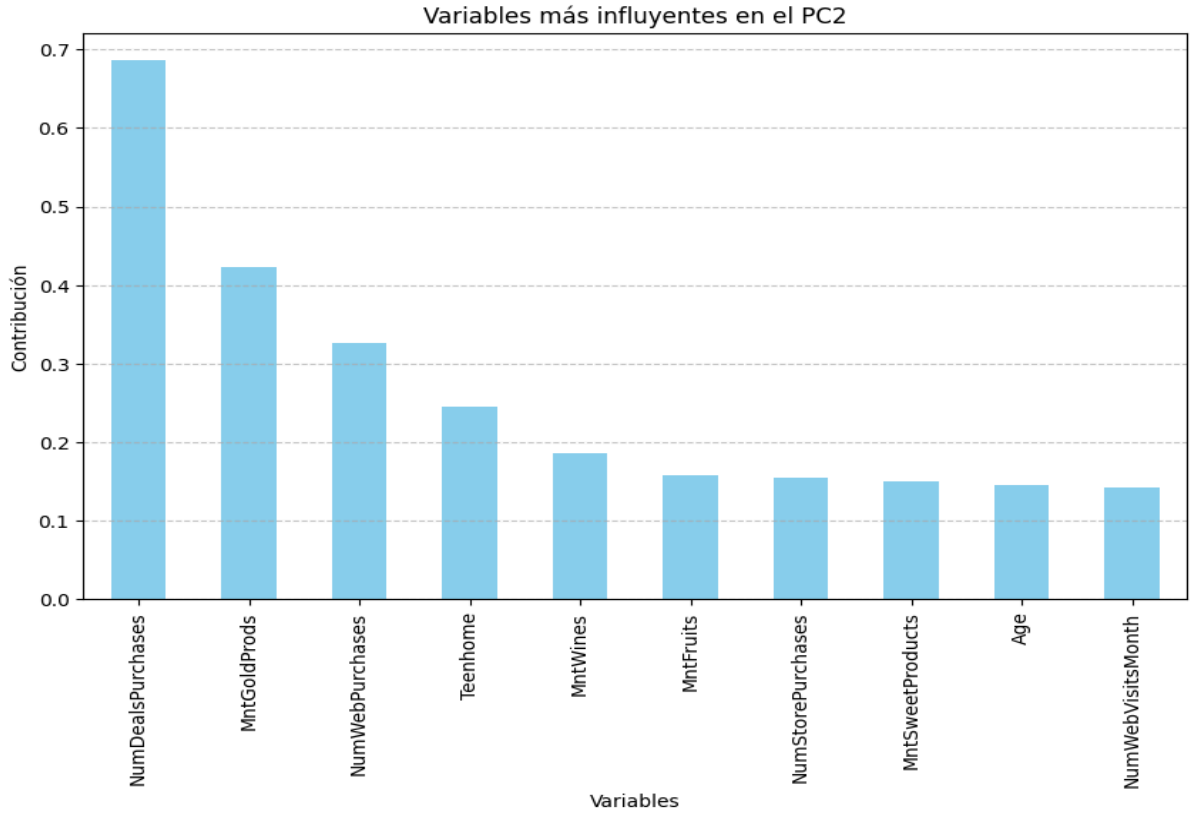
Con RobustScaler y sin t-SNE y luego con t-SNE

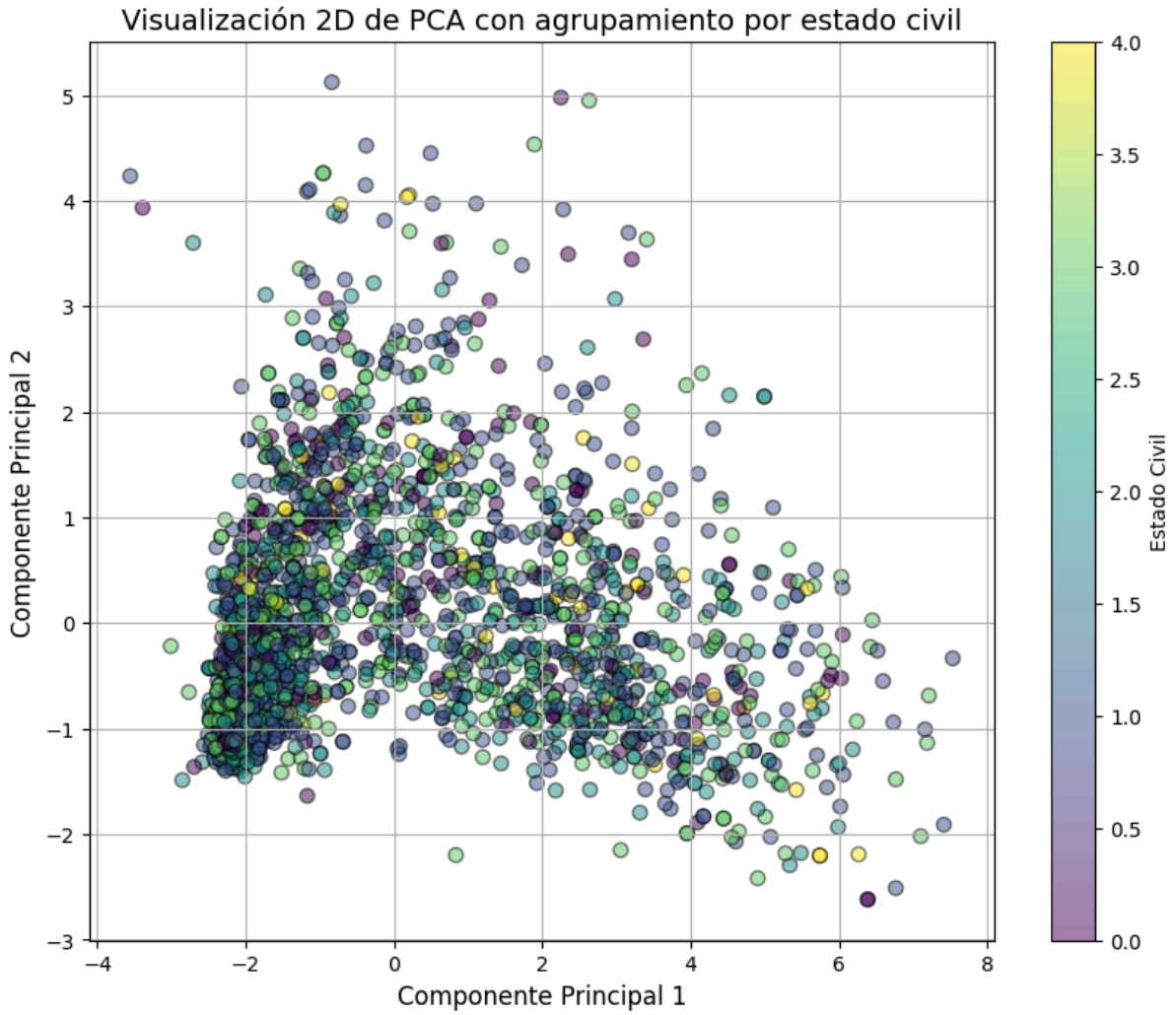




Se nota una mejoría en la agrupación


Variables más influyentes en el PC1
Contribución
0.0
0.1
0.2
0.3
0.4
MntSweetProducts
MntFruits
MntFishProducts
MntMeatProducts
MntGoldProds
NumCatalogPurchases
Income
MntWines
NumStorePurchases
NumWebVisitsMonth
Variables


Variables más influyentes en el PC2
Contribución
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
NumDealsPurchases
MntGoldProds
NumWebPurchases
Teenhome
MntWines
MntFruits
NumStorePurchases
MntSweetProducts
Age
NumWebVisitsMonth
Variables



Resumen

Al aplicar PCA se logró la varianza al 90% pero al ver los dos primeros componentes no se ve una agrupación clara.

Al aplicar el t-SNE mejoró un poco la visualización pero aún así no hay una clara agrupación, no se está capturando las relaciones internas del dataset.

Detalle de la varianza lograda en los dos primeros componentes.

- Componente 1: 38.78% (38.78%)
- Componente 2: 48.02% (9.23%)

Si agrupamos a los clientes en cuanto a su estado civil, TODOS consumen productos o las variables detalladas en los componentes 1 y 2.

Lo que llama la atención es que carnes tiene una menor influencia en el componente 1, está en el cuarto lugar pero en cuanto a montos es de los primeros.

Esto puede ser debido a que es muy costoso y que los clientes lo compren menos y prefieren pescados, que es más barato.

Variables más influyentes en PC1:

MntSweetProducts: 0.442320
MntFruits: 0.425971
MntFishProducts: 0.402110
MntMeatProducts: 0.356036
MntGoldProds: 0.287853
Variables más influyentes en PC2:

NumDealsPurchases: 0.685943
MntGoldProds: 0.422086
NumWebPurchases: 0.325921
Teenhome: 0.245124
MntWines: 0.185925

Modelado MLP

Predecir etiquetas del Volumen Total de Ventas.

Categorías.

1. Baja: 0-99
2. Media: 100-199
3. Alta: 200-499
4. Muy Alta: 500 en adelante.

Balanceo de Categorías con Smote.

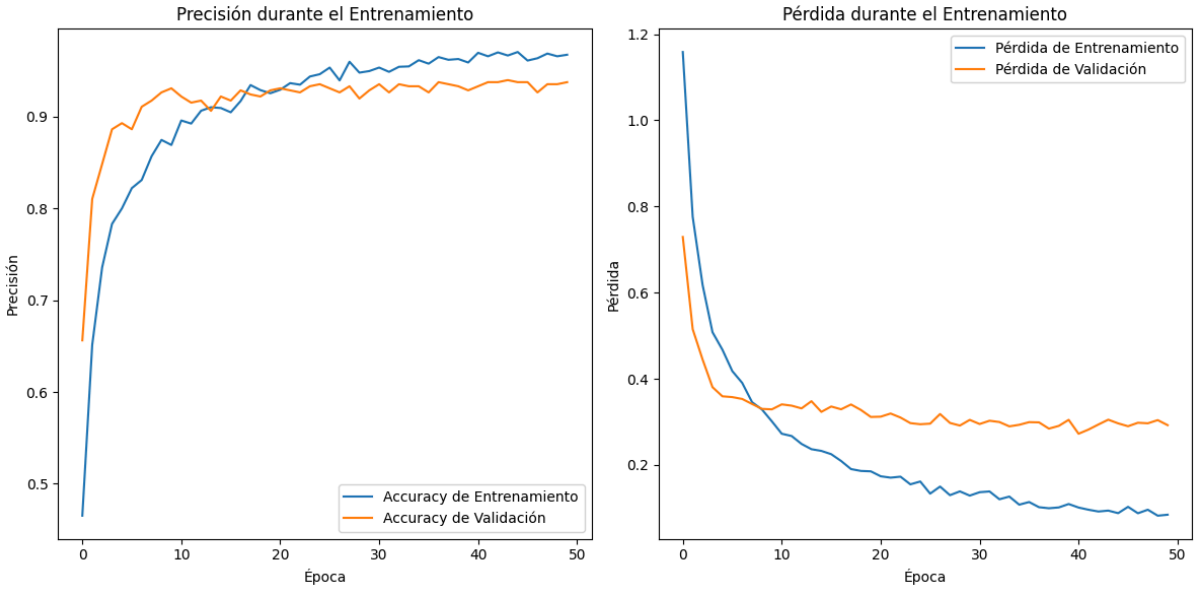
Conjuntos de datos.

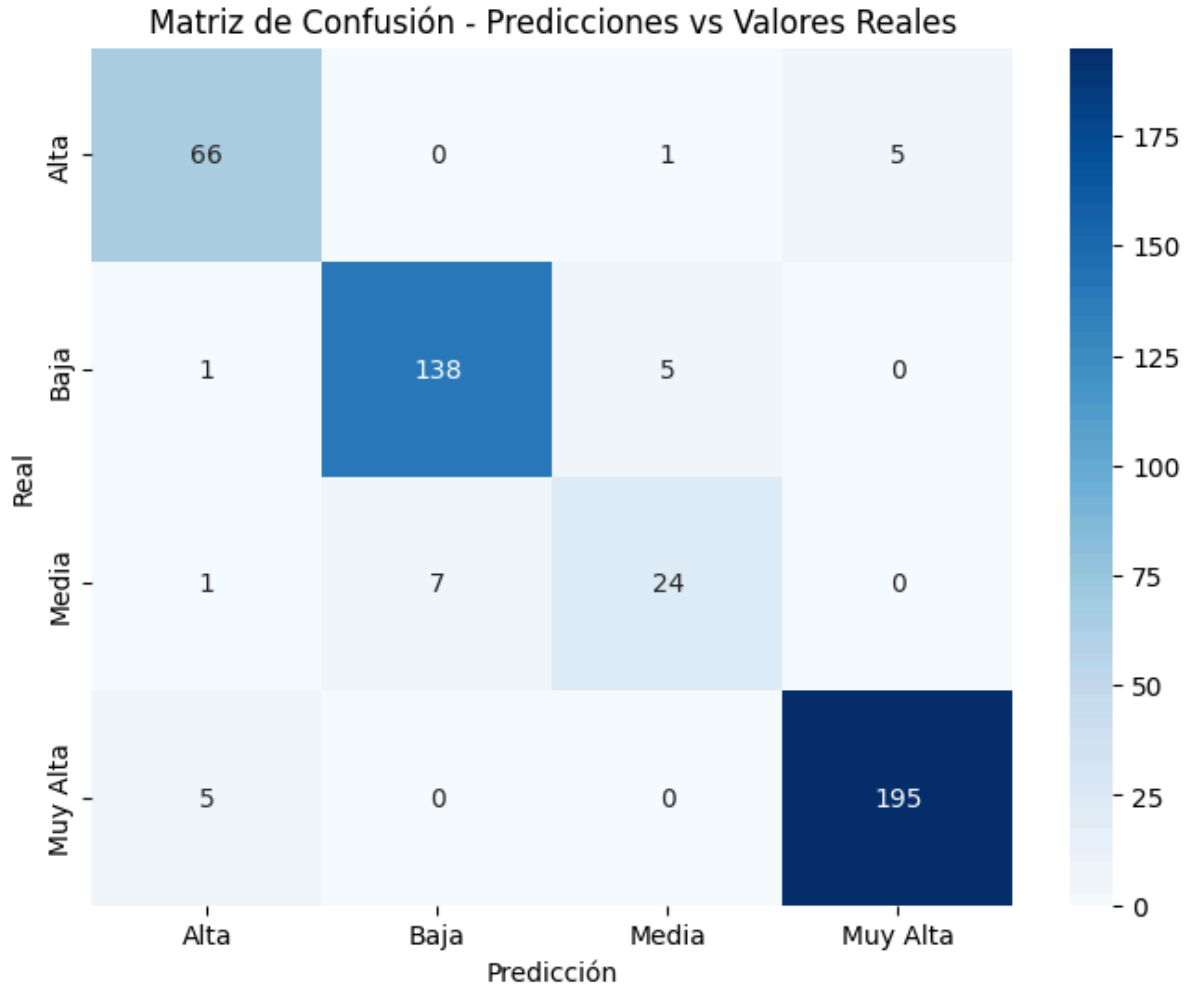
Datos de Entrenamiento, Test y Validación.

Configuración

Modelo de densidad con 2 capas ocultas de 64 y 32 con activación ReLU.
Capa de salida con 4 clases. Activación SofMax
Prevención de sobre ajuste:
Dropout en ambas capas internas al 30%
Early Stopping

Gráficos.


Precisión durante el Entrenamiento
Precisión
Época
Accuracy de Entrenamiento
Accuracy de Validación
Pérdida durante el Entrenamiento
Pérdida
Época
Pérdida de Entrenamiento
Pérdida de Validación


Matriz de Confusión - Predicciones vs Valores Reales
Real
Predicción
Alta
Baja
Media
Muy Alta
66
0
1
5
1
138
5
0
1
7
24
0
5
0
0
195

Evaluación del modelo y conclusiones

Con respecto al PCA con el RobusScaler solo preciso 15 componentes para explicar la varianza al menos del 90%. Pero al ver como quedó en el espacio, no vi claramente separación entre los componentes. No logra exponer adecuadamente la agrupación. Creo que hay relaciones no lineales fuera del alcance del PCA.

Ante esta situación tuve que pensar en una nueva característica para poder predecir algo importante para este dataset. No me fue posible estimar gastos por productos. Entonces, se me ocurrió en sumar todas las ventas y segmentar por Baja, Media, Alta y Muy Alta y con esos sí pude tener unas predicciones mucho mejores.

Me pareció necesario realizar un balanceo de las clases y al aplicar la nueva característica tuve un rendimiento general del 94%

Se podría intentar mejorar con una paciencia menor al 10 y ver otros hyperparámetros, e investigar y probar más.