# Segmentación de Clientes y Predicción Basada en Clustering: Aplicación de K-Means y Árboles de Decisión

**2025**

**Técnicas Aprendizaje de Máquina**

**Juan Torres, Ney Peñuela y Juliana Rubio**

# Objetivo

El objetivo de este proyecto es aplicar y analizar técnicas de segmentación de clientes mediante métodos de clustering no supervisado, específicamente utilizando K-Means, con el propósito de identificar patrones de comportamiento y características comunes entre distintos clientes. Además, se busca utilizar los segmentos generados como etiquetas para entrenar un modelo supervisado basado en árboles de decisión, con el fin de predecir el grupo al que pertenecerá un cliente nuevo.

Estas herramientas permitirán a la empresa desarrollar estrategias personalizadas para cada segmento, optimizando su enfoque en campañas de marketing, diseño de productos y servicios, así como en la atención al cliente. Al comprender las necesidades específicas de cada grupo, la empresa podrá tomar decisiones más informadas que contribuyan al fortalecimiento de la fidelidad de los clientes y la mejora de la experiencia del usuario.

# Análisis Exploratorio y Preprocesamiento de Datos

**Presentación del dataset**

El dataset utilizado contiene un total de 13 variables y 1833331 observaciones que describen diferentes aspectos del comportamiento y características de los usuarios. Entre estas variables se encuentran datos numéricos como la edad, los ingresos anuales, la cantidad de compras, el valor promedio de compra, la frecuencia de compras mensual y el valor total gastado, además de

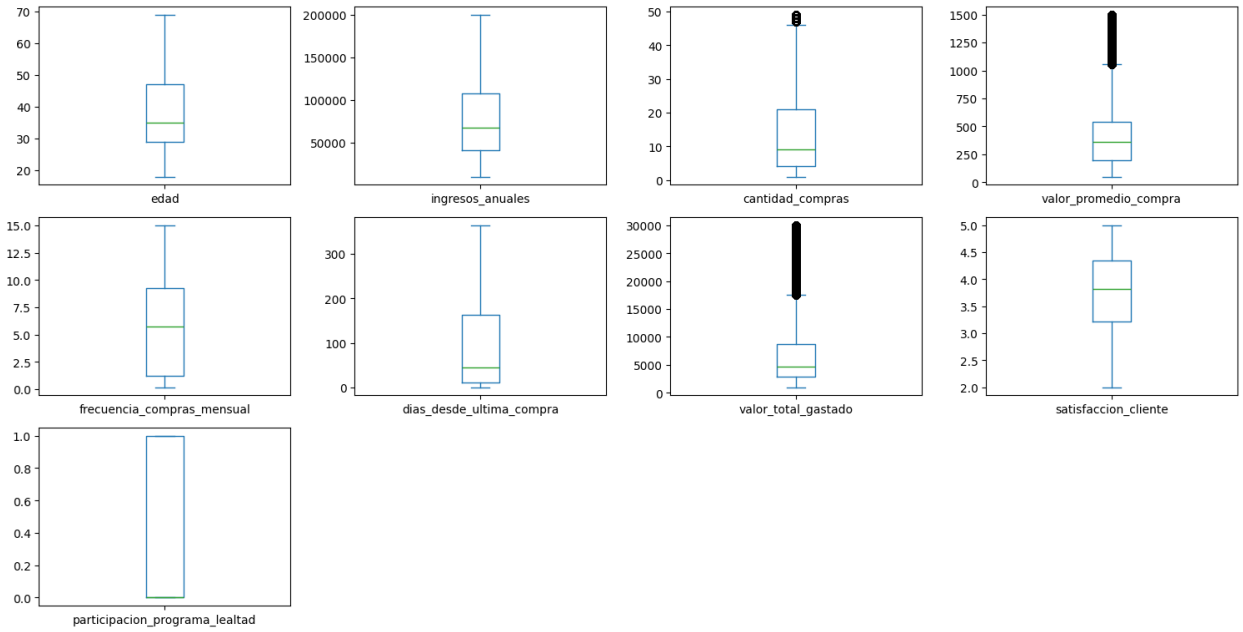
aspectos temporales como los días desde la última compra. También incluye variables categóricas como el dispositivo utilizado, la fuente de tráfico, el método de pago y los productos adquiridos, que aportan contexto sobre las preferencias de los clientes. Finalmente, variables como la satisfacción del cliente y la participación en programas de lealtad permiten evaluar el compromiso y la experiencia del cliente con la empresa.

## Análisis descriptivo del dataset

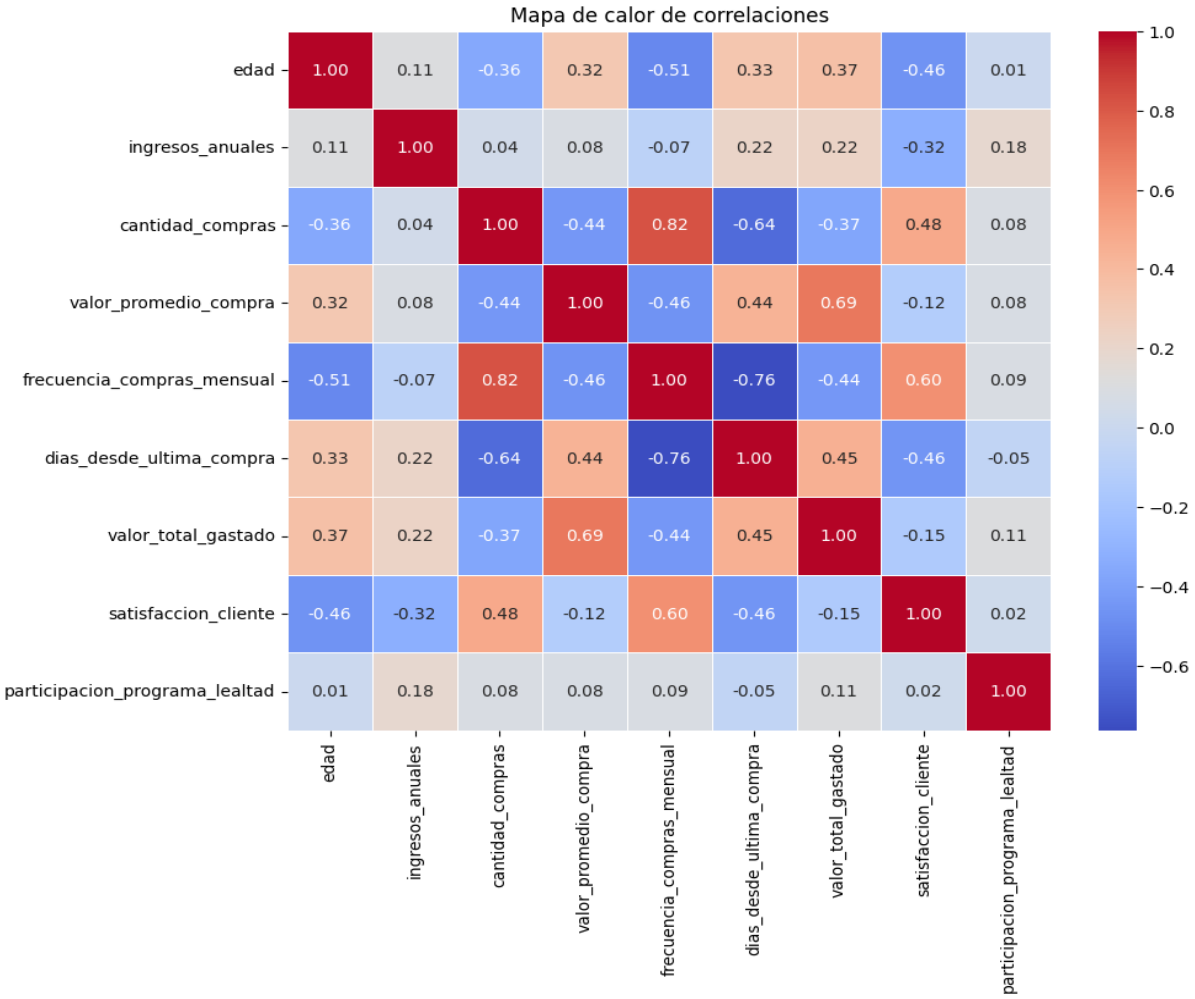
| | edad | ingresos_anuales | cantidad_compras | valor_promedio_compra | frecuencia_compras_mensual | dias_desde_ultima_compra | valor_total_gastado | satisfaccion_cliente | participacion_programa_lealtad |
|---|---|---|---|---|---|---|---|---|---|
| count | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 | 1.833331e+06 |
| mean | 3.877552e+01 | 7.614756e+04 | 1.408932e+01 | 4.301820e+02 | 5.487638e+00 | 9.651723e+01 | 7.045614e+03 | 3.761349e+00 | 4.636860e-01 |
| std | 1.358327e+01 | 4.419407e+04 | 1.182157e+01 | 3.086627e+02 | 4.209958e+00 | 1.065618e+02 | 6.204356e+03 | 7.281546e-01 | 4.986797e-01 |
| min | 1.800000e+01 | 1.000000e+04 | 1.000000e+00 | 5.000000e+01 | 2.000125e-01 | 1.000000e+00 | 1.000000e+03 | 2.000002e+00 | 0.000000e+00 |
| 25% | 2.900000e+01 | 4.131500e+04 | 4.000000e+00 | 1.960000e+02 | 1.235148e+00 | 1.200000e+01 | 2.944000e+03 | 3.220253e+00 | 0.000000e+00 |
| 50% | 3.500000e+01 | 6.750700e+04 | 9.000000e+00 | 3.610000e+02 | 5.749663e+00 | 4.500000e+01 | 4.606000e+03 | 3.823675e+00 | 0.000000e+00 |
| 75% | 4.700000e+01 | 1.078480e+05 | 2.100000e+01 | 5.400000e+02 | 9.280341e+00 | 1.640000e+02 | 8.770000e+03 | 4.348068e+00 | 1.000000e+00 |
| max | 6.900000e+01 | 1.999990e+05 | 4.900000e+01 | 1.499000e+03 | 1.499999e+01 | 3.640000e+02 | 2.999900e+04 | 5.000000e+00 | 1.000000e+00 |

- **Edad**: Los clientes tienen una edad promedio de 38.8 años, con un rango que va desde los 18 hasta los 69 años. La desviación estándar de 13.6 indica que hay una dispersión considerable en la edad de los clientes.
- **Ingresos anuales**: El ingreso promedio es de 76,147, con un rango que varía entre 10,000 y 199,999. Los ingresos presentan una desviación estándar alta (44,194), lo que refleja una diversidad significativa en la capacidad adquisitiva de los clientes.
- **Cantidad de compras**: En promedio, los clientes realizan 14 compras, aunque la cantidad varía desde 1 hasta 49 compras. Esto sugiere que hay tanto compradores ocasionales como recurrentes.
- **Valor promedio de compra**: El valor promedio por compra es de 430, con un rango que va desde 50 hasta 1,499. La desviación estándar de 308 muestra una variación significativa en los montos gastados por los clientes en una sola transacción.
- **Frecuencia de compras mensual**: La frecuencia promedio es de 5.5 compras mensuales, con clientes que realizan desde 0.2 hasta 15 compras al mes. Esto destaca la diferencia en la actividad de compra entre los clientes.
- **Días desde la última compra**: En promedio, los clientes compraron hace 96.5 días, aunque algunos hicieron compras hace solo 1 día y otros hasta hace 364 días. Esto muestra diferentes niveles de actividad reciente.
- **Valor total gastado**: Los clientes tienen un gasto total promedio de 7,045, con un rango que va desde 1,000 hasta 29,999. Esto refleja una diferencia marcada entre clientes de bajo y alto valor para la empresa.

- **Satisfacción del cliente**: La satisfacción promedio es de 3.76, con clientes que reportan valores entre 2 y 5. Esto indica una mayoría de clientes moderadamente satisfechos, pero también da espacio para mejorar.
- **Participación en programas de lealtad**: Solo el 46.4 % de los clientes participa en programas de lealtad, lo que destaca la importancia de incentivar este compromiso para fortalecer la relación con la marca.




Como se puede observar en los boxploty, las variables 'cantidad_compras', 'valor_promedio_compra' y 'valor_total_gastado' presentan una gran cantidad de valores atípicos hacia arriba (valores altos). Sin embargo, debido a la naturaleza del problema y la distribución de estas variables, no necesariamente se pueden considerar como errores. Estos valores podrían representar comportamientos legítimos de clientes con un gasto elevado o un mayor número de compras, lo cual es importante para el análisis.




La matriz de correlaciones muestra relaciones clave entre las variables del dataset. La frecuencia de compras mensual tiene una fuerte correlación positiva con la cantidad de compras (0.822) y con la satisfacción del cliente (0.600), lo que sugiere que los clientes más activos suelen estar más satisfechos. Por el contrario, está negativamente correlacionada con los días desde la última compra (-0.763), indicando que quienes compran más frecuentemente hacen compras más recientes.

Variables como la edad y los ingresos anuales tienen correlaciones moderadas con el valor total gastado (0.366 y 0.224, respectivamente), mostrando que los clientes mayores o con mayores ingresos tienden a gastar más en total. Sin embargo, la participación en programas de lealtad tiene correlaciones bajas con la mayoría de las variables, lo que indica que no está directamente influenciada por características de compra o demográficas. Estas correlaciones ayudan a identificar patrones clave para el análisis y segmentación.

A pesar de las correlaciones significativas entre algunas variables, decidimos mantener todas en el análisis, incluso aquellas con alta correlación. Esto se debe a que cada variable podría aportar

información única y relevante para la segmentación del modelo K-Means. Eliminar alguna de ellas podría resultar en una pérdida de detalles importantes que podrían enriquecer la calidad y la interpretación de los clusters generados.

# Preprocesamiento de datos

### Transformación de variables numéricas

Se transformaron las variables numéricas edad, ingresos_anuales, cantidad_compras, valor_promedio_compra, frecuencia_compras_mensual, dias_desde_ultima_compra, valor_total_gastado, satisfaccion_cliente (la variable participación_programa_lealtad fue excluida del proceso ya que es binaria), utilizando un escalado estándar mediante *StandardScaler*, lo cual normaliza las variables para que tengan una media de 0 y una desviación estándar de 1. Esto nos asegura que ninguna variable predomine sobre otra cuando se apliquen las técnicas de clustering y de árboles de decisión.

| | edad | ingresos_anuales | cantidad_compras | valor_promedio_compra | frecuencia_compras_mensual | dispositivo_utilizado | fuente_trafico | dias_desde_ultima_compra | valor_total_gastado | satisfaccion_cliente |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | -0.793294 | 0.323741 | -0.853467 | -0.956326 | -1.084620 | tablet | redes sociales | 1.252633 | -0.592586 | -1.142011 |
| 1 | -0.277954 | -0.459464 | 0.415400 | -0.204696 | -0.039432 | móvil | búsqueda orgánica | -0.746208 | 0.198310 | -0.432325 |
| 2 | -0.646054 | -1.196531 | 2.107223 | -0.561072 | 0.839362 | móvil | búsqueda orgánica | -0.783745 | -0.778423 | 0.375742 |
| 3 | -0.277954 | -0.346349 | -0.515103 | 1.398997 | -0.622752 | móvil | búsqueda orgánica | -0.230075 | 0.154309 | -0.404628 |
| 4 | -1.014154 | 0.550695 | 0.415400 | 0.044767 | 0.446139 | móvil | redes sociales | -0.370839 | 0.242473 | 0.753970 |

### Codificación de variables categóricas

Dado que el dataset incluye las siguientes variables categóricas con sus respectivos valores:

```
Valores únicos en 'dispositivo_utilizado':
['tablet' 'móvil' 'PC']
--------------------------------------------------
Valores únicos en 'fuente_trafico':
['redes sociales' 'búsqueda orgánica' 'email']
--------------------------------------------------
Valores únicos en 'metodo_pago':
['tarjeta crédito' 'paypal' 'transferencia']
--------------------------------------------------
Valores únicos en 'productos_adquiridos':
['computadoras' 'accesorios' 'electrodomésticos' 'móviles']
--------------------------------------------------
```

Se procedió a transformar las variables categóricas del dataset (dispositivo_utilizado, fuente_trafico, metodo_pago y productos_adquiridos) utilizando la técnica de One-Hot Encoding. Esta genera una columna binaria (0 o 1) para cada categoría, permitiendo que las variables categóricas sean interpretadas correctamente en los modelos analíticos.

| | dispositivo_utilizado_PC | ... | fuente_trafico_búsqueda orgánica | \ |
|---|---|---|---|---|
| 0 | 0.0 | ... | 0.0 | |
| 1 | 0.0 | ... | 1.0 | |
| 2 | 0.0 | ... | 1.0 | |
| 3 | 0.0 | ... | 1.0 | |
| 4 | 0.0 | ... | 0.0 | |

| | fuente_trafico_email | fuente_trafico_redes sociales | metodo_pago_paypal | \ |
|---|---|---|---|---|
| 0 | 0.0 | 1.0 | 0.0 | |
| 1 | 0.0 | 0.0 | 0.0 | |
| 2 | 0.0 | 0.0 | 0.0 | |
| 3 | 0.0 | 0.0 | 1.0 | |
| 4 | 0.0 | 1.0 | 1.0 | |

| | metodo_pago_tarjeta crédito | metodo_pago_transferencia | \ |
|---|---|---|---|
| 0 | 1.0 | 0.0 | |
| 1 | 1.0 | 0.0 | |
| 2 | 1.0 | 0.0 | |
| 3 | 0.0 | 0.0 | |
| 4 | 0.0 | 0.0 | |

| | productos_adquiridos_accesorios | productos_adquiridos_computadoras | \ |
|---|---|---|---|
| 0 | 0.0 | 1.0 | |
| 1 | 1.0 | 0.0 | |
| 2 | 1.0 | 0.0 | |
| 3 | 0.0 | 0.0 | |
| 4 | 0.0 | 0.0 | |

| | productos_adquiridos_electrodomésticos | productos_adquiridos_móviles |
|---|---|---|
| 0 | 0.0 | 0.0 |
| 1 | 0.0 | 0.0 |
| 2 | 0.0 | 0.0 |
| 3 | 1.0 | 0.0 |
| 4 | 0.0 | 1.0 |

[5 rows x 22 columns]

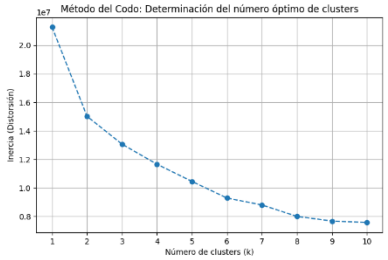
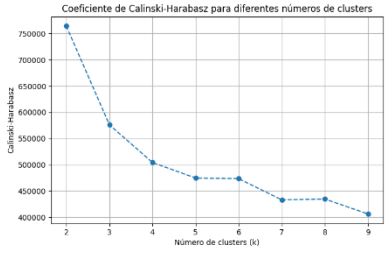
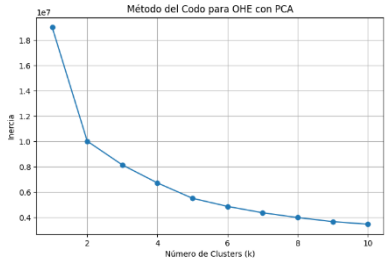
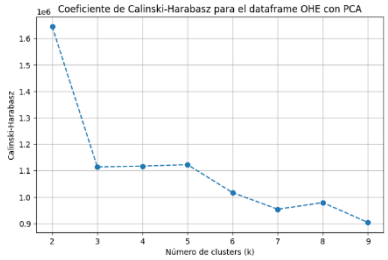
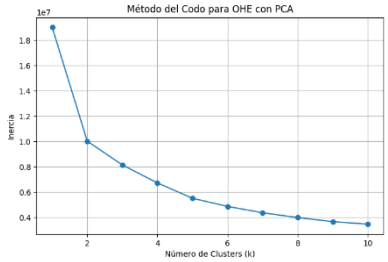
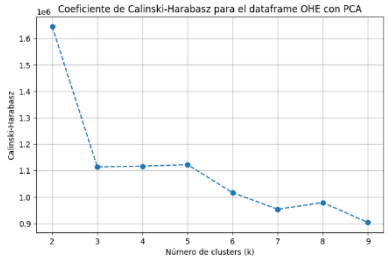
Tras la transformación quedan un total de 22 columnas en el dataset (9 columnas adicionales).

# Aplicación de K-Means

## Análisis sobre la selección de números de clústeres

Inicialmente, para determinar el número óptimo de clústeres, se evaluó el impacto de la alta dimensionalidad generada por la codificación one-hot encoding (OHE) en el agrupamiento. Dado los valores que torna este proceso (0 o 1) son absolutistas, pudiendo influir en la agrupación euclidiana, se optó por un análisis comparativo. Este estudio se dividió en tres enfoques para examinar la afectación de la dimensionalidad: utilizando solo las variables numéricas, los datos procesados con OHE, y los datos procesados con OHE seguido de reducción de dimensionalidad con PCA. Para un análisis robusto, se emplearon el método del codo, el índice de silueta y el coeficiente de Calinski-Harabasz.

El método del codo ayuda a identificar el punto en el que agregar más clusters deja de mejorar significativamente la agrupación (reducir la suma de las distancias cuadradas dentro de los clusters), mientras que el coeficiente de Calinski-Harabasz mide qué tan compactos y separados están los clusters.

| Metodología de manejo dataframe | Método del Codo | Calinski-Harabasz |
|---|---|---|
| OHE sin PCA |  |  |
| OHE con PCA |  |  |
| Numérico |  |  |

Por otro lado, se usó el método de la silueta, la cuál es una técnica utilizada para evaluar la calidad de un agrupamiento en análisis de clustering. Se basa en medir qué tan bien cada punto está asignado a su grupo en comparación con otros clusters. Para ello, calcula un puntaje que refleja la cohesión dentro de un mismo cluster y la separación respecto a los demás grupos. Los valores de silueta varían entre -1 y 1; Un valor cercano a 1 indica que los puntos están bien agrupados dentro de su cluster y alejados de los demás. Un valor cercano a 0 sugiere que los puntos están en el límite entre clusters, lo que puede indicar superposición.

Este método permite determinar el número óptimo de clusters analizando el promedio de los puntajes de silueta. Un valor alto sugiere una mejor agrupación, ayudando a elegir la cantidad ideal de clusters para el dataset.

Los resultados obtenidos fueron los siguientes:

| Número de Clusters (n_clusters) | Score de Silueta Promedio (OHE sin PCA) | Score de Silueta Promedio (OHE con PCA) | Score de Silueta Promedio (Sin Categoricos) |
|---|---|---|---|
| 2 | 0.2885 | 0.4159 | 0.6069 |
| 3 | 0.2726 | 0.3687 | 0.5152 |
| 4 | 0.2129 | 0.2886 | 0.5356 |
| 5 | 0.2517 | 0.3147 | 0.5265 |
| 6 | 0.2605 | 0.3153 | 0.512 |
| 7 | 0.2058 | 0.2982 | 0.4712 |
| 8 | 0.2136 | 0.302 | 0.4626 |
| 9 | 0.1955 | 0.3008 | 0.4548 |
| 10 | 0.1963 | 0.2971 | 0.4606 |

Por lo tanto, con base en el método de la silueta, el coeficiente de Calinski-Harabasz, y el método del codo coinciden en que el número óptimo de clusters para este análisis es **2**. Y por lo tanto,se recomienda utilizar 2 clústeres como la mejor opción para este conjunto de datos, sin importar el manejo que se dé.

Además, considerando la pérdida de agrupabilidad observada en los diferentes métodos y, en particular, en el análisis de la silueta, se concluye que la mejor estrategia es utilizar el dataframe donde las variables categóricas fueron reducidas en dimensión mediante PCA. Esto permite mantener la información clave, los valores categóricos, sin introducir un exceso de dimensiones que puedan afectar la calidad de la agrupación.

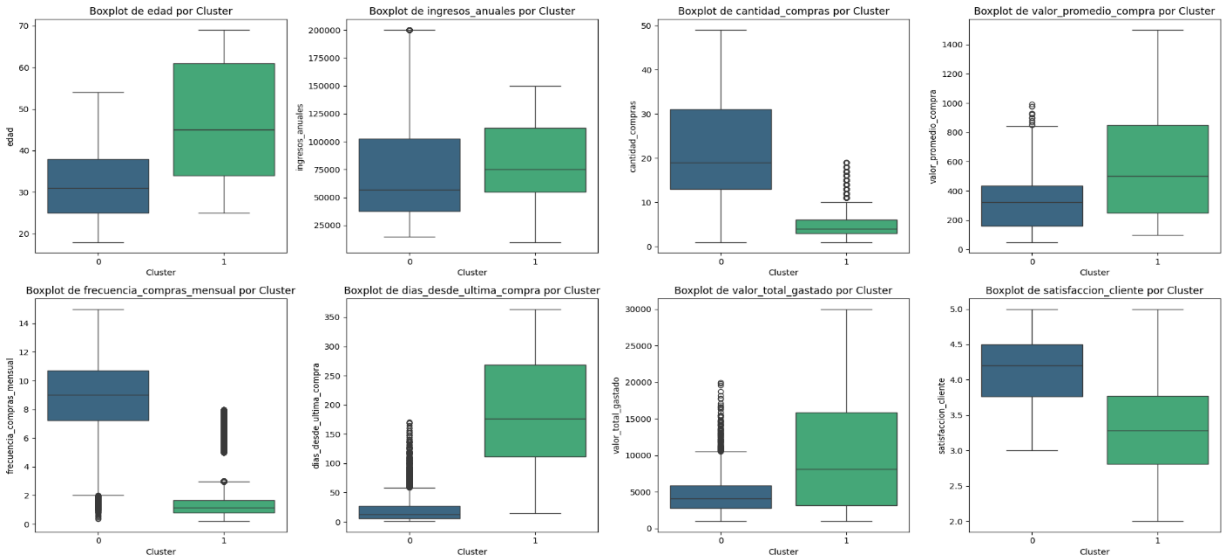
Sin embargo, una limitación de esta metodología es que, al aplicar PCA, la interpretación de los clústeres se vuelve menos clara. Específicamente, no será posible identificar directamente qué

variables categóricas influyeron en la asignación de los clústeres, ya que la información quedará condensada en las nuevas dimensiones generadas por el PCA.

Por esta razón, con el objetivo de garantizar una categorización más transparente y facilitar el análisis de la composición de los clústeres, se optará por utilizar el dataframe sin reducción de dimensionalidad.

## Análisis de Clústeres

Después de aplicar **K-Means** con un número de clusters igual a 2, se identifican los siguientes resultados.

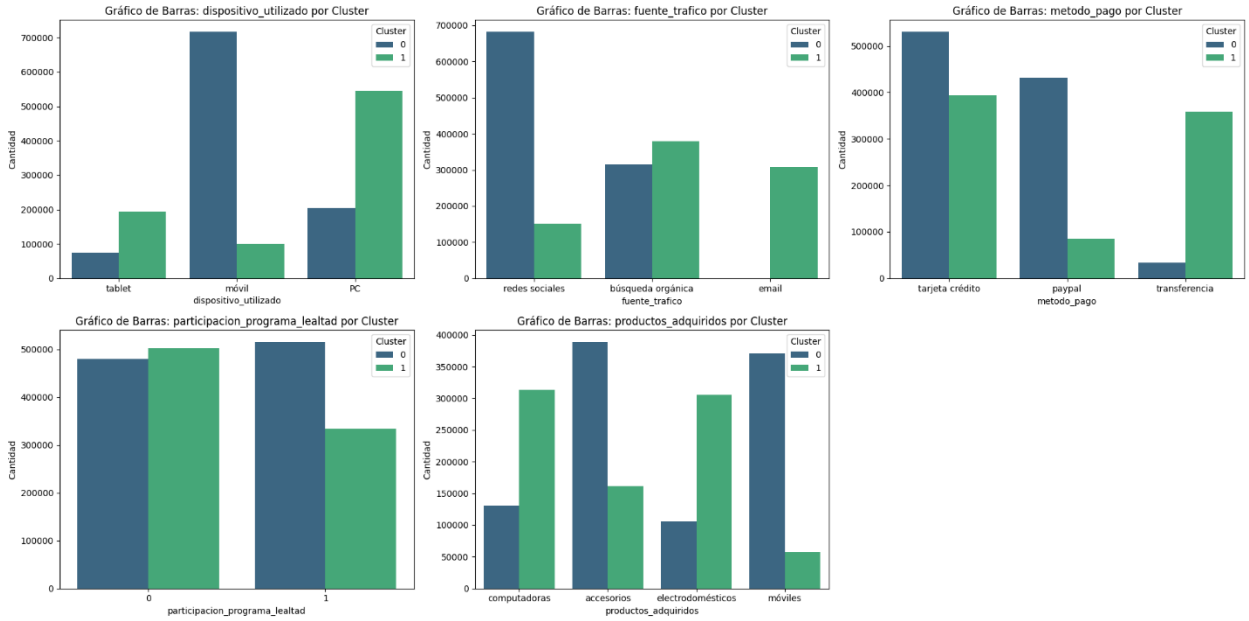



El Cluster 0 está conformado principalmente por compradores jóvenes, con una edad promedio de 31.89 años. Este grupo tiene una frecuencia alta de compras, realizando aproximadamente 9 compras mensuales en promedio, lo que demuestra un comportamiento activo y comprometido con la marca. Aunque sus ingresos anuales son relativamente moderados en comparación con el otro grupo (73,707 en promedio), su cantidad total de compras (22.13) refleja un claro patrón de consumo enfocado en compras pequeñas pero constantes. Este tipo de cliente tiende a generar ingresos recurrentes debido a su regularidad, aunque el valor promedio por compra (306.99) es más bajo en comparación con otros segmentos.

Además, el Cluster 0 se caracteriza por su alta satisfacción (4.13 en promedio), lo que indica que tienen una percepción positiva de su experiencia con la marca. Este grupo también muestra una participación moderada en los programas de lealtad, con un 51.77% de inclusión en estos programas, lo cual refuerza la idea de que son clientes relativamente leales y comprometidos. Estos hallazgos sugieren que este segmento representa una base sólida de clientes frecuentes y satisfechos, con un importante potencial para ser fidelizados aún más a través de estrategias como descuentos regulares o promociones por volumen.

Por otro lado, el Cluster 1 incluye a compradores de mayor edad, con un promedio de 46.97 años, lo que podría reflejar un público más maduro y posiblemente con un mayor poder adquisitivo. Este grupo realiza compras con baja frecuencia (1.3 compras mensuales en promedio) y han pasado 189 días desde su última compra, lo que los posiciona como clientes más esporádicos en comparación con el Cluster 0. A pesar de esta menor frecuencia, el Cluster 1 destaca por su alto gasto total acumulado (10,090.49 en promedio) y un valor promedio por compra significativamente mayor ($576.66). Esto sugiere que este segmento tiende a realizar menos transacciones, pero estas son de alto valor, aportando significativamente a los ingresos totales.

En términos de satisfacción, este segmento se encuentra algo rezagado en comparación con el Cluster 0, con un promedio de 3.32, lo que podría ser un área de oportunidad para mejorar su experiencia y, potencialmente, su compromiso con la marca. Asimismo, su participación en los programas de lealtad es menor (39.95%), lo que indica que, aunque son clientes valiosos, no necesariamente tienen un fuerte vínculo con la marca.




En términos de dispositivo utilizado, el Cluster 0 muestra una marcada preferencia por el uso de dispositivos móviles, representando más del 71% de las compras en este grupo. Esto refuerza la

percepción de que estos compradores, más jóvenes y con una frecuencia alta de compras mensuales, priorizan la conveniencia y rapidez que ofrecen los dispositivos móviles. Este comportamiento está en línea con su perfil dinámico, caracterizado por un mayor nivel de interacción con la marca y una mayor actividad de compra. En contraste, el Cluster 1 está dominado por usuarios de PC (65%) y tabletas (23%), lo que sugiere un enfoque más metódico y tradicional en sus decisiones de compra. Estos clientes, al realizar compras de alto valor con menor frecuencia, parecen preferir interfaces más detalladas y seguras, posiblemente relacionadas con la naturaleza más planificada y reflexiva de sus transacciones.

Al analizar la fuente de tráfico, las diferencias entre ambos clusters son igualmente notables. El Cluster 0 tiende a interactuar principalmente a través de redes sociales (68%), lo cual es coherente con su perfil juvenil y digitalmente activo. Estas plataformas parecen ser un canal clave para captar su atención y fomentar su comportamiento recurrente. Por otro lado, el Cluster 1 muestra una mayor dependencia de búsquedas orgánicas (45%) y campañas de correo electrónico (37%), lo que refleja un proceso de compra más investigativo y un enfoque en la información previa a la decisión. La baja participación en redes sociales (18%) en este grupo indica una menor afinidad con estrategias de marketing digital directo, lo que sugiere que campañas de email marketing bien diseñadas podrían ser más efectivas para atraerlos.

En cuanto a los métodos de pago, los clientes del Cluster 0 utilizan predominantemente tarjetas de crédito (53%) y PayPal (43%), métodos asociados con la rapidez y facilidad en las transacciones. Esto se alinea con su alto volumen de compras y su preferencia por métodos ágiles que respalden su comportamiento frecuente. En contraste, el Cluster 1 muestra una mayor inclinación por transferencias bancarias (43%), seguido de tarjetas de crédito (47%), reflejando una preferencia por opciones de pago más tradicionales y seguras, especialmente considerando el valor significativamente mayor de sus compras. Este patrón está estrechamente relacionado con su perfil más maduro y cauteloso.

En términos de participación en programas de lealtad, el Cluster 0 presenta una mayor proporción de clientes inscritos (52%), lo que refuerza su compromiso con la marca y su predisposición a aprovechar beneficios continuos. Sin embargo, el Cluster 1, a pesar de su alto poder adquisitivo, tiene una participación menor (40%), lo que representa una oportunidad importante para fortalecer la conexión con este grupo. Dado su menor frecuencia de compra, incentivar su inclusión en estos programas podría fomentar una mayor lealtad y aumentar la recurrencia en sus transacciones.

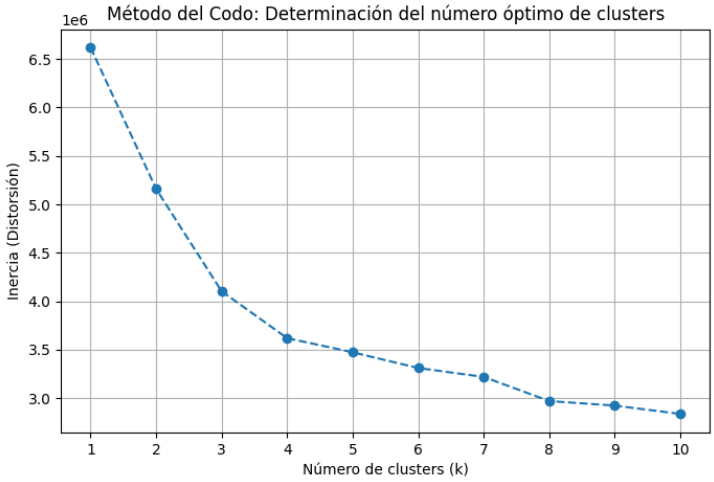
Finalmente, los productos adquiridos reflejan aún más estas diferencias. En el Cluster 0, los productos principales incluyen accesorios (39%) y móviles (37%), categorías asociadas con un comportamiento de compra más frecuente, impulsado por tendencias o necesidades inmediatas. En contraste, el Cluster 1 se orienta hacia productos más costosos y menos frecuentes como computadoras (37%) y electrodomésticos (36%), lo que refuerza su perfil de cliente que valora

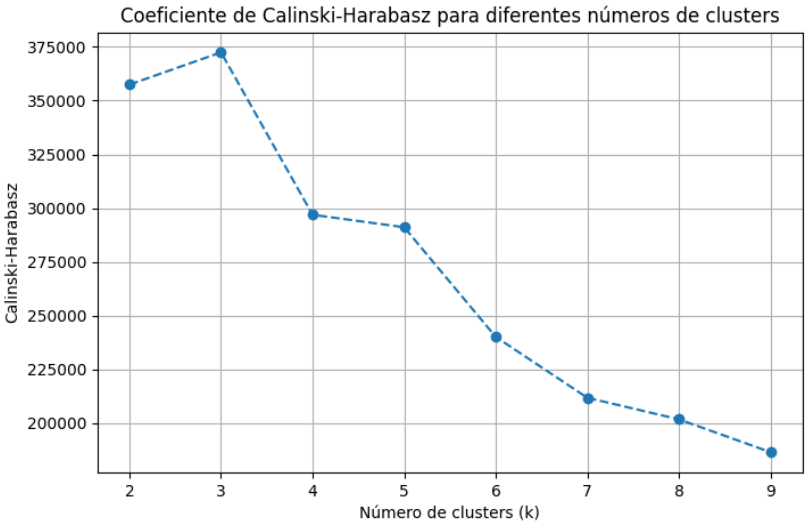
transacciones significativas y planeadas. Esto también explica su menor satisfacción promedio, ya que las expectativas tienden a ser más altas cuando se realizan compras de mayor valor.

## Cluster 0 - Compradores dinámicos y digitales

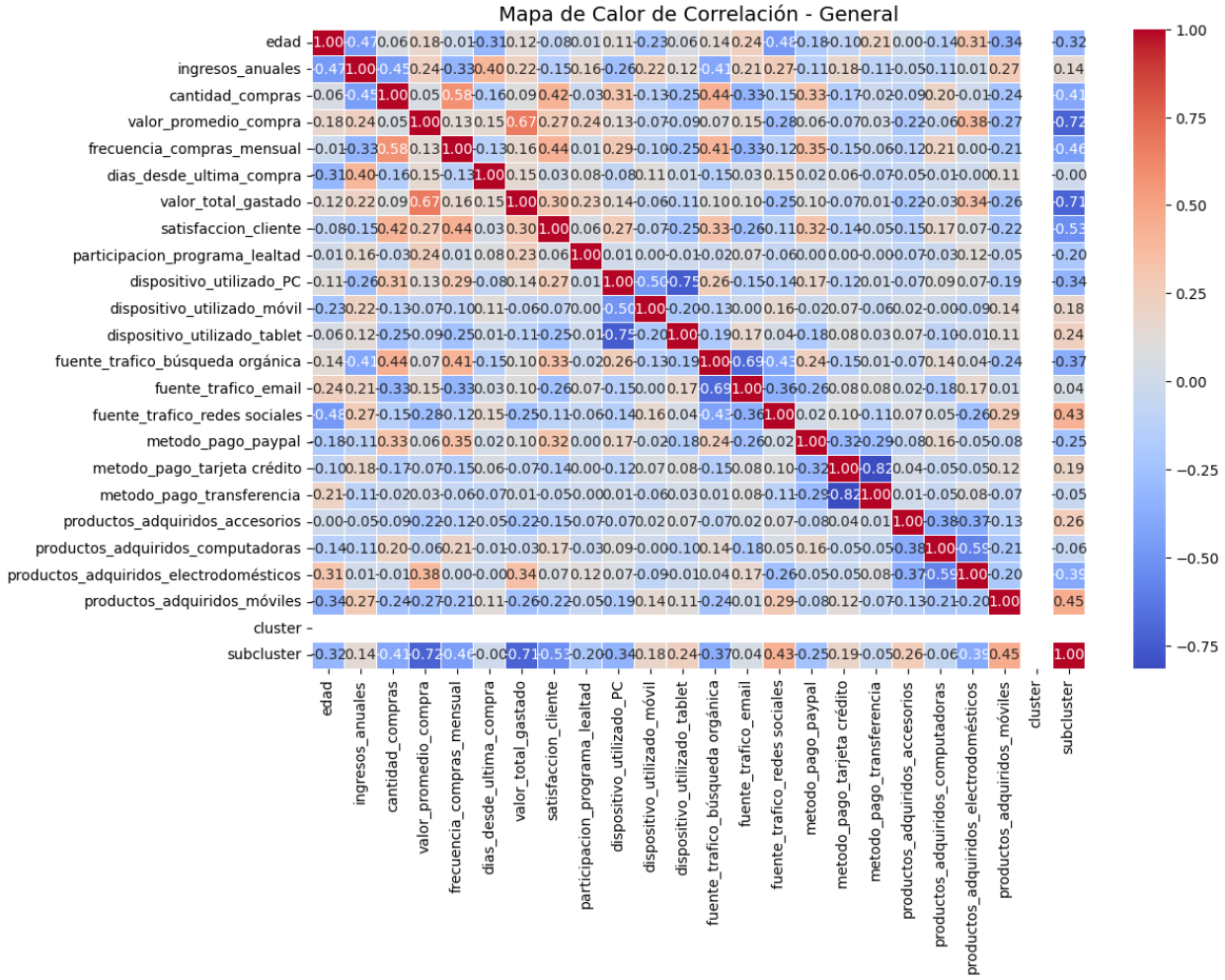
Para poder identificar mayor separación entre los clústeres y hacer el estudio de manera más granular, se decidió hacer el estudio individual de cada clúster y a estos aplicarle su propia metodología de K-means, usando los mismos pasos y criterios tomados anteriormente.

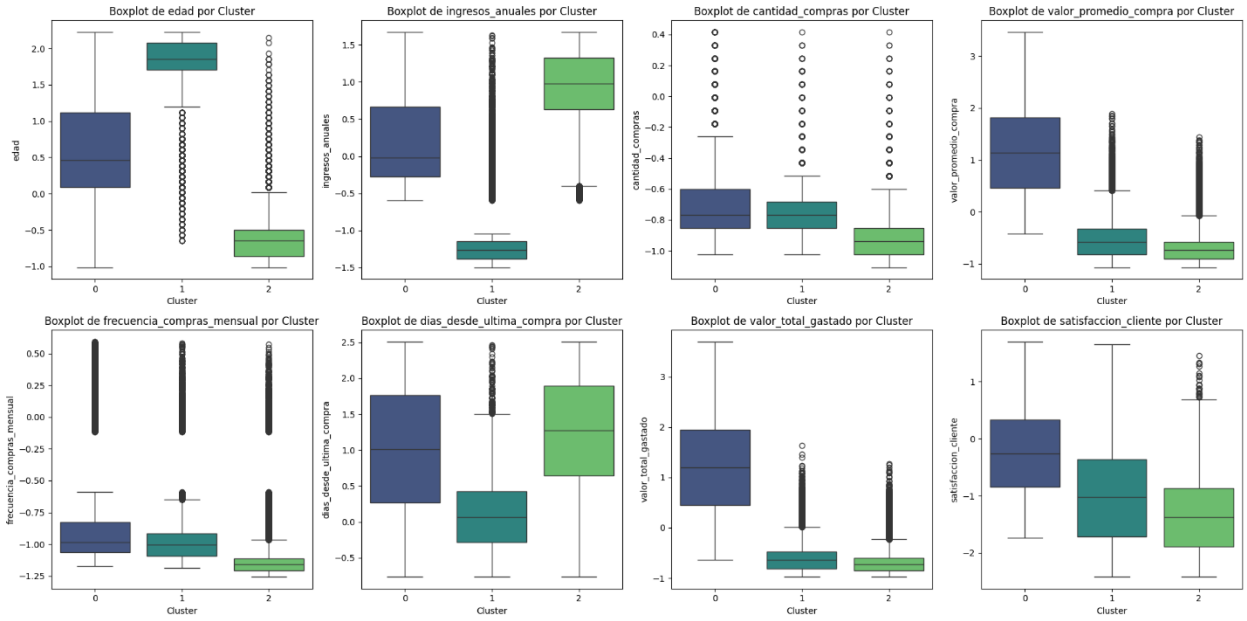
Para entender mejor los grupos segmentados entre los compradores dinámicos y digitales , se realizaron subclusters dentro de cada cluster principal utilizando técnicas adicionales de agrupamiento. Esto permitió identificar patrones más específicos y diferencias internas dentro de los clusters, ofreciendo una visión aún más detallada de los comportamientos y características únicas de cada segmento, lo cual facilita la implementación de estrategias más precisas y personalizadas. Esto, nos puede ayudar a identificar las tendencias y comportamientos entre las personas cuyas principales características son  ser compradores dinámicos y digitales.

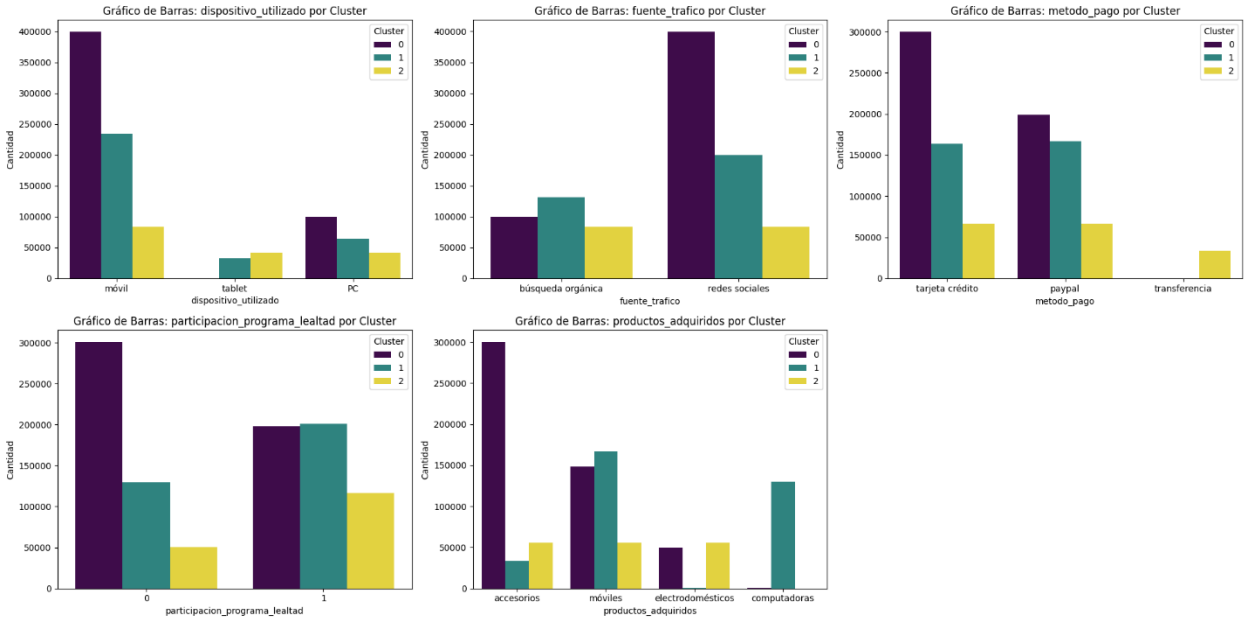





En este caso el coeficiente de Calinski-Harabasz, y el método del codo coinciden en que el número óptimo de clusters para este análisis es 3. Y por lo tanto, se recomienda utilizar 3 clústeres como la opción óptima.

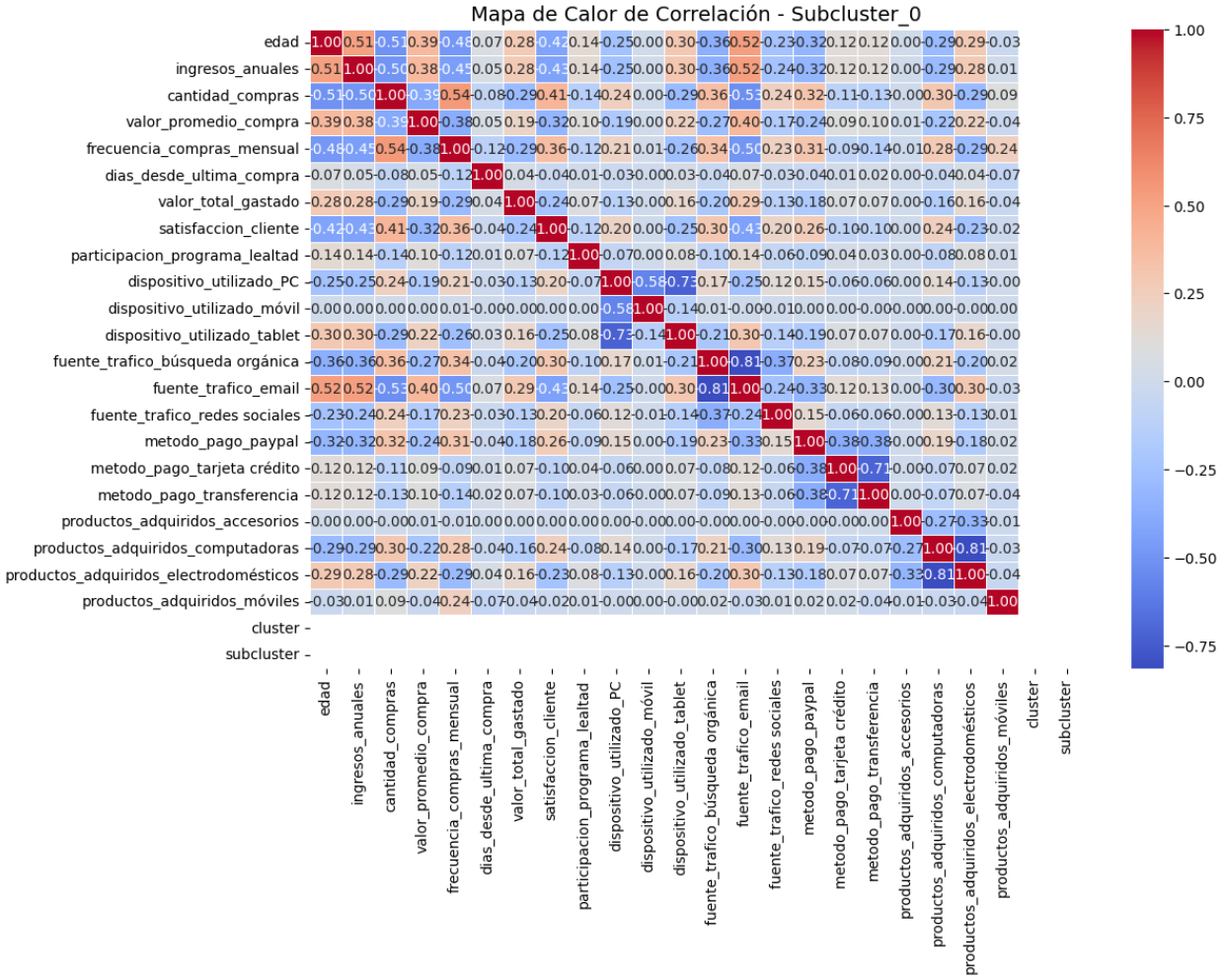

Mapa de Calor de Correlación - General
edad
ingresos_anuales
cantidad_compras
valor_promedio_compra
frecuencia_compras_mensual
dias_desde_ultima_compra
valor_total_gastado
satisfaccion_cliente
participacion_programa_lealtad
dispositivo_utilizado_PC
dispositivo_utilizado_móvil
dispositivo_utilizado_tablet
fuente_trafico_búsqueda orgánica
fuente_trafico_email
fuente_trafico_redes sociales
metodo_pago_paypal
metodo_pago_tarjeta crédito
metodo_pago_transferencia
productos_adquiridos_accesorios
productos_adquiridos_computadoras
productos_adquiridos_electrodomésticos
productos_adquiridos_móviles
cluster
subcluster
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75



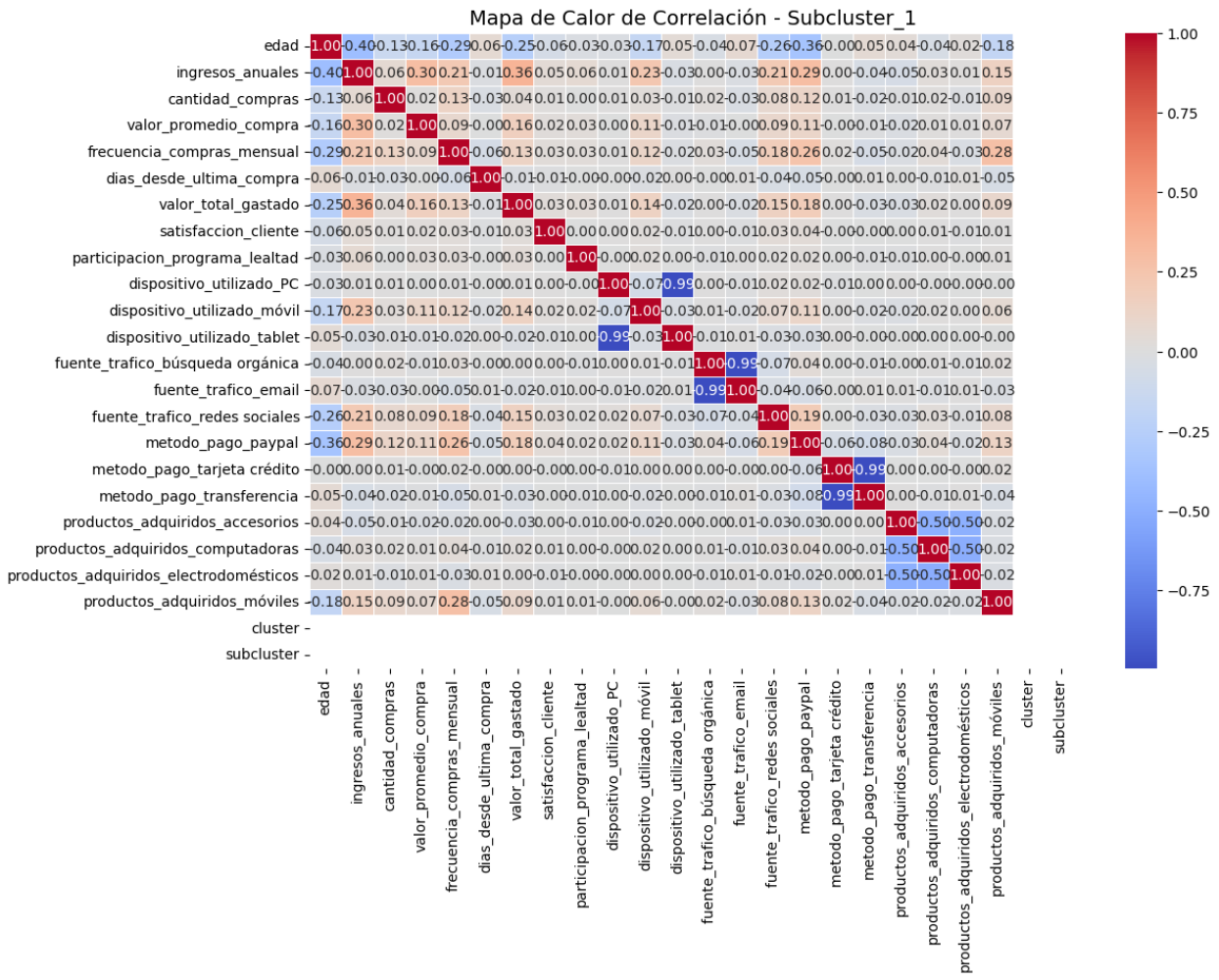

A primera vista podemos encontrar una tendencia a realizar compras pequeñas, desde las redes sociales y mayoritariamente mediante celulares. Siendo una relación directa con la facilidad que representa comprar con una tarjeta de crédito; sintiéndose más seguro con respecto mejor experiencia haya tenido con respecto al producto.

### *Sub Cluster 0*



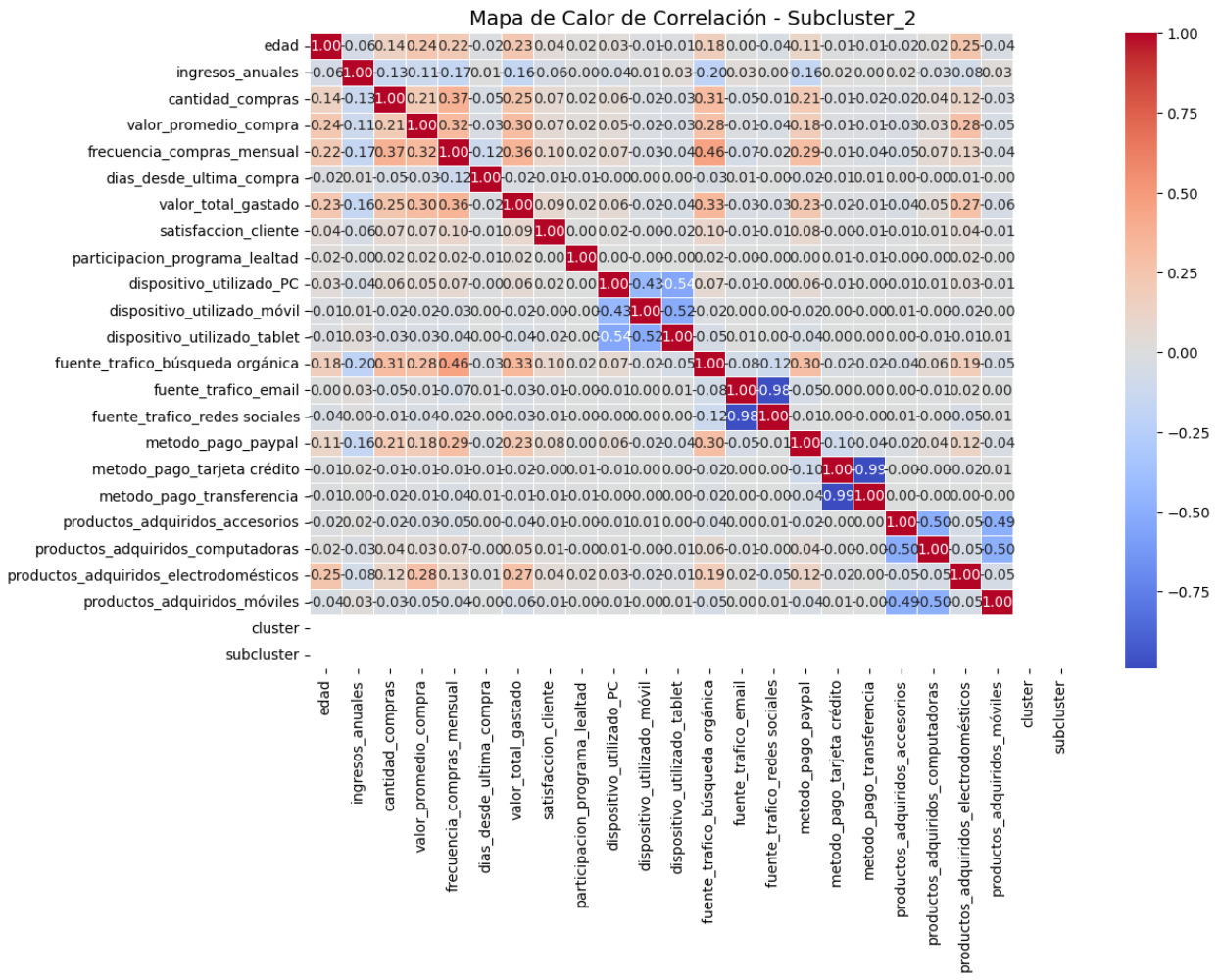

Edad promedio: 26 años | Ingreso anual: 37533) Este grupo representa a los compradores más jóvenes del Cluster 0, caracterizados por su ritmo acelerado y alta interacción con la marca. Realizan 24 compras al mes, con casi 10 transacciones mensuales, lo que los convierte en compradores frecuentes e impulsivos. Su gasto promedio por transacción es de 274, acumulando un total mensual de 3003. Utilizan predominantemente dispositivos móviles (80 %) y confían en las redes sociales (80 %) como su principal fuente de inspiración para comprar. Esto refleja una fuerte conexión con tendencias y productos de fácil acceso, siendo los accesorios (60 %) y los móviles (30 %) sus categorías favoritas. A pesar de estar altamente satisfechos (4.5), solo un 39.6 % participa en programas de lealtad, lo que deja espacio para estrategias dirigidas a fomentar un mayor compromiso con la marca.

## *Sub Cluster 1*




(Edad promedio: 34 años | Ingreso anual: 84748) Este subgrupo está compuesto por compradores adultos jóvenes que combinan dinamismo con un enfoque más equilibrado en sus decisiones. Compran 13 veces al mes, con 6.5 transacciones mensuales, pero destacan por un gasto promedio elevado de 448 por compra, alcanzando un total mensual de 6463. Prefieren productos más relevantes como móviles (50 %) y computadoras (39 %), lo que indica un interés por bienes de mayor valor. Aunque los dispositivos móviles lideran su uso (70 %), las tabletas (10 %) tienen una presencia significativa, mostrando cierto grado de sofisticación tecnológica. Este grupo no solo confía en redes sociales (60 %), sino que también emplea búsquedas orgánicas (40 %) para investigar antes de comprar. Con una satisfacción moderada (3.76) y un 60.8 % de participación en programas de lealtad, este subcluster tiene gran potencial para campañas que destaquen exclusividad y beneficios para compradores selectivos.

## *Sub Cluster 2*




(Edad promedio: 44 años | Ingreso anual: 160068) El perfil de este subgrupo incluye a consumidores maduros con el ingreso anual más alto dentro del Cluster 0. Realizan un promedio de 35 compras al mes, con 11 transacciones mensuales, pero su gasto promedio por compra es el más bajo (125), totalizando 4994 al mes. Esto sugiere un patrón de compras recurrentes y diversificado. Este grupo equilibra el uso de dispositivos: móviles (50 %), PC (25 %) y tabletas (25 %), reflejando flexibilidad tecnológica. Sus fuentes de tráfico están equilibradas entre redes sociales y búsquedas orgánicas (50 % cada una), indicando una estrategia híbrida en su interacción digital. Adquieren una variedad de productos en proporciones similares: accesorios, electrodomésticos y móviles (33 % cada uno). Con una alta participación en programas de lealtad (69.9 %) pero una satisfacción moderada (3.75), este subgrupo representa una oportunidad para personalizar ofertas y mejorar la experiencia del cliente

En general, los subclusters dentro del Cluster 0 presentan características diferenciadas que permiten entender mejor los perfiles de estos consumidores. El Subcluster 0 se compone principalmente de personas jóvenes con ingresos anuales relativamente bajos. Su comportamiento de compra es marcado por una alta frecuencia mensual y transacciones de valor moderado. Este grupo tiene una clara inclinación por el uso de dispositivos móviles y redes sociales, reflejando su perfil dinámico y su preferencia por la rapidez y conveniencia. Los productos más adquiridos son accesorios y móviles, lo cual está alineado con su estilo de vida digital. Sin embargo, su participación en programas de lealtad es baja, lo que podría representar una oportunidad para fidelizar aún más a estos compradores recurrentes.

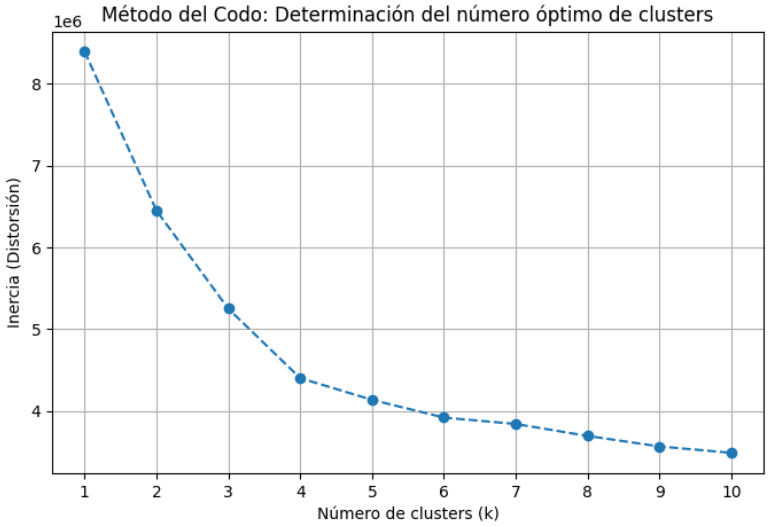
El Subcluster 1 está compuesto por individuos adultos jóvenes con ingresos medios-altos. Aunque realizan compras menos frecuentes, el valor promedio por transacción es notablemente alto, lo que indica un enfoque más reflexivo y planificado. Este grupo diversifica su uso entre móviles y tabletas, mostrando un perfil más sofisticado en términos tecnológicos. Además, aunque redes sociales siguen siendo relevantes, hay una mayor proporción de interacción a través de búsquedas orgánicas. Los productos que destacan son móviles y computadoras, reflejando la preferencia por bienes de mayor valor. La satisfacción del cliente en este grupo es moderada, pero su participación en programas de lealtad es considerablemente alta, lo cual sugiere que las estrategias de fidelización pueden ser particularmente efectivas para este perfil.

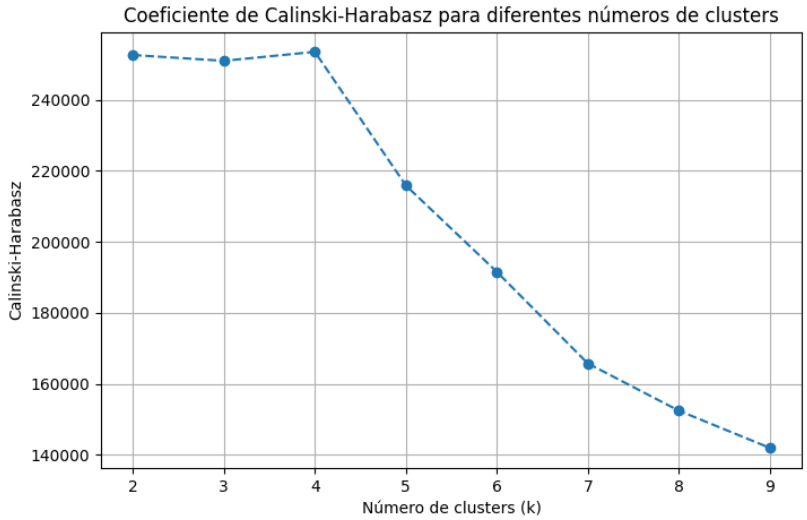
Por último, el Subcluster 2 se caracteriza por un perfil más maduro y con ingresos altos. A pesar de ser el grupo con mayor frecuencia de compras mensuales, el valor promedio de sus transacciones es bajo, lo que podría reflejar interés en compras recurrentes de menor valor. Este subcluster utiliza una mezcla equilibrada de dispositivos como móviles, PC y tabletas, lo que denota flexibilidad tecnológica. En términos de tráfico, este grupo muestra igual interés por redes sociales y búsquedas orgánicas, lo que abre oportunidades para estrategias híbridas de marketing. Los productos adquiridos están distribuidos equitativamente entre accesorios, electrodomésticos y móviles, reflejando una preferencia diversificada. Aunque la satisfacción promedio es moderada, la participación en programas de lealtad es alta, lo cual fortalece su vínculo con la marca.

| **Tipo de Correlación** | **Subcluster 0 y 1** | **Subcluster 0 y 2** | **Subcluster 1 y 2** | **Todos los Subclusters** |
|---|---|---|---|---|

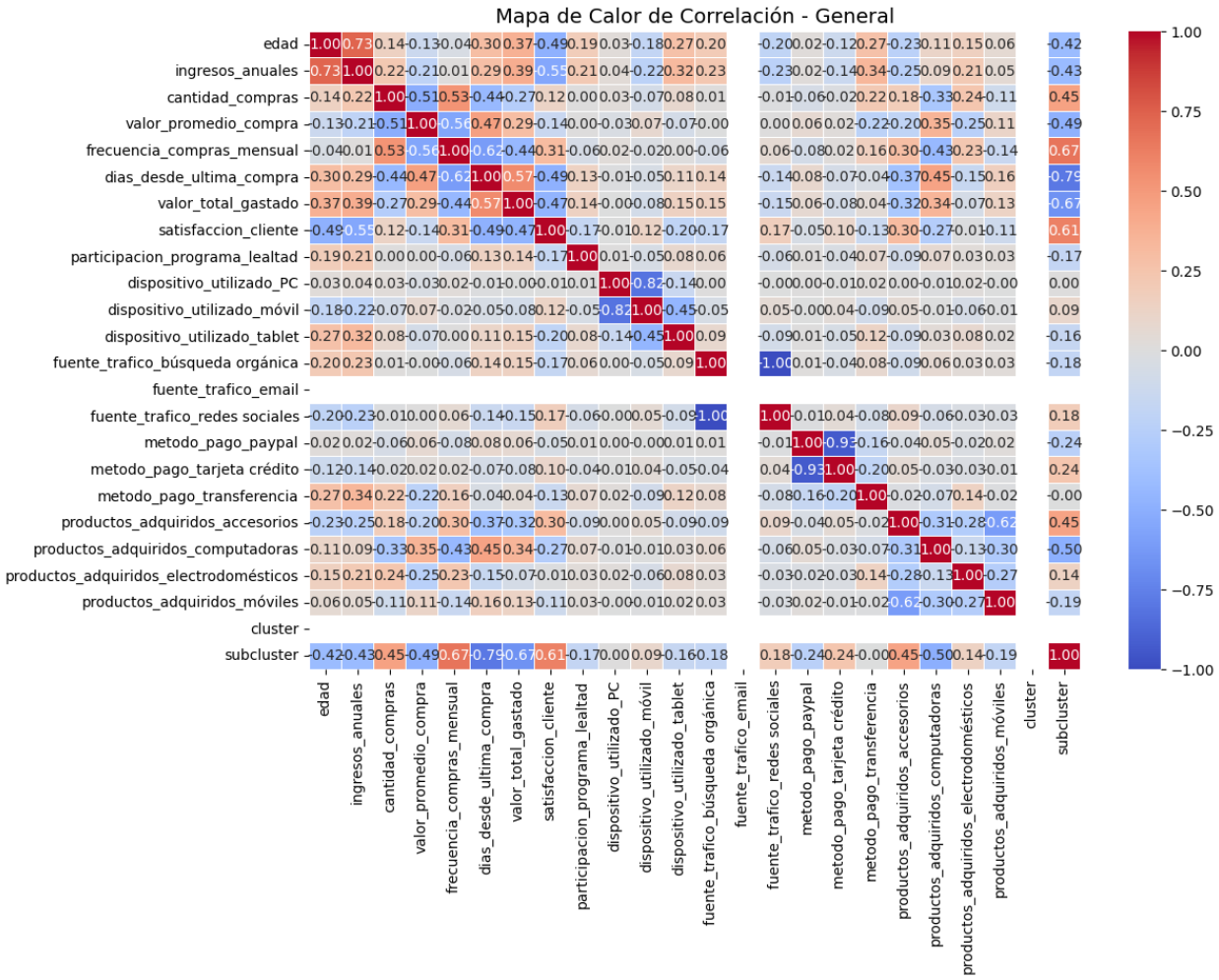
| | | | | |
|---|---|---|---|---|
| **Fuertes (0.7 <=corr)** | ('metodo_pago_tarjeta crédito', 'metodo_pago_transferencia'), ('fuente_trafico_búsqueda orgánica', 'fuente_trafico_email'), ('dispositivo_utilizado_tablet', 'dispositivo_utilizado_PC'), ('dispositivo_utilizado_PC', 'dispositivo_utilizado_tablet'), ('fuente_trafico_email', 'fuente_trafico_búsqueda orgánica'), ('metodo_pago_transferencia', 'metodo_pago_tarjeta crédito') | ('metodo_pago_transferencia', 'metodo_pago_tarjeta crédito'), ('metodo_pago_tarjeta crédito', 'metodo_pago_transferencia') | ('metodo_pago_transferencia', 'metodo_pago_tarjeta crédito'), ('metodo_pago_tarjeta crédito', 'metodo_pago_transferencia') | ('metodo_pago_transferencia', 'metodo_pago_tarjeta crédito'), ('metodo_pago_tarjeta crédito', 'metodo_pago_transferencia') |
| **Moderadas(0.5 <=corr <0.7)** | No hay | No hay | No hay | No hay |
| **Débiles (0.3 <=corr <0.5)** | ('productos_adquiridos_electrodomésticos', 'productos_adquiridos_accesorios'), ('metodo_pago_paypal', 'edad'), ('productos_adquiridos_accesorios', 'productos_adquiridos_electrodomésticos'), ('edad', 'metodo_pago_paypal') | ('cantidad_compras', 'fuente_trafico_búsqueda orgánica'), ('frecuencia_compras_mensual', 'valor_promedio_compra'), ('fuente_trafico_búsqueda orgánica', 'frecuencia_compras_mensual'), ('frecuencia_compras_mensual', 'fuente_trafico_búsqueda orgánica'), ('valor_promedio_compra', 'frecuencia_compras_mensual'), ('fuente_trafico_búsqueda orgánica', 'cantidad_compras') | No hay | No hay |

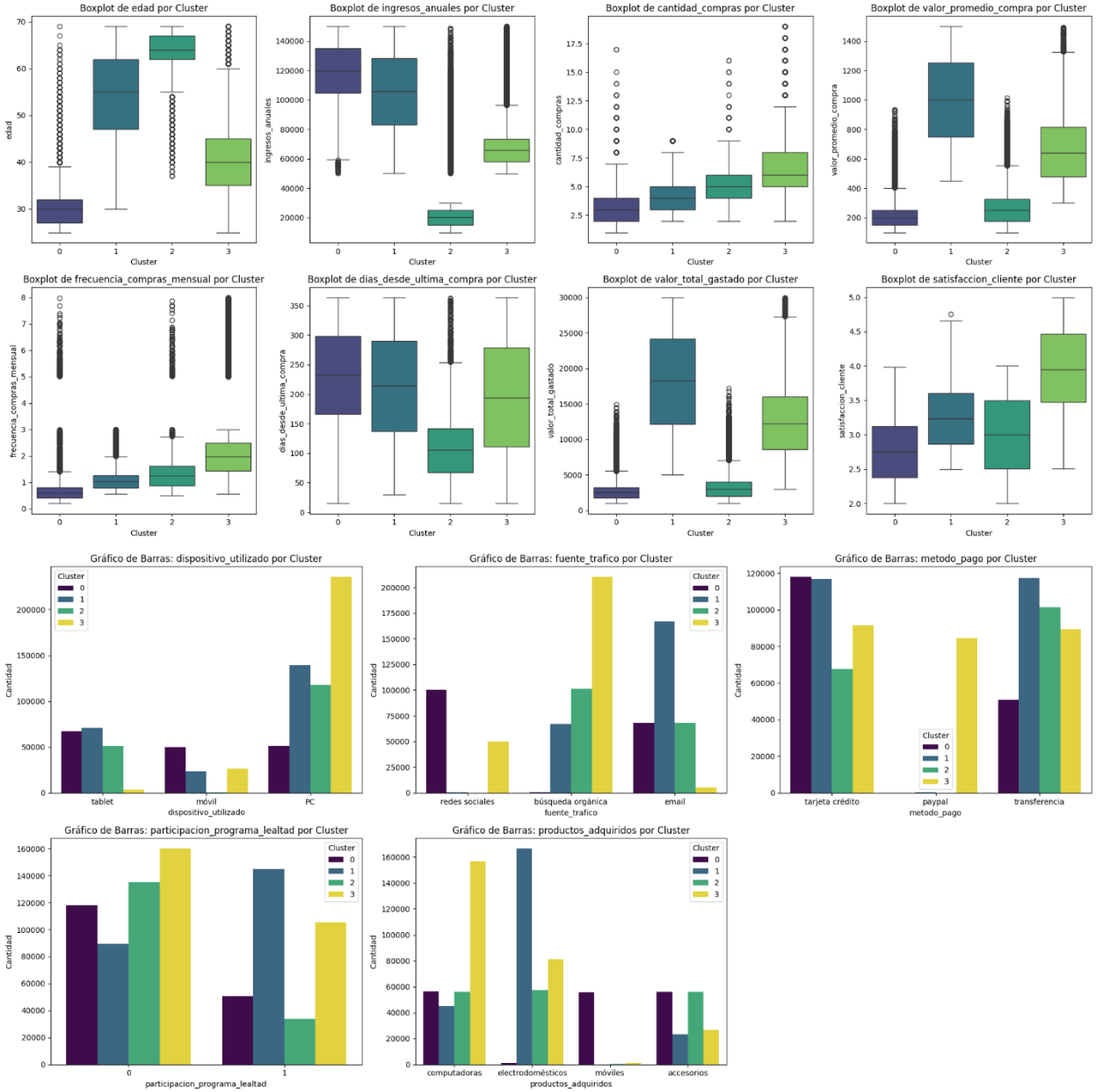
## Cluster 1- Compradores reflexivos y tradicionales



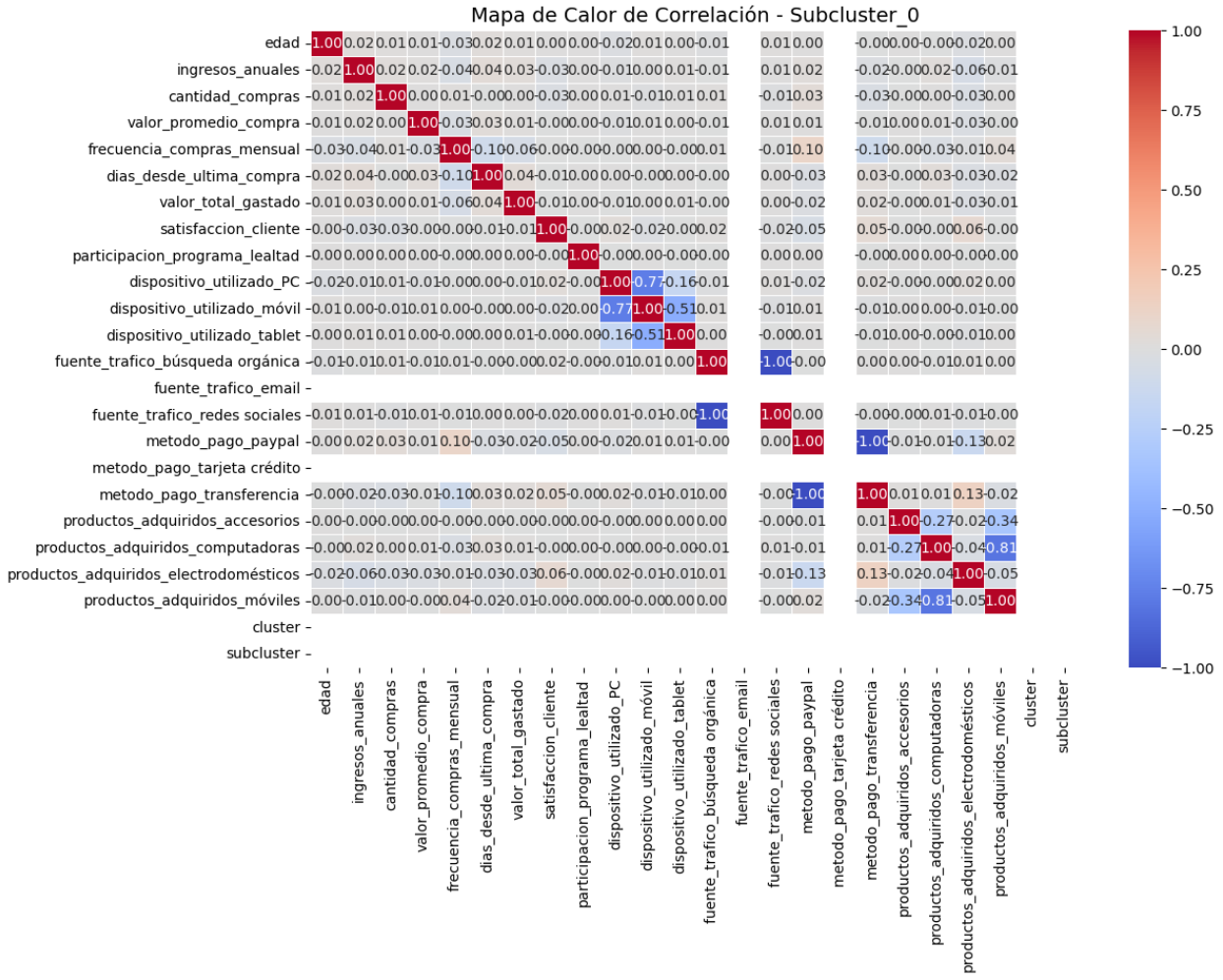

En este caso el valor de k(número de clusters óptimo) es de 4.


Mapa de Calor de Correlación - General
edad
ingresos_anuales
cantidad_compras
valor_promedio_compra
frecuencia_compras_mensual
dias_desde_ultima_compra
valor_total_gastado
satisfaccion_cliente
participacion_programa_lealtad
dispositivo_utilizado_PC
dispositivo_utilizado_móvil
dispositivo_utilizado_tablet
fuente_trafico_búsqueda orgánica
fuente_trafico_email
fuente_trafico_redes sociales
metodo_pago_paypal
metodo_pago_tarjeta crédito
metodo_pago_transferencia
productos_adquiridos_accesorios
productos_adquiridos_computadoras
productos_adquiridos_electrodomésticos
productos_adquiridos_móviles
cluster
subcluster
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00




Con ayuda de las anteriores gráficas, podemos encontrar medianamente patrones definidos entre las personas que hacen uso de sus celulares para la compra, tal como con sus patrones de búsqueda y su interés en tomar un buen tiempo para tomar la decisión sobre sus compras

## *Sub Clúster 0*

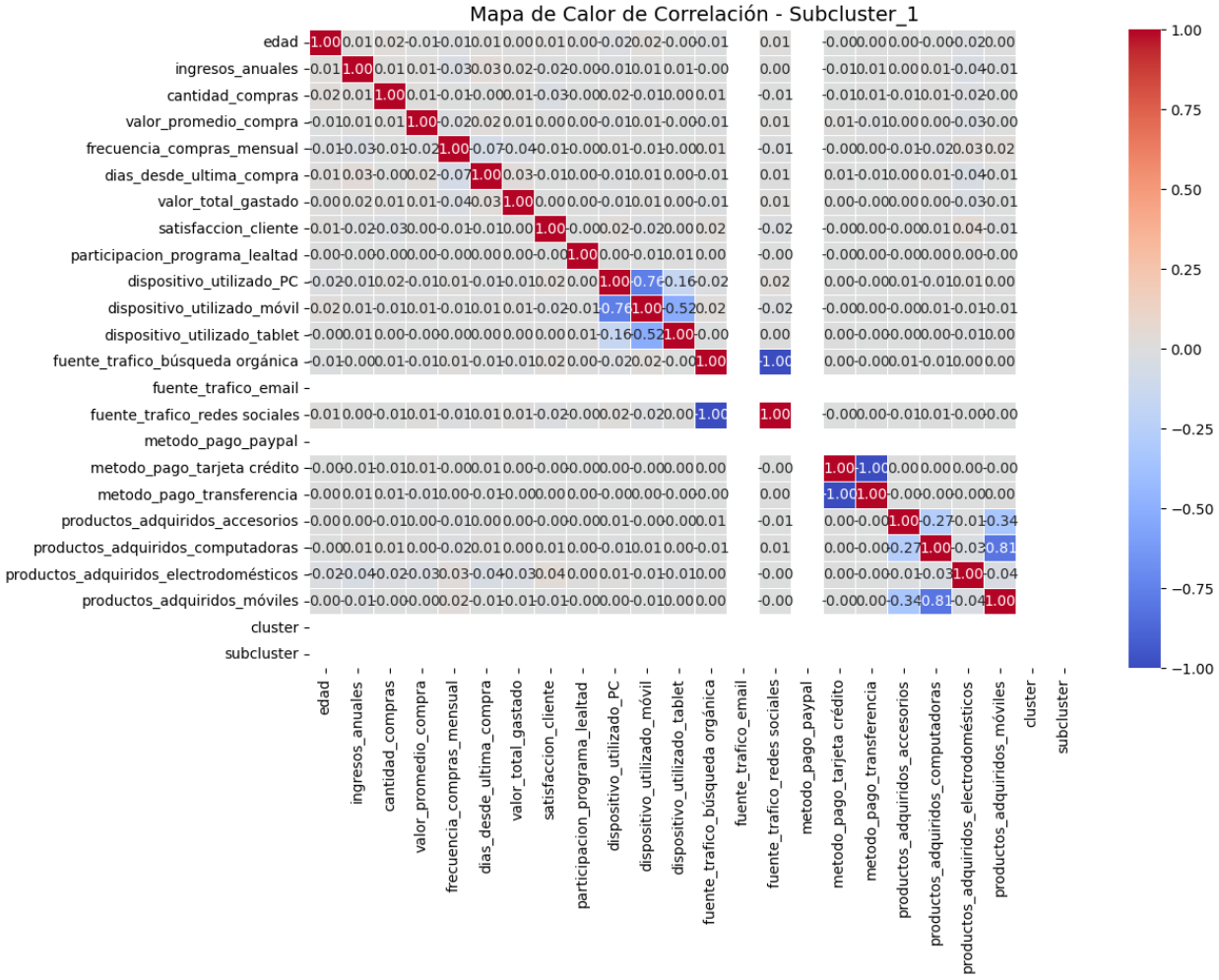



(Edad promedio: 29 años | Ingreso anual: 119849) Este grupo es el más joven dentro del Cluster 1, pero, a pesar de sus altos ingresos, tienen un comportamiento de compra relativamente pasivo. Realizan 2.5 compras al mes, con una frecuencia mensual baja (0.6 transacciones) y un gasto promedio de 203 por compra, que suma un total mensual de 2551. Confían principalmente en tabletas (40 %) y redes sociales (59 %) como canal de inspiración, seguido por email marketing (40 %). Adquieren productos en proporciones similares: accesorios, computadoras y móviles. Su satisfacción es baja (2.75), y solo un 29.9 % participa en programas de lealtad, lo que indica que se necesita un enfoque renovado para captar su interés y mejorar su experiencia.

En general, está compuesto por clientes que utilizan tanto dispositivos móviles como computadoras para realizar compras y tienen una alta participación en programas de lealtad. Sus ingresos y gastos son moderados, y muestran una preferencia por métodos de pago

digitales como PayPal. Su comportamiento indica que buscan comodidad y seguridad en sus transacciones, además de estar influenciados por redes sociales como fuente de tráfico principal.

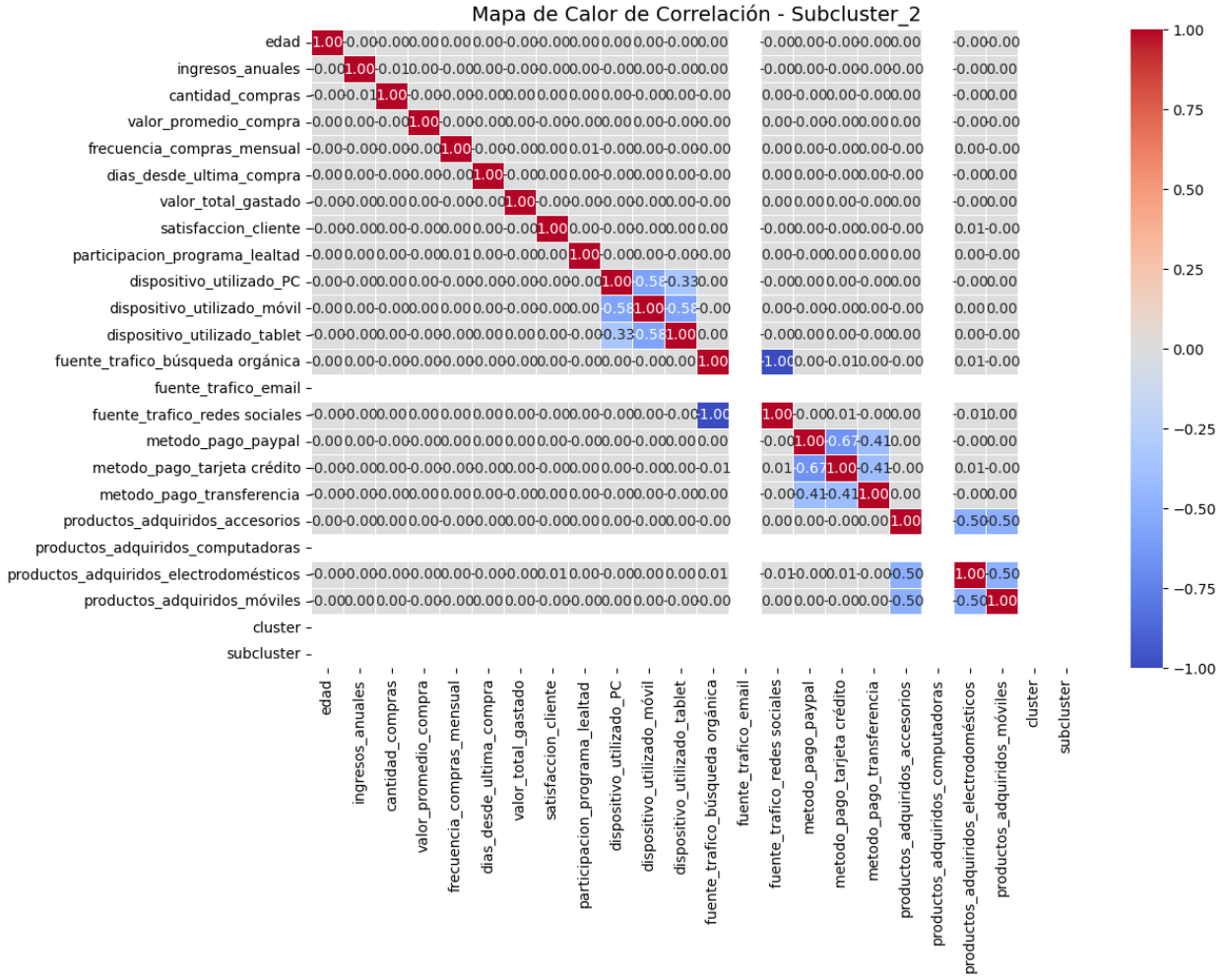
### *Sub Cluster 1*




(Edad promedio: 54 años | Ingreso anual: 105577) Este perfil incluye a consumidores mayores que realizan 3.5 compras al mes, con una frecuencia mensual de 1 transacción. Destacan por su gasto promedio elevado de 998 por compra, acumulando 18058 al mes. Prefieren bienes como electrodomésticos (71 %) y computadoras (19 %). Usan principalmente PC (59 %) y confían en email marketing (71 %) como su canal principal. La satisfacción promedio es moderada (3.23) y su participación en programas de lealtad alcanza el 61.9 %, lo que sugiere que una comunicación detallada y beneficios exclusivos podrían fortalecer aún más su vínculo con la marca.

Este subcluster se asemeja al primero en varios aspectos, pero con una diferencia clave en los métodos de pago, ya que estos clientes prefieren utilizar tarjetas de crédito en lugar de plataformas digitales. Esto sugiere que pueden tener un perfil financiero más estable o que buscan aprovechar beneficios adicionales como acumulación de puntos o pagos diferidos.

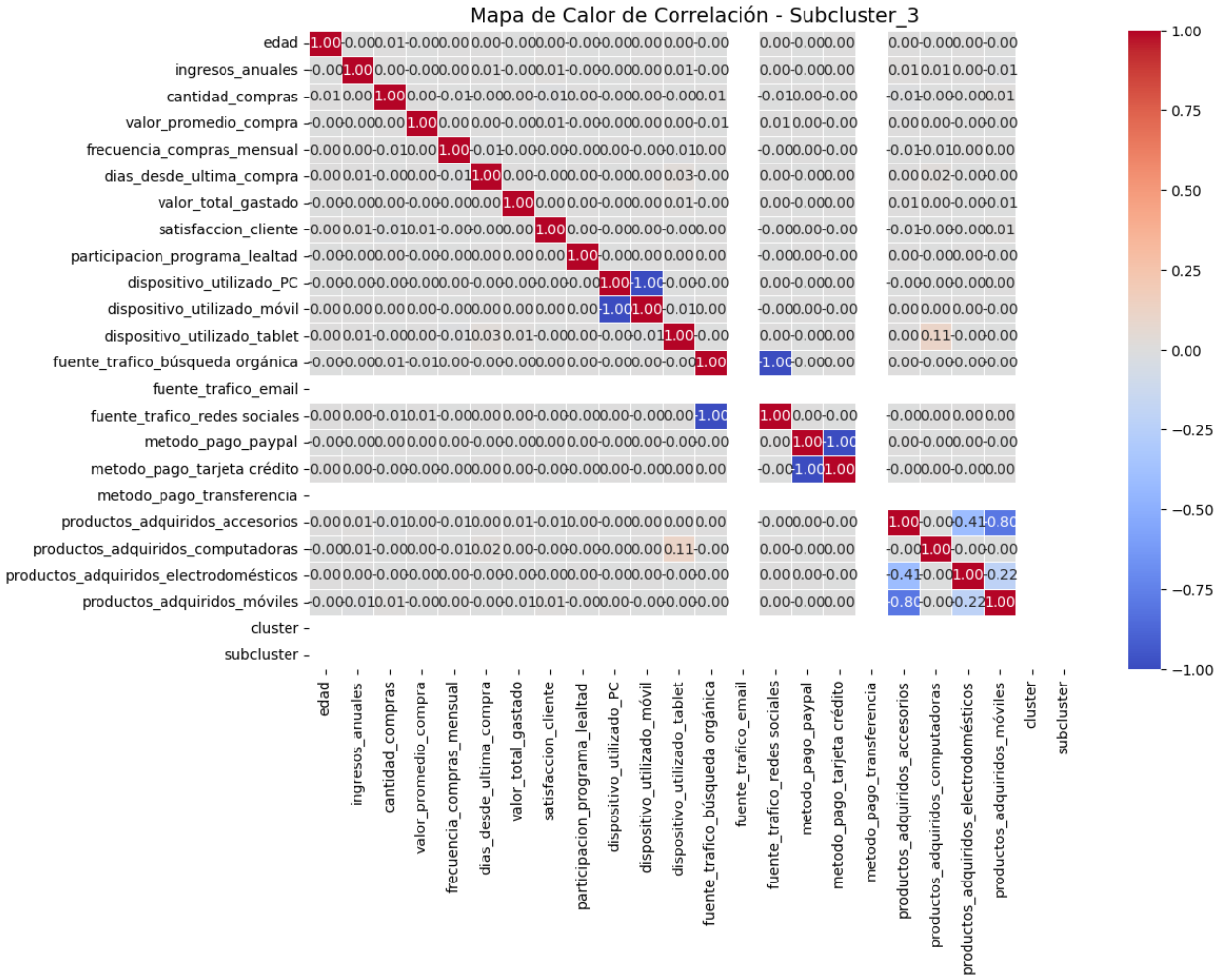
### *Sub Cluster 2*




(Edad promedio: 64 años | Ingreso anual: 20832) Este grupo de consumidores mayores tiene el ingreso más bajo dentro del Cluster 1, pero realizan más compras (5 al mes) y tienen una frecuencia mensual de 1.25 transacciones. Con un gasto promedio de 254 por compra, suman un total mensual de 3061. Sus intereses principales son los electrodomésticos (34 %) y las computadoras (33 %). Utilizan principalmente PC (69 %) y confían en búsquedas orgánicas (59 %) para tomar decisiones. Su satisfacción es moderada (3.00), y su participación en programas de lealtad es baja (19.9 %), lo que representa una oportunidad para diseñar estrategias que los conecten mejor con la marca.

Este grupo toma clientes con ingresos significativamente más altos, aunque con una menor frecuencia de compra. Sin embargo, cuando realizan una transacción, el valor individual de sus compras es mayor y suelen adquirir una mayor variedad de productos, incluyendo electrodomésticos y accesorios. Aunque participan en programas de lealtad, no muestran tanta dependencia de ellos como los primeros grupos. Sus métodos de pago son diversos, combinando tanto tarjetas de crédito como plataformas digitales, lo que sugiere un enfoque más flexible y estratégico al momento de comprar.

### *Sub Cluster 3*




(Edad promedio: 39 años | Ingreso anual: 66755) Este grupo de compradores adultos realiza 6.4 compras al mes, con una frecuencia mensual de 2 transacciones. Con un gasto promedio de 647

por compra, acumulan un total mensual de 12319. Prefieren computadoras (59 %) y electrodomésticos (31 %) como sus principales productos. Usan principalmente PC (89 %) y se apoyan en búsquedas orgánicas (79 %) como fuente clave de información antes de comprar. Su satisfacción es relativamente alta (3.96), y el 39.6 % participa en programas de lealtad, lo que indica que todavía hay oportunidades para mejorar su fidelidad a través de beneficios atractivos y personalizados.

Para este caso, se compone de clientes con ingresos más bajos en comparación con los demás, pero que realizan compras con mayor frecuencia, aunque en montos más pequeños. Presentan una alta satisfacción con su experiencia de compra y tienen una menor participación en programas de lealtad. Su comportamiento muestra una preferencia por métodos de pago más tradicionales y un consumo orientado principalmente a accesorios y dispositivos móviles.

| **Tipo de Correlación** | **Subcluster 0 y 1** | **Subcluster 0 y 2** | **Subcluster 0 y 3** | **Subcluster 1 y 2** | **Subcluster 1 y 3** | **Subcluster 2 y 3** | **TODOS** |
|---|---|---|---|---|---|---|---|
| **Fuertes** | ('dispositivo_utilizado_móvil', 'dispositivo_utilizado_PC'), ('productos_adquiridos_computadoras', 'productos_adquiridos_móviles'), ('fuente_trafico_búsqueda | ('fuente_trafico_búsqueda orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica') | ('dispositivo_utilizado_móvil', 'dispositivo_utilizado_PC'), ('dispositivo_utilizado_PC', 'dispositivo_utilizado_móvil'), ('fuente_trafico_búsqueda | ('fuente_trafico_búsqueda orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica') | ('dispositivo_utilizado_móvil', 'dispositivo_utilizado_PC'), ('dispositivo_utilizado_PC', 'dispositivo_utilizado_móvil'), ('fuente_trafico_búsqueda orgánica', | ('fuente_trafico_búsqueda orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda | ('fuente_trafico_búsqueda orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica') |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica'), ('dispositivo_utilizado_PC', 'dispositivo_utilizado_móvil'), ('productos_adquiridos_móviles', 'productos_adquiridos_computadoras') | orgánica', 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica') | | | 'fuente_trafico_redes sociales'), ('fuente_trafico_redes sociales', 'fuente_trafico_búsqueda orgánica') | orgánica') | |
| **Moderadas** | ('dispositivo_utilizado_móvil', 'dispositivo_utiliz | ('dispositivo_utilizado_móvil', 'dispositivo_utiliz | set() | ('dispositivo_utilizado_móvil', 'dispositivo_utiliza | set() | set() | set() |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| | ado_tablet'), ('dispositivo_utilizado_tablet', 'dispositivo_utilizado_móvil') | ado_tablet'), ('dispositivo_utilizado_tablet', 'dispositivo_utilizado_móvil') | | do_tablet'), ('dispositivo_utilizado_tablet', 'dispositivo_utilizado_móvil') | | | |
| **Débiles** | ('productos_adquiridos_accesorios', 'productos_adquiridos_móviles'), ('productos_adquiridos_móviles', 'productos_adquiridos_accesorios') | ('productos_adquiridos_accesorios', 'productos_adquiridos_móviles'), ('productos_adquiridos_móviles', 'productos_adquiridos_accesorios') | set() | ('productos_adquiridos_accesorios', 'productos_adquiridos_móviles'), ('productos_adquiridos_móviles', 'productos_adquiridos_accesorios') | set() | set() | set() |

Interpretación general

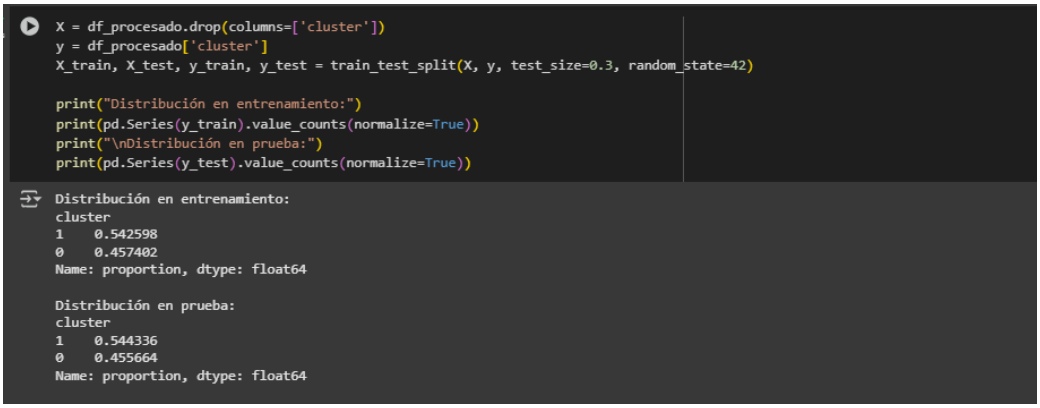
Los subclusters del Cluster 0 presentan un perfil predominantemente joven y digital, con comportamientos de compra frecuentes e impulsivos impulsados por móviles y redes sociales. Por otro lado, los subclusters del Cluster 1 son más reflexivos, priorizando dispositivos tradicionales como PC y tabletas, y toman decisiones basadas en investigación previa a través

de búsquedas orgánicas y email marketing. Mientras el Cluster 0 enfatiza la conveniencia y productos accesibles, el Cluster 1 está orientado a compras más planificadas y de mayor valor. Cada grupo ofrece oportunidades únicas para personalizar campañas y mejorar la conexión con la marca.

# Arboles de decisión

## Árbol general

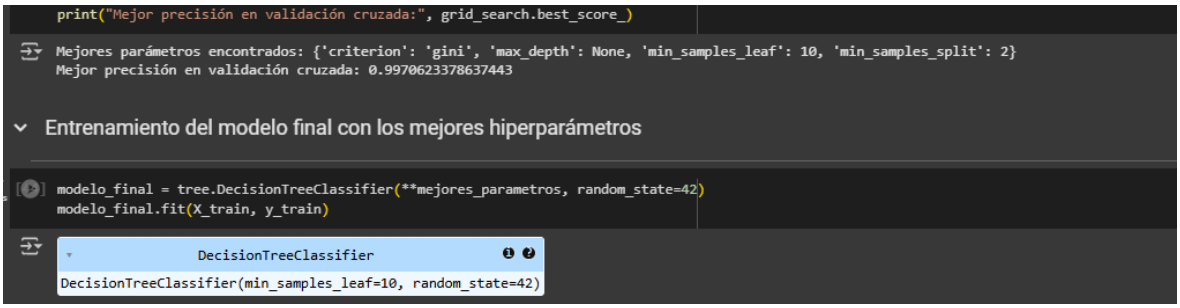
División de Datos (Entrenamiento y Prueba): Inicialmente, el conjunto de datos fue dividido en dos partes: un conjunto de entrenamiento y un conjunto de prueba. La división se realizó asignando el 70% de los datos al entrenamiento y el 30% restante a la prueba.



```
X = df_procesado.drop(columns=['cluster'])
y = df_procesado['cluster']
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)

print("Distribución en entrenamiento:")
print(pd.Series(y_train).value_counts(normalize=True))
print("\nDistribución en prueba:")
print(pd.Series(y_test).value_counts(normalize=True))
```

```
Distribución en entrenamiento:
cluster
1    0.542598
0    0.457402
Name: proportion, dtype: float64

Distribución en prueba:
cluster
1    0.544336
0    0.455664
Name: proportion, dtype: float64
```

Elección de Hiperparámetros con Grid Search: Para optimizar el rendimiento del árbol de decisión, se empleó la técnica de Grid Search.Nos permite realizar un  grid de hiperparámetros, evaluando el rendimiento del modelo para cada combinación de parámetros mediante validación cruzada.



```
print("Mejor precisión en validación cruzada:", grid_search.best_score_)
```

```
Mejores parámetros encontrados: {'criterion': 'gini', 'max_depth': None, 'min_samples_leaf': 10, 'min_samples_split': 2}
Mejor precisión en validación cruzada: 0.9970623378637443
```

Entrenamiento del modelo final con los mejores hiperparámetros

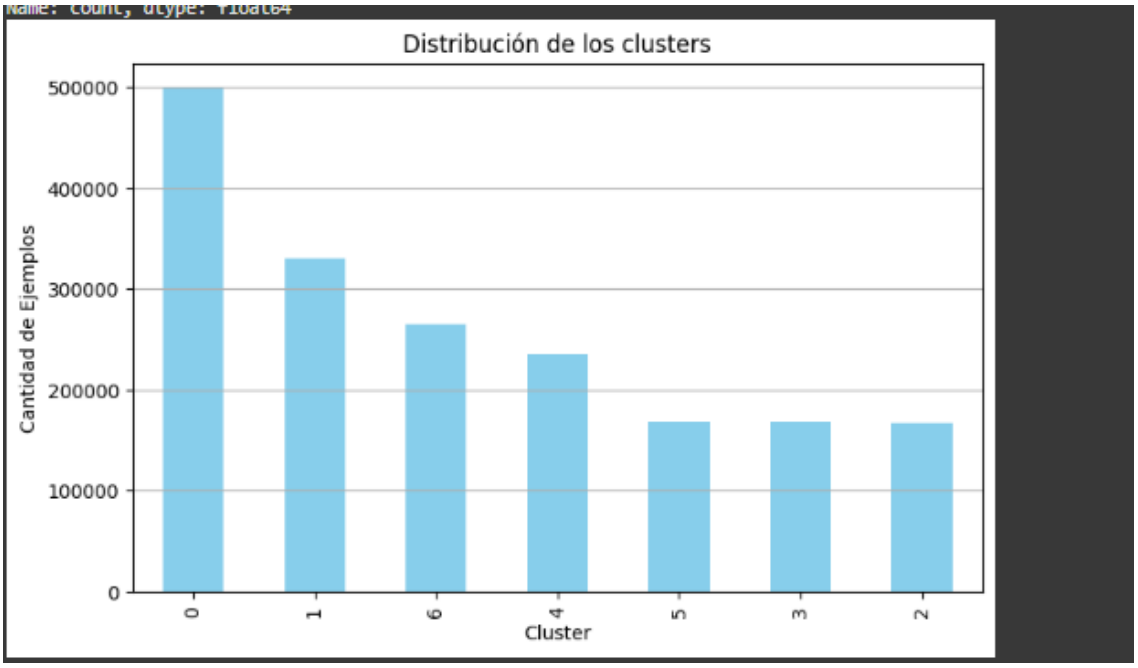
```
modelo_final = tree.DecisionTreeClassifier(**mejores_parametros, random_state=42)
modelo_final.fit(X_train, y_train)
```

```
DecisionTreeClassifier(min_samples_leaf=10, random_state=42)
```

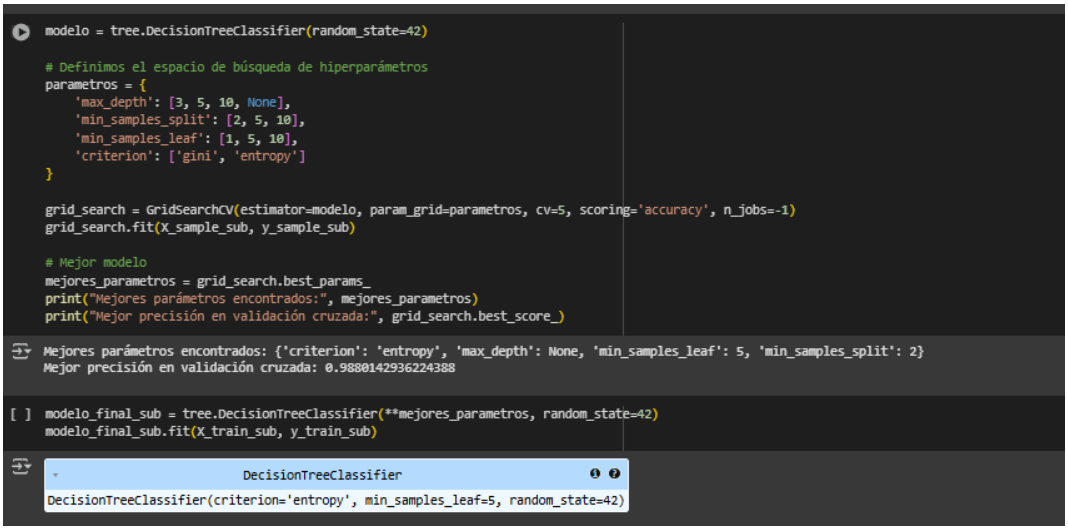
Entrenamiento del Modelo Final con los Mejores Hiperparámetros: Una vez identificada la mejor combinación de hiperparámetros a través de Grid Search, se procedió a entrenar el modelo final usando estos hiperparámetros óptimos. En donde vemos que El modelo de árbol de decisión desarrollado es bastante complejo, lo que podría indicar cierto grado de sobreajuste. Para comprobarlo sería útil probarlo con otros conjuntos de datos y ver cómo se desempeña en diferentes contextos, sin embargo en los datos de prueba el desempeño y las métricas son buenas .

## Árbol con subgrupos

- Balanceo de datos con SMOTE: La distribución inicial de los clusters estaba desequilibrada, por lo que se aplicó SMOTE para generar muestras sintéticas de las clases minoritarias. Esto evita que el modelo se sesgue hacia las clases mayoritarias y mejora su capacidad de generalización.



Ajustamos los hiperparámetros del árbol de decisión (profundidad máxima, muestras por hoja, etc.) mediante Grid Search.



```
modelo = tree.DecisionTreeClassifier(random_state=42)

# Definimos el espacio de búsqueda de hiperparámetros
parametros = {
    'max_depth': [3, 5, 10, None],
    'min_samples_split': [2, 5, 10],
    'min_samples_leaf': [1, 5, 10],
    'criterion': ['gini', 'entropy']
}

grid_search = GridSearchCV(estimator=modelo, param_grid=parametros, cv=5, scoring='accuracy', n_jobs=-1)
grid_search.fit(X_sample_sub, y_sample_sub)

# Mejor modelo
mejores_parametros = grid_search.best_params_
print("Mejores parámetros encontrados:", mejores_parametros)
print("Mejor precisión en validación cruzada:", grid_search.best_score_)
```

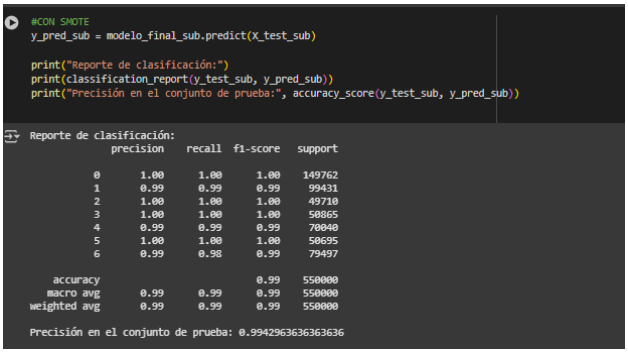
```
Mejores parámetros encontrados: {'criterion': 'entropy', 'max_depth': None, 'min_samples_leaf': 5, 'min_samples_split': 2}
Mejor precisión en validación cruzada: 0.9880142936224388
```

```
modelo_final_sub = tree.DecisionTreeClassifier(**mejores_parametros, random_state=42)
modelo_final_sub.fit(X_train_sub, y_train_sub)
```

```
DecisionTreeClassifier
DecisionTreeClassifier(criterion='entropy', min_samples_leaf=5, random_state=42)
```

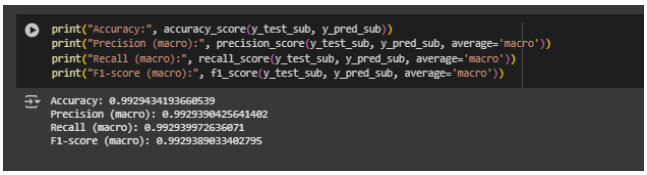
Para los resultados, se probaron dos versiones del modelo: una entrenada con datos balanceados mediante SMOTE y otra con datos originalmente desbalanceados. Este enfoque permitió comparar el impacto del balanceo de datos en el rendimiento del modelo. Los resultados obtenidos fueron los siguientes:

Con SMOTE: Precisión = 0.994



```
#CON SMOTE
y_pred_sub = modelo_final_sub.predict(X_test_sub)

print("Reporte de clasificación:")
print(classification_report(y_test_sub, y_pred_sub))
print("Precisión en el conjunto de prueba:", accuracy_score(y_test_sub, y_pred_sub))
```

```
Reporte de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    149762
           1       0.99      0.99      0.99     99431
           2       1.00      1.00      1.00     49710
           3       1.00      1.00      1.00     50865
           4       0.99      0.99      0.99     70040
           5       1.00      1.00      1.00     50695
           6       0.99      0.98      0.99     79497

    accuracy                           0.99    550000
   macro avg       0.99      0.99      0.99    550000
weighted avg       0.99      0.99      0.99    550000

Precisión en el conjunto de prueba: 0.9942963636363636
```



```
print("Accuracy:", accuracy_score(y_test_sub, y_pred_sub))
print("Precision (macro):", precision_score(y_test_sub, y_pred_sub, average='macro'))
print("Recall (macro):", recall_score(y_test_sub, y_pred_sub, average='macro'))
print("F1-score (macro):", f1_score(y_test_sub, y_pred_sub, average='macro'))
```

```
Accuracy: 0.9929434193660539
Precision (macro): 0.9929390425641402
Recall (macro): 0.992939972636071
F1-score (macro): 0.9929389033402795
```

Sin SMOTE: Precisión = 0.995

```
#SIN SMOTE
y_pred_sin_smote = modelo_final_sub.predict(X_test_sin_smote)

print("Reporte de clasificación:")
print(classification_report(y_test_sin_smote, y_pred_sin_smote))
print("Precisión en el conjunto de prueba:", accuracy_score(y_test_sin_smote, y_pred_sin_smote))
```

```
Reporte de clasificación:
              precision    recall  f1-score   support

           0       1.00      1.00      1.00    149493
           1       0.99      0.99      0.99     99333
           2       1.00      1.00      1.00     49860
           3       1.00      1.00      1.00     50589
           4       0.99      0.99      0.99     70488
           5       1.00      1.00      1.00     50779
           6       0.99      0.99      0.99     79458

    accuracy                           1.00    550000
   macro avg       1.00      1.00      1.00    550000
weighted avg       1.00      1.00      1.00    550000

Precisión en el conjunto de prueba: 0.9950327272727273
```

```
print("Accuracy:", accuracy_score(y_test_sin_smote, y_pred_sin_smote))
print("Precision (macro):", precision_score(y_test_sin_smote, y_pred_sin_smote, average='macro'))
print("Recall (macro):", recall_score(y_test_sin_smote, y_pred_sin_smote, average='macro'))
print("F1-score (macro):", f1_score(y_test_sin_smote, y_pred_sin_smote, average='macro'))
```

```
Accuracy: 0.9950327272727273
Precision (macro): 0.9950677010548326
Recall (macro): 0.9951310859025408
F1-score (macro): 0.995098686560233
```

Aunque la diferencia es mínima, SMOTE garantiza un tratamiento más equitativo de todas las clases, como confirman los informes de clasificación. Además, El modelo de árbol de decisión, debido a su alta complejidad, presenta dificultades en su interpretación y potencial riesgo de sobreajuste. Por tanto, es esencial probar este modelo en otros conjuntos de datos para verificar su capacidad de generalización y confirmar que su rendimiento no se limita únicamente al conjunto actual. Sin embargo, en los datos de prueba el desempeño y las métricas son buenas.