Anamaria Leguizamón Alarcón

Juan David Lopez Becerra

Daniel Andrés Becerra Sierra

**Introducción:**

La segmentación es una práctica esencial en el marketing moderno que busca personalizar la experiencia del cliente. Este proceso busca crear grupos de clientes homogéneos con características similares, lo que permite adaptar productos, servicios y comunicaciones a las necesidades específicas de cada grupo. Además, contribuye a la eficiencia en la asignación de recursos, centrándose en los segmentos más valiosos y mejora significativamente la experiencia del cliente al ofrecer una comprensión más profunda de cada grupo, lo que a su vez incrementa la satisfacción y fomenta la lealtad hacia la empresa. Por otro lado, la segmentación puede revelar nichos de mercado desatendidos y nuevas oportunidades de negocio, permitiendo el desarrollo de productos innovadores que respondan a las preferencias de cada segmento. Al identificar estas oportunidades, se pueden descubrir áreas de mejora en los productos existentes, lo que contribuye a una oferta más competitiva y alineada con las expectativas del mercado.

Con el fin de aplicar las técnicas de segmentación de manera efectiva, es esencial contar con un dataset que contenga información relevante sobre los clientes. En esta sección, se presentará el conjunto de datos que se utilizará para llevar a cabo el análisis. Este dataset incluye variables clave que permiten explorar las características de los clientes y su comportamiento, facilitando así la identificación de patrones y segmentos.

**Presentación del dataset:**

**Edad:** Representa la edad del cliente en años. De acuerdo a los diferentes grupos de edad podemos encontrar patrones en preferencias y comportamientos de compra.

**Ingresos Anuales:** Refleja los ingresos totales del cliente en un año. Es un indicador clave para segmentar a los clientes en función de su capacidad de gasto.

**Cantidad de Compras:** Indica el número total de compras realizadas por el cliente. Con ella podemos evaluar la frecuencia de compra y la lealtad del cliente.

**Valor Promedio de Compra:** Representa el gasto promedio de cada compra. Es útil para entender el comportamiento de compra de los clientes y su disposición a gastar.

**Frecuencia de Compras Mensual:** Mide la cantidad de veces que un cliente realiza compras en un mes. Con esta información podemos identificar clientes frecuentes.

**Dispositivo Utilizado:** Indica el tipo de dispositivo que el cliente utiliza para realizar compras (por ejemplo, móvil, ordenador).

**Fuente de Tráfico:** Refleja cómo el cliente llegó a la plataforma de compra (por ejemplo, redes sociales, búsqueda orgánica). Importante para evaluar la efectividad de canales de marketing.

**Días Desde la Última Compra:** Mide el tiempo transcurrido en días desde la última vez que el cliente realizó una compra.

**Valor Total Gastado:** Indica el total gastado por el cliente.

**Satisfacción del Cliente:** Mide el nivel de satisfacción del cliente de uno a cinco con los productos o servicios ofrecidos. Es crucial para evaluar la experiencia del cliente.

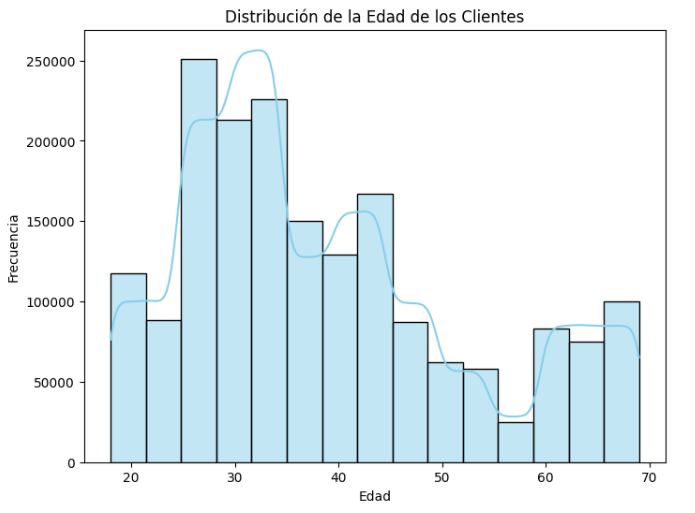
**Método de Pago:** Indica el método utilizado por el cliente para realizar sus compras (por ejemplo, tarjeta de crédito, PayPal o transferencia).

**Participación en Programa de Lealtad:** Refleja si el cliente forma parte de algún programa de lealtad.

**Productos Adquiridos:** Lista de productos que el cliente ha comprado. Permite analizar las preferencias de productos y realizar recomendaciones personalizadas.

**Visualización y análisis de las características del dataset:**

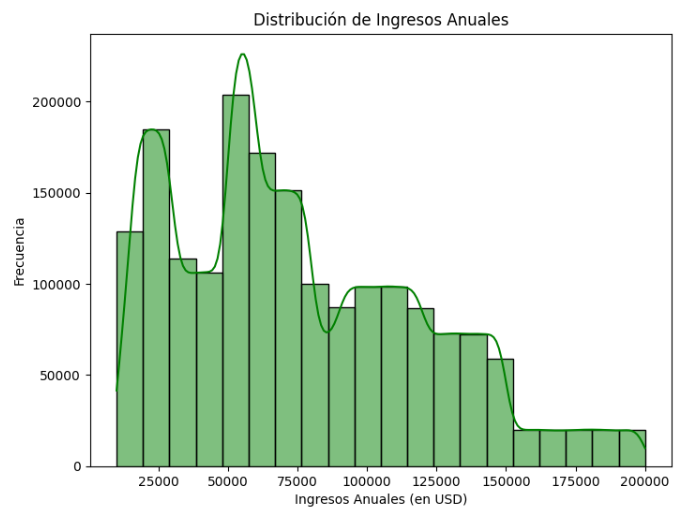
Para comprender de manera más profunda las características del dataset y facilitar la identificación de patrones relevantes, se procederá a realizar una serie de visualizaciones descriptivas. Estas representaciones gráficas nos permitirán examinar las distribuciones de las variables, las posibles correlaciones entre ellas, y la existencia de grupos que presenten características similares, lo cual es esencial para llevar a cabo un proceso de segmentación efectivo.

1. Gráfico: el siguiente histograma nos ayudar a identificar si existen grupos predominantes de edad, lo cual podría influir en la creación de segmentos con necesidades y preferencias diferenciadas.

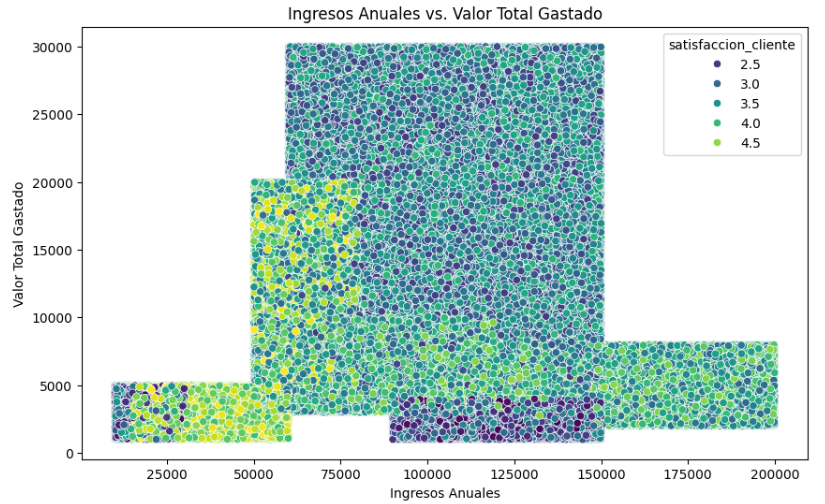
En el eje horizontal se representa la edad de los clientes, dividida en rangos de edad. En el eje y se muestra la cantidad de clientes que se encuentran en cada rango de edad.

Se observan dos picos en la gráfica que sugiere que existen dos grupos de edad predominante entre los clientes. El primer pico se encuentra alrededor de los 30 años y el segundo se ubica aproximadamente a los 40 años. La distribución tiene una cola que se extiende hacia las edades más avanzadas, lo que significa que hay una cantidad menor, pero no despreciable, de clientes mayores de 50 años.

1. Gráfico: al igual que en el gráfico anterior, esta representación visual nos permite entender cómo se distribuyen los ingresos anuales de un grupo de personas.

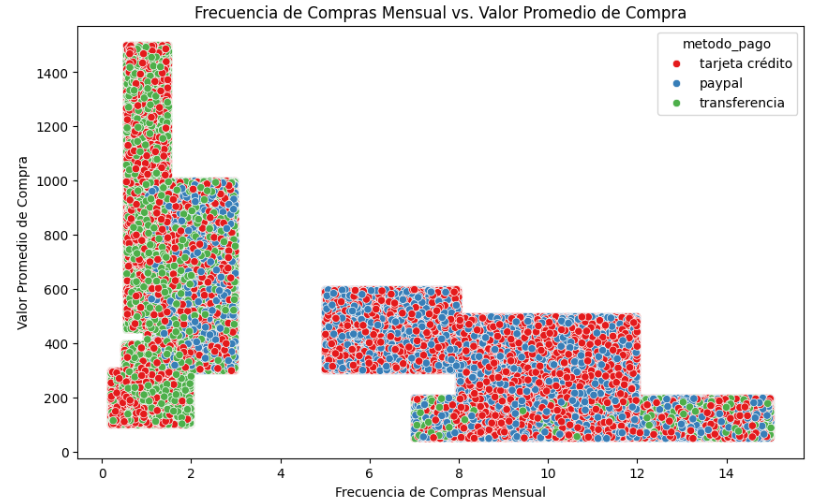


En el eje horizontal (x) se muestran los diferentes rangos de ingresos anuales, expresados en dólares estadounidenses. En el eje vertical (y) se indica la cantidad de personas que se encuentran dentro de cada rango de ingresos. La distribución de los ingresos anuales muestra una forma ligeramente asimétrica hacia la derecha. Esto significa que la mayoría de las personas tienen ingresos relativamente bajos o moderados, mientras que una minoría tiene ingresos muy altos. También podríamos decir que el nivel de ingresos más común en este grupo de personas es alrededor de 50000 dólares anuales.

1. Gráfica: este gráfico se utiliza para visualizar la relación entre ingresos anuales y valor total gastado, y además, se agrega una tercera variable categórica (satisfacción del cliente) representada por diferentes colores.

Los ingresos anuales se representan en el eje x y muestra los diferentes niveles de ingresos anuales de los individuos, el valor total gastado se representa en el eje y y muestra la cantidad total gastada por cada individuo. La satisfacción del cliente, se representa por el color de cada punto. Cada color corresponde a un nivel diferente de satisfacción del cliente, lo que permite visualizar cómo esta variable se relaciona con los ingresos y el gasto.

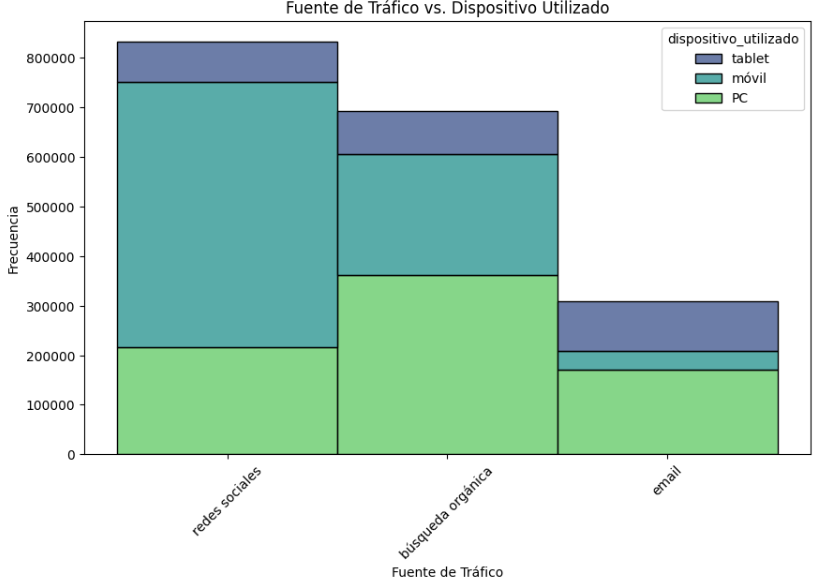
El análisis revela una correlación positiva entre los ingresos anuales y el valor total gastado, lo cual es lógico dado que un mayor poder adquisitivo suele traducirse en un mayor gasto. La satisfacción del cliente (indicada por el color) no parece seguir un patrón claro en función de los ingresos anuales o el valor gastado.

1. Gráfico: similar al gráfico anterior, este se utiliza para visualizar la relación entre dos frecuencia de compras mensuales y valor promedio de compra, en este caso también agregamos una variable categórica representada por diferentes colores la cual es método de pago.

En el eje x y y tenemos la cantidad de compras que realiza un cliente promedio en un mes y el promedio de dinero que un cliente gasta en cada compra respectivamente.

No se observa una relación lineal clara entre la frecuencia de compras y el valor promedio de compra. Los puntos se distribuyen de manera más bien aleatoria en todo el gráfico, lo que sugiere que la frecuencia con la que un cliente compra no necesariamente está relacionada con el monto que gasta por compra. Los clientes que utilizan tarjeta de crédito y PayPal presentan patrones de compra similares y los clientes que utilizan transferencias tienden a realizar compras menos frecuentes y de menor valor.

1. Gráfico: el siguiente gráfico es de barras apiladas y representa la relación entre las diferentes fuentes de tráfico (eje x) y los dispositivos utilizados por los clientes (colores dentro de las barras). Las barras están apiladas para mostrar la frecuencia con la que se utilizan tres tipos de dispositivos: tablet, móvil, y PC.



La mayoría del tráfico proveniente de redes sociales se realiza a través de dispositivos móviles, representando más de la mitad de la barra. Los PCs también son utilizados con frecuencia, pero en menor proporción, mientras que el uso de tablets es significativamente menor. El tráfico por búsqueda orgánica tiene una distribución más balanceada entre móviles y PCs, aunque los dispositivos móviles aún tienen una ligera mayoría. Este canal es mixto en cuanto a los dispositivos utilizados, lo que puede indicar que los usuarios acceden tanto desde dispositivos móviles como desde computadoras de escritorio. En el caso del tráfico por email, la mayoría de los usuarios acceden a través de PCs, lo que difiere de las otras dos fuentes de tráfico. Los móviles también tienen una representación significativa, pero menor que los PCs. Este patrón sugiere que las campañas de email marketing son más efectivas para los usuarios de PC, probablemente porque este tipo de contenido se consume más cómodamente desde una computadora de escritorio.

**Transformación de variables numéricas y codificación de variables categóricas:**

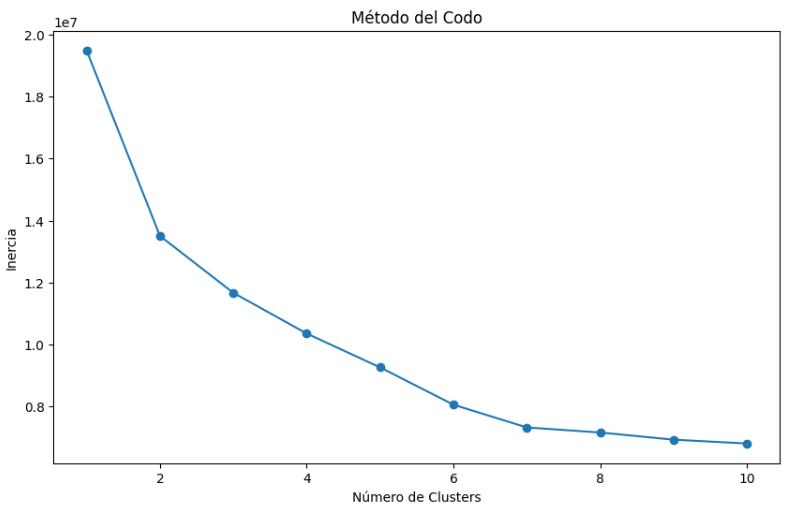
El algoritmo K-Means se basa en la minimización de la distancia euclidiana entre puntos, lo que implica que las variables con mayores escalas pueden tener un impacto desproporcionado en los resultados del modelo. Por ejemplo, en nuestro conjunto de datos, las variables como ingresos anuales y valor total gastado tienen valores significativamente más grandes que otras como frecuencia de compras mensual o días desde la última compra.

Para mitigar este riesgo, hemos decidido aplicar la estandarización de las variables numéricas utilizando la técnica de StandardScaler, que transforma cada variable para que tenga una media cero y una desviación estándar unitaria.

Si bien K-Means está diseñado originalmente para trabajar con variables numéricas, en este proyecto hemos decidido incluir variables categóricas como parte del proceso de segmentación por su importancia en la descripción de las características y comportamientos de los clientes. Variables como el dispositivo utilizado, la fuente de tráfico o el método de pago ofrecen información valiosa sobre las preferencias y hábitos de los clientes que no se reflejarán únicamente a través de las variables numéricas.

Para que el algoritmo K-Means pueda procesar variables categóricas, hemos aplicado la técnica de codificación One-Hot. Esta técnica convierte cada categoría en una variable binaria, lo que permite que el modelo considere la pertenencia o no de un cliente a cada categoría de forma independiente.

**Selección del número óptimo de clusters por método del codo:**

Para identificar el número óptimo de clusters en el algoritmo, hemos implementado el método del codo que evalúa el comportamiento de la inercia a medida que incrementa el número de clusters. 

Como se puede observar, la inercia disminuye a medida que aumenta el número de clústers, pero esta disminución se vuelve menos pronunciada después de cierto punto, que en este caso parece estar alrededor de 7. Este comportamiento nos indica que agregar más clusters no mejora significativamente la segmentación, por lo cual tomamos 7 clusters como la opción más eficiente para el modelo de segmentación.

**Aplicación de K-Means al dataset:**

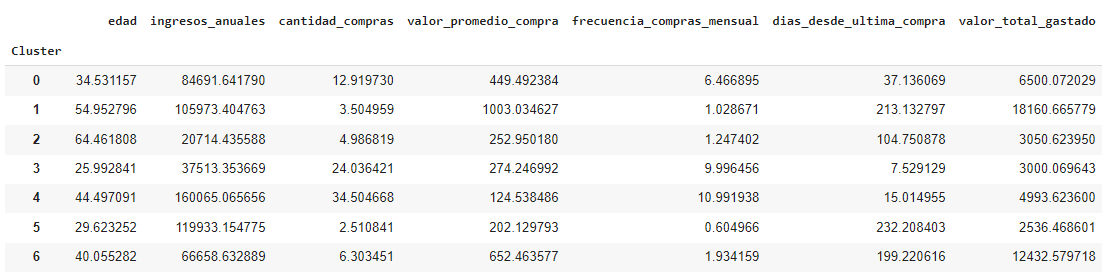
Con base en los resultados del método del codo, se procedió a aplicar el algoritmo K-Means con el número óptimo de clusters, fijando n clusters=7. Al elegir 7 clústeres, nos aseguramos de que el modelo captura la estructura subyacente de los datos sin complicar innecesariamente. Un número menor de clústeres podría simplificar demasiado la segmentación, agrupando de manera inadecuada a clientes que tienen comportamientos distintos. Por el contrario, un número excesivo de clústeres podría sobre ajustar el modelo, dividiendo a los clientes en demasiados grupos y dificultando la interpretación práctica de los resultados.

Además se procedió a integrar los resultados del modelo de K-Means al dataset original, agregando una columna denominada "Cluster" que contiene las etiquetas resultantes de la asignación a cada grupo. Este paso es fundamental, ya que permite identificar a qué clúster pertenece cada cliente dentro del conjunto de datos, facilitando el análisis posterior de las características específicas de cada segmento.

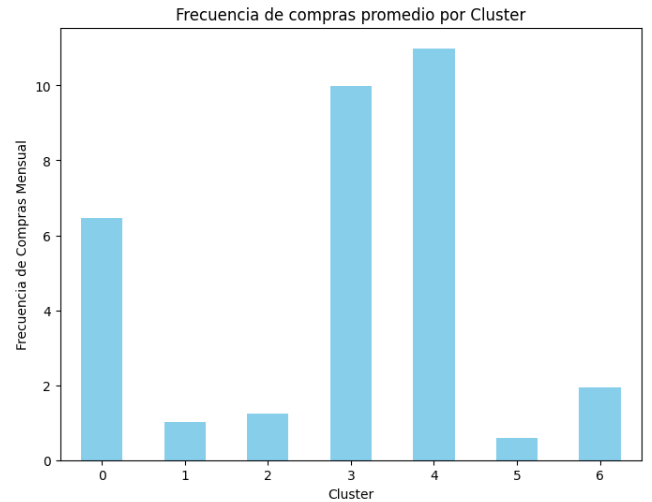
**Descripción detallada de cada uno de los grupos encontrados (segmentación de clientes):**

En esta sección, explicaremos en detalle los siete grupos de clientes que hemos identificado utilizando el algoritmo K-Means. Nos basaremos en los promedios de las principales variables numéricas y visualizaciones para comprender mejor las características y comportamientos que definen a cada clúster.

1. **Tabla:** la siguiente tabla muestra el promedio de cada variable para los 7 clusters.



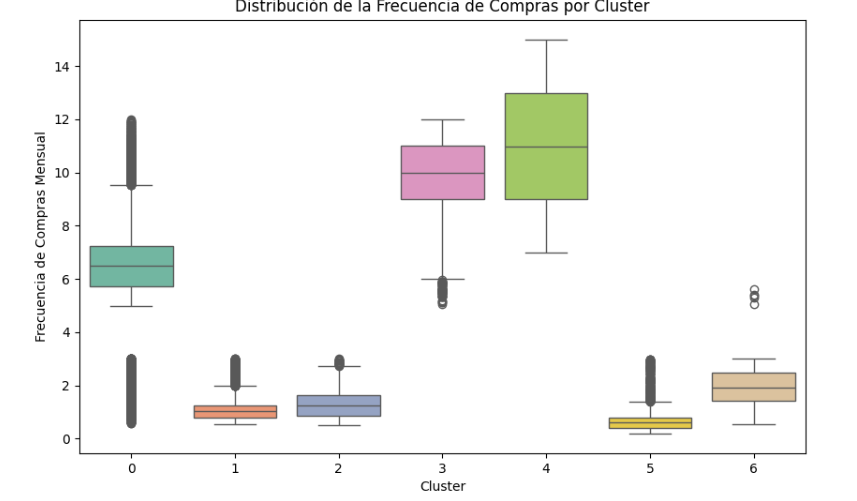
* Cluster 0: Este grupo está compuesto por clientes de mediana edad (promedio de 34 años), con ingresos anuales relativamente altos (84,691) y un comportamiento de compra moderado. Realizan en promedio alrededor de 13 compras, con un valor promedio por compra de 449.5, y una frecuencia mensual de compras de 6.5 veces. Estos clientes han gastado un total de 6,500 y tienden a regresar después de un período promedio de 37 días desde la última compra. Este perfil sugiere un grupo de consumidores activos y con una disposición a realizar compras frecuentes.
* Cluster 1: En contraste, este grupo está compuesto por clientes mayores (promedio de 54 años) con ingresos anuales considerablemente altos (105,973). Aunque realizan pocas compras al año (3.5 en promedio), el valor promedio de cada compra es elevado (1,003), lo que indica que prefieren realizar menos compras pero de alto valor. Su frecuencia mensual de compras es muy baja (1.03), y el tiempo desde la última compra es de 213 días, lo que sugiere un perfil de clientes de grandes transacciones pero poco frecuentes.
* Cluster 2: Los clientes de este grupo tienen una edad promedio mayor (64 años), con ingresos anuales mucho más bajos (20,714). Realizan menos compras al año (5 en promedio) con un valor promedio de compra de 252.95, y tienen una frecuencia de compra baja (1.25 veces al mes). Estos clientes también muestran una tendencia a espaciar más sus compras, con un tiempo promedio desde la última compra de 104 días. Este perfil puede indicar un grupo de clientes conservadores en sus hábitos de consumo, posiblemente con menos capacidad adquisitiva.
* Cluster 3: Este grupo está compuesto principalmente por clientes jóvenes (promedio de 25 años) con ingresos anuales moderados (37,513). Sin embargo, destacan por su alta cantidad de compras (24 en promedio) y una frecuencia mensual de casi 10 compras. El valor promedio por compra es relativamente bajo (274.2), pero el volumen alto de transacciones sugiere un perfil de compradores recurrentes y muy activos.
* Cluster 4: Este grupo es particularmente interesante por su alto volumen de compras (34.5 en promedio) a pesar de tener un valor promedio de compra bajo (124.5). Los clientes de este clúster tienen ingresos anuales altos (160,065) y una frecuencia de compra elevada (11 veces al mes). Con un tiempo promedio de 15 días desde la última compra, estos consumidores son los más activos y recurrentes en términos de frecuencia, pero optan por realizar compras de bajo valor.
* Cluster 5: Este clúster agrupa a clientes jóvenes (promedio de 29 años) con ingresos elevados (119,933), pero que realizan muy pocas compras al año (2.5 en promedio). El valor promedio de compra es relativamente bajo (202.1), y su frecuencia mensual de compras es extremadamente baja (0.6), lo que sugiere que, aunque tienen alta capacidad adquisitiva, realizan compras de manera esporádica, con un tiempo promedio de 232 días desde la última compra.
* Cluster 6: Este grupo está compuesto por clientes de mediana edad (40 años) con ingresos moderados (66,658), quienes realizan un promedio de 6 compras al año, pero con un valor elevado por compra (652.5). Su frecuencia mensual de compras es baja (1.9 veces), y el tiempo desde la última compra es de 199 días, lo que sugiere un perfil de consumidores que realizan compras ocasionales, pero de alto valor.

1. **Gráfico:** 

El gráfico presenta un diagrama de barras que compara la frecuencia promedio de compras realizadas por mes en los diferentes clusters de clientes. Cada barra representa un cluster específico, y la altura de la barra indica cuántas compras en promedio realiza ese grupo por mes.

Como vimos en la tabla anterior, es evidente que no todos los clusters se comportan de la misma manera en términos de frecuencia de compra. Hay grupos que realizan un número significativamente mayor de compras mensuales en comparación con otros. Por ejemplo, el cluster número 4, destaca como el grupo que realiza, en promedio, el mayor número de compras mensuales. Esto sugiere que este segmento podría estar compuesto por clientes altamente comprometidos o que tienen una necesidad más frecuente de adquirir productos o servicios.

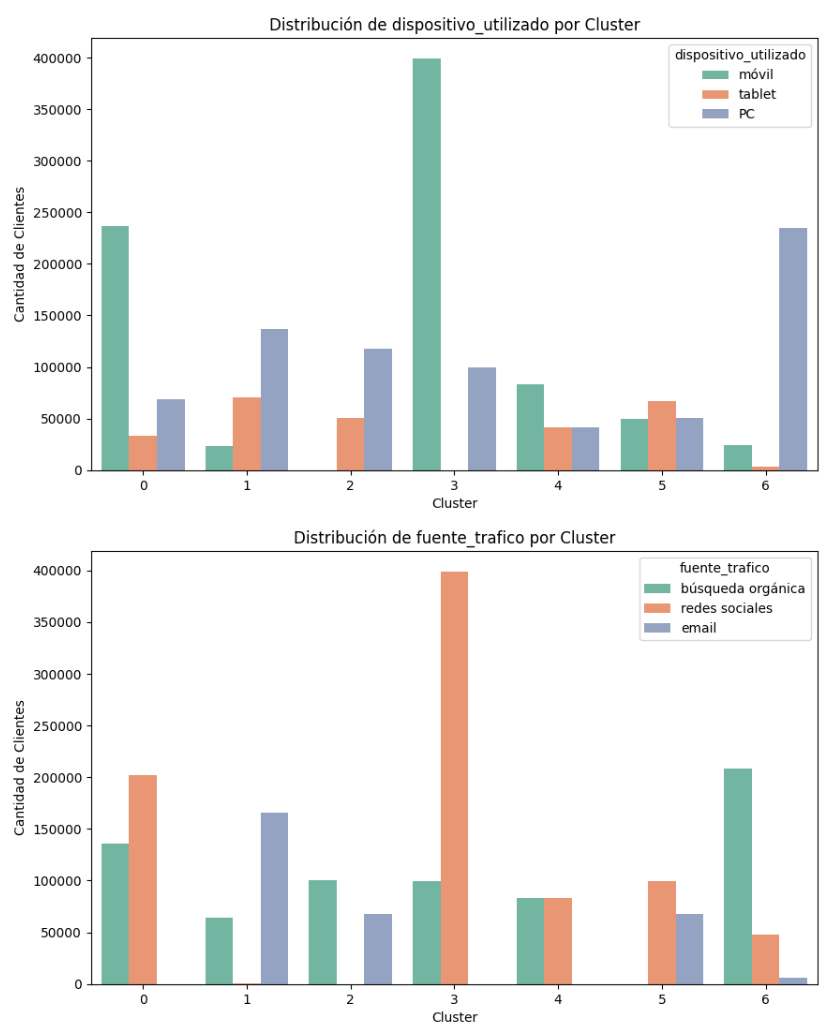
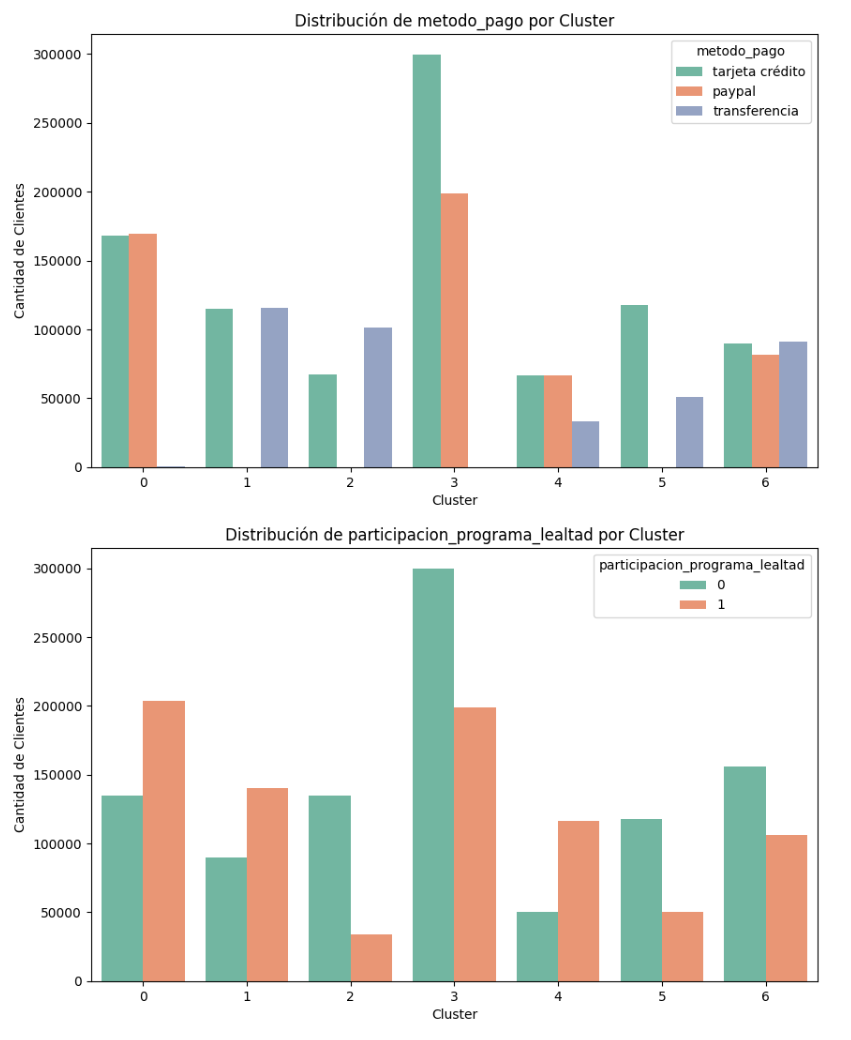
Los clusters 0 y 3 presentan una frecuencia de compra promedio que se encuentra por encima de la media general, pero por debajo del máximo establecido por el cluster 4. Estos grupos podrían representar un segmento de clientes con un nivel de compromiso moderado. El resto, muestran una frecuencia de compra promedio menor.

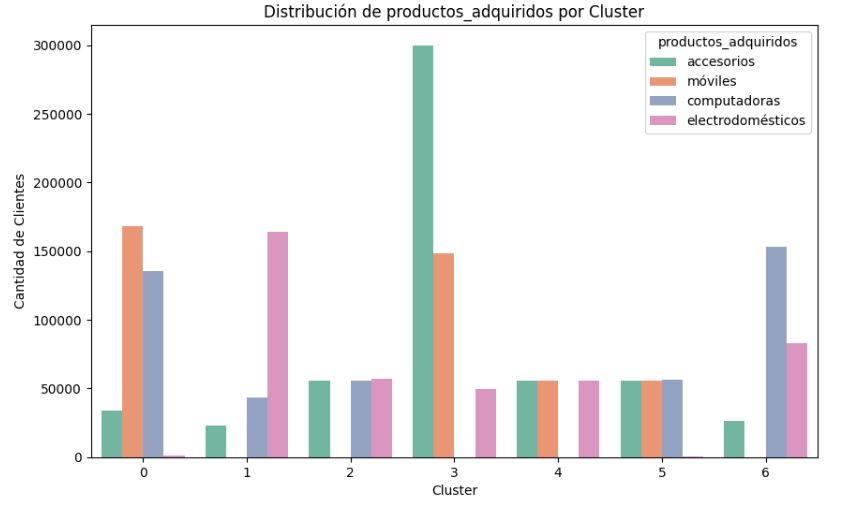
1. **Gráfico:** 

El gráfico de caja y bigotes muestra la distribución de la Frecuencia de Compras Mensual para cada uno de los clusters generados. Los clusters 3 y 4 destacan como los grupos con mayor actividad, presentando medianas elevadas (alrededor de 10 y 11 compras mensuales, respectivamente) y una dispersión amplia, lo que sugiere que estos clientes son altamente activos y pueden ser considerados como los más valiosos. En contraste, los clusters 1, 2 y 5 presentan frecuencias de compra muy bajas (medianas cercanas a 0.5 y 2 compras mensuales), indicando que estos clientes son de menor compromiso o tienen un patrón de compra esporádico. El Cluster 0 muestra una mediana intermedia, mientras que el Cluster 6 tiene una frecuencia moderada pero con baja dispersión.

Adicionalmente, se observan valores atípicos en la mayoría de los clusters, lo que sugiere la presencia de clientes con comportamientos excepcionales dentro de cada grupo. Estos hallazgos permiten identificar claramente qué grupos requieren estrategias de fidelización y cuáles podrían beneficiarse de estrategias de reactivación para maximizar el valor de cada segmento.

1. **Gráfico:** para complementar el análisis anterior, se han creado las siguientes visualizaciones que muestran una representación del comportamiento de compra de los siete clusters.

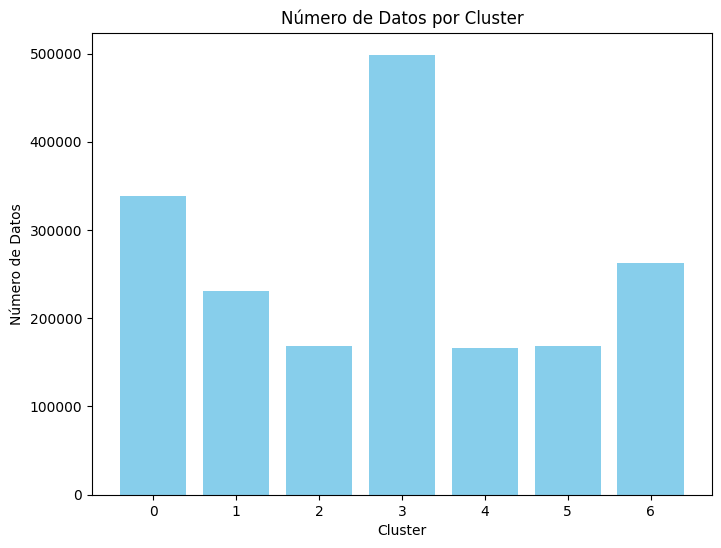


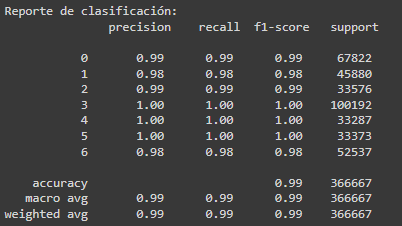


1. **Cluster 0:** Estos clientes utilizan principalmente móviles y acceden principalmente a través de redes sociales. Esto indica que buscan activamente productos, lo que es consistente con su patrón de compras frecuentes y valor medio de compra. Su método de pago preferido es PayPal y tarjeta de crédito, la mayoría de estas personas hacen parte del programa lealtad.
2. **Cluster 1:** Estos clientes utilizan más PC y acceden a la plataforma principalmente a través de email. Además, tienen una inclinación por pagar con tarjeta de crédito o transferencia igual, lo que es coherente con el valor elevado de sus compras.
3. **Cluster 2:** Este clúster utiliza principalmente PC y accede a la plataforma mediante búsqueda orgánica. Además, al igual que el cluster 1, son personas de edad avanzada y ninguna accede por medio de redes sociales. Prefieren transferencias como método de pago, lo que puede reflejar una preferencia por evitar el uso de crédito.
4. **Cluster 3:** Este grupo prefiere el móvil como dispositivo, ninguno por tablet y accede principalmente a través de redes sociales. El método de pago preferido es la tarjeta de crédito, lo que sugiere que están cómodos con el crédito para sus compras frecuentes.
5. **Cluster 4:** Al igual que el Cluster 3, prefieren móviles como dispositivo y acceden a través de redes sociales. Sin embargo, su método de pago preferido es PayPal. También su mayoría de personas hace parte del programa lealtad y tienen una preferencia igual por comprar accesorios, móviles y electrodomésticos, pero ninguno se interesa por computadoras.
6. **Cluster 5:** Este grupo accede principalmente desde PC y su principal fuente de tráfico son las redes sociales, ninguno accede por búsqueda orgánica. Prefieren tarjeta de crédito como método de pago.
7. **Cluster 6:** Estos clientes utilizan más PC y acceden principalmente a través de búsqueda orgánica. Prefieren pagar con tarjeta de crédito y transferencia.

**Entrenamiento y Evaluación del Árbol de Decisión**

Por más de que hay un desbalance inicial en las clases generadas por el k-means quisimos primero probar con las etiquetas tal como están por lo que en un primer momento tomando los datos ya etiquetados por los 7 clusters (segmentos de clientes) que definimos anteriormente, para este caso se tomó un split del 80% para datos de entrenamiento y un 20% para datos de testeo, tanto en las variables de clasificación como en la variable objetivo (cluster), tomando un random\_state=42 como semilla para siempre tener la misma división de los datos.



Una vez hecha la división llevamos a cabo el entrenamiento del árbol de decisión y empleamos para su respectiva evaluación las medidas precision, recall, f1-score, support y accuracy, dándonos el siguiente reporte:  


* **Precisión (Precision):** Es la proporción de verdaderos positivos entre todas las instancias clasificadas como positivas.

Las clases con mayor precisión son las clases 3, 4 y 5, con un 100%. Las demás clases tienen precisiones ligeramente menores pero siguen siendo bastante altas (98% a 99%).

* **Recall (Sensibilidad):** Es la proporción de verdaderos positivos entre todas las instancias que deberían haber sido clasificadas como positivas.

Todas las clases tienen un buen recall (98% a 100%). La clase 3 tiene un recall perfecto (1.00), lo que significa que todas las instancias de la clase 3 fueron correctamente identificadas.

* **F1-Score:** Es la media armónica entre la precisión y el recall, útil cuando hay un equilibrio entre estos dos.

El F1-score es alto para todas las clases (98% a 100%), lo que significa que el modelo tiene un buen equilibrio entre precisión y recall en todas las clases.

* **Support:** Número de instancias reales de cada clase.

La clase 3 tiene el mayor número de instancias (100,192), y la clase 1 tiene el menor número (45,880).

* **Exactitud global (accuracy):** 99%, lo que indica que el modelo clasifica correctamente el 99% de las instancias en general.
* **Weighted avg:** Promedio ponderado según el número de instancias en cada clase.

También es 99%, lo que significa que el rendimiento del modelo está muy bien distribuido en función del número de instancias de cada clase.

**Pasando a la matriz de confusión:**

1. **Clase 0 (≈ 350,000 instancias):**

**Predicciones correctas:** 67,265 (Correctamente clasificadas como clase 0).

**Errores de predicción:** 299 instancias de clase 0 fueron clasificadas como clase 3, y 234 como clase 6, lo que significa que la mayoría de los errores se distribuyen entre esas dos clases.

**Observación:** Aunque clase 0 tiene un buen desempeño (precisión de 99%), la clase 0 es la tercera clase más grande del conjunto de datos, por lo que incluso un pequeño porcentaje de errores implica un número considerable de instancias mal clasificadas.

**2. Clase 1 (≈ 230,000 instancias):**

**Predicciones correctas:** 44,867 (Correctamente clasificadas como clase 1).

**Errores de predicción:** La clase 1 sufre más confusión con la clase 6, con 876 instancias clasificadas incorrectamente. Esto podría deberse a que las características que diferencian estas dos clases no son suficientemente claras para el modelo.

**Observación:** Con una precisión del 98%, los errores en clase 1 son más notorios comparados con otras clases de tamaño similar. Además, dada la cantidad de datos, el impacto de esta confusión no es menor.

**3. Clase 2 (≈ 200,000 instancias):**

**Predicciones correctas:** 33,407.

**Errores de predicción:** Se confunde con clase 1 (68 casos) y clase 6 (90 casos).

**Observación:** Esta clase es más pequeña que las clases 0 y 3, y sus errores parecen estar relativamente bien distribuidos, lo que indica que el modelo distingue bastante bien esta clase.

**4. Clase 3 (≈ 500,000 instancias, la más grande):**

**Predicciones correctas:** 99,934 (prácticamente todas las instancias).

**Errores de predicción:** 258 casos fueron clasificados incorrectamente como clase 0.

**Observación:** Clase 3 tiene el mejor rendimiento con una precisión y recall perfectos (100%). Dado su tamaño, esto tiene un gran impacto positivo en las métricas globales del modelo.

**5. Clase 4 (≈ 200,000 instancias):**

**Predicciones correctas:** 33,277 (prácticamente sin errores).

**Observación**: Clase 4 tiene un desempeño perfecto (precisión de 100%), aunque su tamaño es relativamente pequeño comparado con la clase 3. Esto indica que el modelo generaliza bien en esta clase.

**6. Clase 5 (≈ 200,000 instancias):**

**Predicciones correctas:** 33,246.

**Errores de predicción:** Hay algunos errores pequeños hacia clase 1 (38 instancias) y clase 6 (86 instancias).

**Observación:** Aunque la clase 5 tiene algunos errores, la mayoría de las instancias se clasifican correctamente. Con una precisión de 100%, es probable que el modelo esté bastante ajustado para esta clase.

**7. Clase 6 (≈ 300,000 instancias):**

**Predicciones correctas:** 51,250.

**Errores de predicción:** 863 instancias fueron clasificadas como clase 1, lo que representa la mayor fuente de error en la clasificación de la clase 6.

**Observación:** La clase 6 tiene más errores que otras clases, lo que sugiere que hay características similares con la clase 1, lo que lleva a errores considerables en la predicción.

**Impacto del Tamaño de las Clases en el Modelo:**

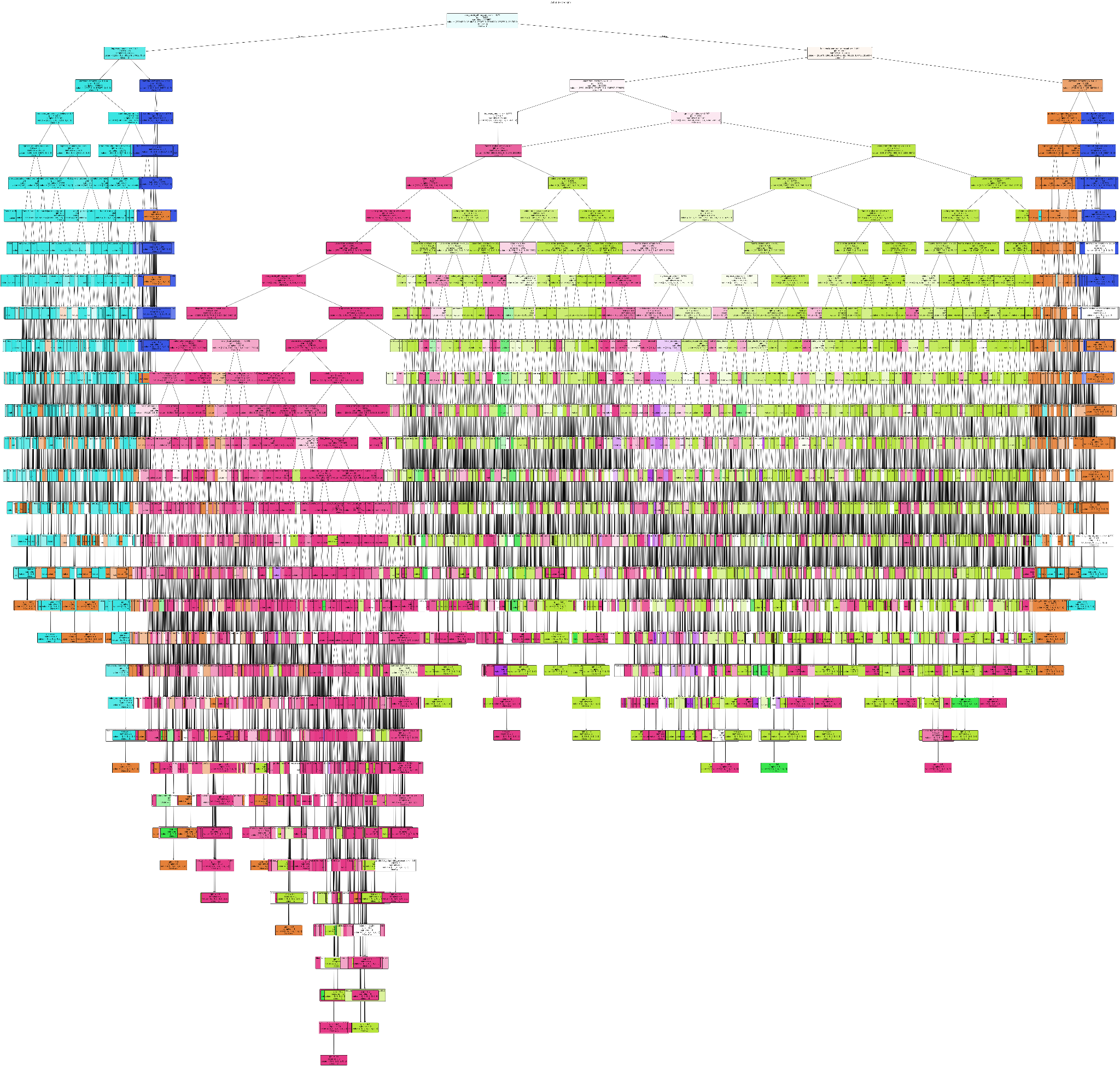
**Clases desbalanceadas:** La clase 3 es, con diferencia, la clase más grande (500,000 instancias), lo que implica que tiene un impacto considerable en las métricas globales. Cualquier error en esta clase sería más significativo en términos de precisión ponderada (weighted avg). La alta precisión en clase 3 beneficia enormemente al rendimiento general del modelo.

**Errores más comunes entre clases:**

La mayor confusión se observa entre la clase 1 y la clase 6, donde 876 instancias de clase 1 fueron clasificadas como clase 6, y 863 instancias de clase 6 fueron clasificadas como clase 1. Esto indica que el modelo tiene dificultades para separar estas dos clases, probablemente debido a características similares o insuficientes diferencias en los datos de estas clases.

**Desempeño en clases pequeñas:** Clases como 2, 4 y 5 tienen relativamente pocos datos comparados con las clases 0 y 3, pero el modelo parece manejarlas bien, con pocos errores y buenos resultados en precisión, recall y F1-score.

**Sobre Ajuste:** Además de las métricas que ya hemos evaluado otro indicativo que nos muestra un cierto sobreajuste en el modelo (por lo que no tiene una amplia capacidad de generalizar las clasificaciones) es la profundidad tan amplia que tiene el árbol, lo que además nos limita en la interpretabilidad del modelo mismo

****

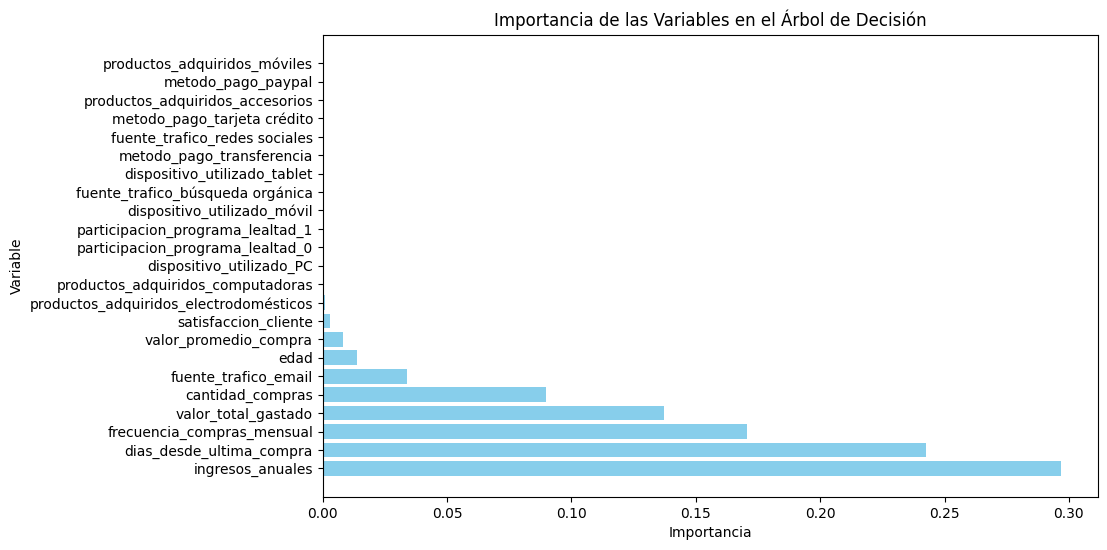
Ademas viendo los resultados del modelo tanto en el set de entrenamiento como en el de testeo:



**Exactitud de 1.0 en el entrenamiento:** podemos denotar que el modelo ha memorizado completamente el conjunto de entrenamiento, lo cual es una señal clara de sobreajuste (overfitting).

**Exactitud del 99.06% en el conjunto de prueba:** Aunque la exactitud en el conjunto de prueba es alta, el hecho de que el modelo tenga un 100% en entrenamiento puede significar que el modelo es demasiado específico para los datos de entrenamiento y podría no generalizar bien a nuevos datos en otros contextos.

**Variables importantes:**



#### **Variables más importantes (mayor peso en la predicción):**

1. **Ingresos Anuales**:
   * **Importancia**: Aproximadamente 0.30
   * **Análisis**: Es la variable más importante en el modelo, lo que sugiere que los ingresos anuales de los clientes son un factor clave para predecir el comportamiento (en este caso, los "clusters" o grupos a los que pertenecen). Es posible que los ingresos anuales influyan en el tipo de productos que adquieren, la frecuencia de compra, o el canal de pago que prefieren.
2. **Días desde la última compra**:
   * **Importancia**: Aproximadamente 0.20
   * **Análisis**: Esta variable indica cuántos días han pasado desde la última compra del cliente. Es lógico que sea una variable importante porque puede estar relacionada con la fidelidad o el compromiso del cliente con la empresa. Los clientes que han realizado compras recientes podrían pertenecer a un cluster más activo, mientras que aquellos que no han comprado en mucho tiempo podrían estar en un cluster menos activo.
3. **Frecuencia de compras mensual**:
   * **Importancia**: Aproximadamente 0.15
   * **Análisis**: La frecuencia con la que un cliente realiza compras en un mes parece ser un factor determinante para el modelo. Clientes que compran más seguido podrían clasificarse en clusters de mayor valor o fidelidad, mientras que aquellos con una frecuencia baja podrían estar en un cluster diferente. Esta variable también está relacionada con el comportamiento de compra regular.
4. **Valor total gastado**:
   * **Importancia**: Aproximadamente 0.10
   * **Análisis**: El valor total que ha gastado el cliente parece ser otro factor clave. Los clientes que gastan más dinero tienden a clasificarse en clusters diferentes que aquellos con menor gasto. Esto podría estar vinculado a los productos que adquieren o a los métodos de pago utilizados.
5. **Cantidad de compras**:
   * **Importancia**: Aproximadamente 0.08
   * **Análisis**: La cantidad total de compras realizadas también es una variable importante. Similar a la frecuencia de compras, puede ayudar a identificar a los clientes más activos y colocarlos en un cluster diferente de aquellos que compran menos.

#### **Variables de importancia media:**

1. **Fuente de tráfico (email)**:
   * **Importancia**: Aproximadamente 0.06
   * **Análisis**: El hecho de que un cliente llegue al sitio web a través de un correo electrónico (email) es un indicativo de su nivel de compromiso con la marca. Los clientes que responden bien a los correos electrónicos podrían estar en clusters más leales, mientras que aquellos que utilizan otras fuentes de tráfico pueden tener comportamientos diferentes.
2. **Edad**:
   * **Importancia**: Aproximadamente 0.05
   * **Análisis**: Aunque la edad no tiene un peso tan alto como otras variables, sigue siendo relevante. Diferentes grupos de edad podrían tener comportamientos de compra distintos, lo que podría influir en los productos que compran, su frecuencia de compra o el método de pago.

Ahora para poder ver una forma de tratar los problemas que hemos venido teniendo como lo es el sobreajuste, desbalanceo de clases y falta de interpretabilidad del arbol realizamos:

#### **1. Tratamiento del Desbalanceo (Undersampling)**

* El conjunto de datos presentaba clases desbalanceadas, donde algunas clases tenían muchas más instancias que otras. Este desbalance podía hacer que el modelo se sesgara hacia las clases más grandes, ignorando las clases menos representadas.
* Para mitigar este problema, realizamos **undersampling**, que reduce el número de instancias de las clases más grandes para igualarlas al número de instancias de la clase más pequeña. Esto permitió que todas las clases tuvieran el mismo peso en el entrenamiento, reduciendo el sesgo.
* Después del undersampling, todas las clases tenían exactamente **133,278 instancias**.

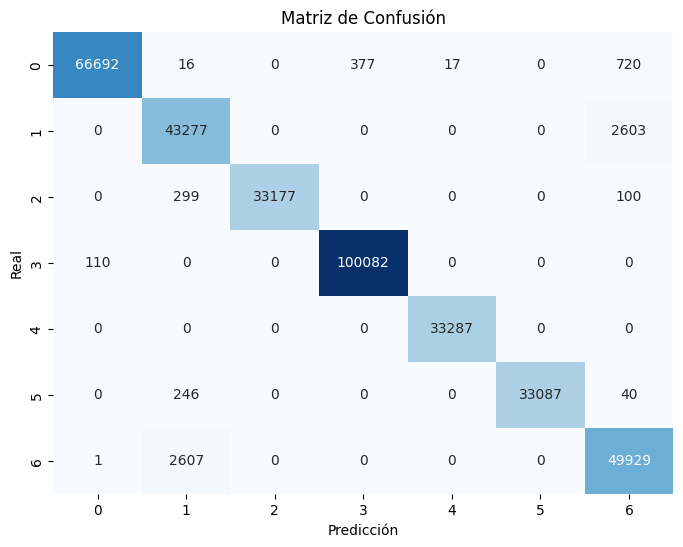
#### **2. Poda del Árbol de Decisión**

* Para tratar el sobreajuste (overfitting), limitamos la complejidad del árbol de decisión mediante la **poda**. Esto implica reducir la profundidad máxima del árbol max\_deph=5y ajustar el número mínimo de muestras requeridas para dividir un nodo min\_samples\_split=10
* La poda evita que el árbol se vuelva demasiado específico para los datos de entrenamiento, lo que mejora su capacidad de generalización a datos no vistos.

### **Análisis de las Métricas del Nuevo Árbol**

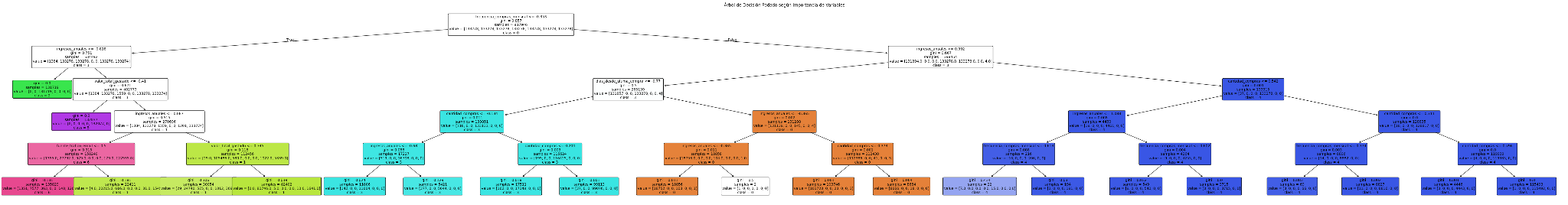
Las métricas presentadas en el reporte de clasificación y la matriz de confusión indican cómo se desempeña el modelo podado en el conjunto de prueba:

1. **Reporte de Clasificación**:
   * **Precisión**: Indica el porcentaje de predicciones correctas para cada clase.
     + El modelo tiene una precisión muy alta, variando entre 0.98 y 1.00 para todas las clases, lo que indica que las predicciones son muy precisas.
   * **Recall (sensibilidad)**: Mide la capacidad del modelo para detectar todas las instancias de una clase.
     + Todas las clases tienen un recall alto, aunque la clase 1 y la clase 6 presentan un recall ligeramente más bajo (0.94 y 0.95, respectivamente), lo que indica que el modelo no detecta todas las instancias de estas clases con la misma efectividad.
   * **F1-Score**: Es la media armónica entre la precisión y el recall.
     + Los F1-scores son consistentemente altos, con valores entre 0.98 y 1.00, lo que indica que el modelo tiene un buen balance entre precisión y recall.
   * **Exactitud global (accuracy)**: Es del 98%, lo que indica que el modelo clasifica correctamente el 98% de las instancias del conjunto de prueba.
2. **Matriz de Confusión**:



* + La matriz de confusión detalla cómo se clasificaron las instancias para cada clase.
  + La mayoría de las instancias se clasifican correctamente en su clase correspondiente, pero hay algunos errores notables:
    - **Clase 0**: 720 instancias se clasificaron incorrectamente como clase 6.
    - **Clase 1**: Hay confusión significativa con la clase 6, donde 2,603 instancias de la clase 1 se clasificaron como clase 6.
    - **Clase 6**: 2,607 instancias de la clase 6 se clasificaron como clase 1, lo que sugiere que el modelo tiene problemas para diferenciar entre estas dos clases.

### **Análisis Detallado del Árbol de Decisión Podado**



***Nota:*** *se puede ver mejor el gráfico en el notebook clickeando dos veces para hacer zoom*

En el árbol podado que has mostrado, se representan las decisiones que el modelo toma en cada nodo. Este árbol es mucho más sencillo debido a la **poda**, lo que permite analizar con mayor claridad las variables más importantes y cómo influyen en la clasificación.

1. **Estructura General del Árbol**:
   * La primera división del árbol (la raíz) se realiza en función de la variable **"ingresos\_anuales"**. Esto confirma que esta es la variable más importante para determinar en qué clase cae un cliente. Los ingresos anuales parecen ser un criterio clave para dividir a los clientes en diferentes clusters.
2. **Variables más importantes**:
   * **Ingresos Anuales**: Aparece varias veces en el árbol, lo que refuerza su importancia en la segmentación. Se utiliza tanto en la raíz como en los niveles inferiores para afinar las decisiones de clasificación.
   * **Frecuencia de Compras Mensual**: Esta variable aparece temprano en el árbol, justo después de los ingresos anuales, y parece ser un buen indicador del comportamiento de compra de los clientes. Es probable que los clientes que compran con más frecuencia pertenezcan a clusters específicos.
   * **Días desde la Última Compra**: Otra variable clave que aparece en varios niveles del árbol. Los clientes que han realizado una compra recientemente parecen clasificarse de manera diferente que aquellos que llevan más tiempo sin hacerlo.
3. **Decisiones del Modelo**:
   * En cada nodo del árbol, se hace una división basada en los valores de una variable. Por ejemplo, si los **ingresos anuales** son menores o iguales a 43.28, el cliente se clasifica en un grupo, y si son mayores, en otro. Este proceso continúa en cada nodo, refinando la clasificación a medida que se avanza por el árbol.
   * Cada hoja del árbol (los nodos terminales) representa una predicción de clase. Las hojas muestran la distribución de las clases en función de las decisiones anteriores, y la clase con más instancias en una hoja es la predicción final para ese grupo de clientes.
4. **Simplificación del Árbol**:
   * Dado que el árbol tiene una profundidad máxima de 5, no se crean nodos innecesarios o muy específicos, lo que evita el sobreajuste. Este es un beneficio directo de la poda, que mantiene el árbol interpretable y eficiente.

### El modelo tiene la capacidad para identificar patrones clave como los **ingresos anuales**, la **frecuencia de compras** y los **días desde la última compra**, el modelo permite a las empresas implementar estrategias de marketing personalizadas y tomar decisiones basadas en datos. Este tipo de segmentación ayuda a optimizar los inventarios, ofrecer promociones dirigidas, e identificar a los clientes con alto riesgo de churn, lo que permite una gestión más proactiva de la relación con el cliente. Además, al ser un modelo **interpretable**, proporciona claridad en las decisiones comerciales, permitiendo a los gerentes entender exactamente qué factores influyen en los comportamientos de los clientes.

### 

### **Interpretación y Conclusiones:**

* **Balanceo del Modelo**: El undersampling permitió que todas las clases tengan el mismo número de instancias en el conjunto de entrenamiento, lo que redujo el sesgo hacia las clases más grandes. Sin embargo, hay algunos problemas de confusión entre la clase 1 y la clase 6, lo que indica que estas clases pueden tener características similares o que el modelo necesita más refinamiento en esas áreas.
* **Poda del Árbol**: La poda del árbol de decisión redujo la complejidad del modelo y mejoró su capacidad de generalización. El árbol resultante es más interpretable y sigue teniendo un excelente rendimiento.
* **Métricas de Desempeño**: El modelo tiene una precisión y un recall muy altos, con una exactitud general del 98%. Aunque el modelo funciona bien en la mayoría de las clases, los problemas de confusión entre la clase 1 y la clase 6 sugieren que se podría realizar un análisis más profundo para mejorar la diferenciación entre estas clases.
* **Importancia de las Variables**: Las variables más importantes (ingresos anuales, frecuencia de compras mensual, días desde la última compra) juegan un papel clave en la segmentación de los clientes, lo que demuestra que el comportamiento financiero y de compra de los clientes es crucial para la clasificación.

El análisis de segmentación aplicado sobre el conjunto de datos de clientes ha permitido identificar siete grupos distintos (clusters), cada uno con características y comportamientos diferenciados. A través de la aplicación del algoritmo **K-Means**, se pudo segmentar a los clientes considerando variables clave como la **edad**, **ingresos anuales**, **frecuencia de compras**, y **método de pago**, entre otros. Cada cluster revela patrones únicos en el comportamiento de los clientes, desde aquellos que realizan compras frecuentes de bajo valor, hasta clientes que realizan pocas pero costosas transacciones.

Además, la visualización de datos proporcionó insights importantes, tales como:

* **Relación entre ingresos anuales y valor total gastado**, confirmando que los clientes con mayor poder adquisitivo tienden a gastar más.
* **Frecuencia de compras mensual y valor promedio de compra**, que mostró variaciones importantes entre los distintos segmentos.
* **Preferencias por dispositivos y fuentes de tráfico**, revelando la importancia de las redes sociales y dispositivos móviles para ciertos grupos de clientes.

El modelo de segmentación permitió además identificar clientes con potencial para estrategias de fidelización o de reactivación, dependiendo de su comportamiento de compra y nivel de compromiso con la marca.

Finalmente, el **árbol de decisión** entrenado con los clusters como clases objetivo tuvo un buen desempeño en la clasificación de clientes, con una **precisión promedio del 98%**. Sin embargo, se detectaron algunos problemas de sobreajuste en el modelo inicial, lo cual se abordó mediante técnicas como el **undersampling** y la **poda del árbol de decisión** que sin embargo deja la duda de que tan sobreajustado pueda llegar a estar.

**Reflexión sobre el proceso de segmentación y predicción:**

El proceso de segmentación fue crucial para identificar patrones ocultos en los datos y agrupar a los clientes de manera que permita un enfoque más personalizado en la toma de decisiones de marketing. La segmentación permite a las empresas enfocar sus recursos en los grupos más valiosos, mejorar la experiencia del cliente y detectar oportunidades de mercado.

Un aspecto interesante fue el uso de variables categóricas en conjunto con las numéricas, lo que permitió un análisis más completo del comportamiento del cliente. La inclusión de variables como el **dispositivo utilizado** o la **fuente de tráfico** aportó información valiosa sobre las preferencias tecnológicas y de consumo de los clientes, revelando la importancia de ajustar las estrategias de marketing a las plataformas más utilizadas por los diferentes segmentos.

El uso del **árbol de decisión** para predecir la pertenencia a un cluster específico también fue útil para entender las variables que influyen más en el comportamiento del cliente. Sin embargo, el proceso reveló algunas dificultades para diferenciar entre ciertos clusters (particularmente entre los clusters 1 y 6), lo que podría mejorarse con un ajuste más fino o utilizando modelos adicionales o distintos que complementen la segmentación.