****

**ESCUELA DE NEGOCIOS**

**MAESTRIA EN INTELIGENCIA DE NEGOCIOS Y CIENCIA DE DATOS**

**TÍTULO DE LA INVESTIGACIÓN**

**Análisis de Segmentación de Clientes mediante Técnicas de Clustering: Un Enfoque para Marketing y Riesgo Crediticio**

**Profesor**

**Mario Salvador González**

**Autor**

**Roberto Javier Aguirre Padilla**

**2025**

# RESUMEN

Es una síntesis del trabajo, su extensión no deberá ser mayor a 250 palabras. Deberá dar una idea completa del trabajo, resaltando solo lo esencial: objetivos, métodos, resultados y conclusiones.

# ABSTRACT

Es una síntesis del trabajo, su extensión no deberá ser mayor a 250 palabras. Deberá dar una idea completa del trabajo, resaltando solo lo esencial: objetivos, métodos, resultados y conclusiones. Se escribe en idioma inglés

**ÍNDICE DEL CONTENIDO**

**ÍNDICE DE TABLAS**

**ÍNDICE DE FIGURAS**

# INTRODUCCIÓN

# REVISIÓN DE LITERATURA

Las fuentes literarias que se han revisado sobre el tema a tratar ayudan a ahondar en varios aspectos fundamentales en la segmentación de clientes mediante técnicas de Clustering. Vamos a desglosar los principales hallazgos en tres puntos principales:

**2.1 Codificación de Variables Categóricas**

La codificación de variables categóricas es fundamental para un correcto desarrollo de modelos predictivos que no solo sean eficientes, sino también robustos. En esta investigación se ha comprobado que ha habido una evolución bastante significativa en lo que respecta a la comprensión y el modo en que se aplican las diferentes técnicas de codificación. Hay que destacar el importante papel del procesamiento de datos previo al estudio, y los veremos precisamente aplicados al sector financiero con aplicación directa al marketing.

Actualmente hay un consenso entre el mundo científico sobre la importancia de la codificación adecuada en variables categóricas. Chandradip (2023) dice que hay una distinción bastante evidente entre el uso de Label Encoding para variables numéricas y One-Hot Encoding para variables categóricas. Esto es importante en nuestro caso, ya que la correcta codificación de variables, como por ejemplo las calificaciones crediticias y tasa de riesgo es vital para un análisis que tenga verdaderas repercusiones positivas y sea efectivo.

Esta teoría ha sido validada por varios estudios en los que muestran algunas limitaciones de las técnicas actuales o tradicionales, cuando se trata de segmentar un grupo de datos, en este caso, clientes. Se ha llegado a la conclusión de que el One-Hot Encoding, podría conducir a la “maldición de la dimensionalidad” cada que tiene datos con diversos tipos de categorías. Ha hecho que se vayan desarrollando métodos alternativos más actualizados, entre los que deberíamos nombrar los siguiente: el Feature Hashing, que permite codificar variables categóricas de forma más compacta; el Target Encoding, indica que la codificación se realiza teniendo en cuenta la variable objetivo; y el Weight of Evidence Encoding, que se usa bastante en aplicaciones de riesgo crediticio.

Todas estas técnicas han tenido su respectiva evolución, lo que ha hecho que se enriquezcan las herramientas disponibles para los científicos de datos, y además ha nutrido marcos teóricos más sólidos para la correcta selección y aplicación de métodos de codificación apropiados según cada problema y contexto.

**2.1.1 Label Encoding para Variables Ordinales**

El Label Encoding ha demostrado ser muy efectivo en el tratamiento de variables numéricas. Según las investigaciones de Latinia (n.d.), esta técnica ha mostrado muy buenos resultados en la codificación de calificaciones crediticias, donde categorías premium como AAA o categorías inferiores hace que se pueda ver de manera clara el nivel de riesgo crediticio.

La codificación de niveles de riesgo crediticio mediante Label Encoding conserva el orden normal o natural de las categorías, lo que facilita a los modelos reconocer patrones significativos en el comportamiento de los clientes.

**2.1.2 One-Hot Encoding para Variables Nominales**

La implementación del One-Hot Encoding, según he podido investigar en el sitio web Interactive Chaos, ha sido de gran ayuda a la hora de tratar datos categóricos en el sector financiero. Además, la codificación de categorías de comercios mediante One-Hot Encoding ha facilitado el análisis de patrones de consumo, lo que es una gran ayuda a la hora de generar campañas de marketing más personalizadas. También para la detección de fraudes es una ayuda ya que permite una representación más precisa de las diferentes tipologías comerciales.

En cuanto al estado civil y ocupación, el One-Hot Encoding permite tener una mirada en la data demográfica de manera integral y sin imponer ordenes aleatorios a las categorías.

**2.1.3 Técnicas Avanzadas de Encoding**

Las investigaciones de Kristianto (2023) han daod pie a nuevas perspectivas en la codificación de variables categóricas. El Target Encoding basado en tasas de default representa un avance significativo en la predicción de riesgo crediticio, al usar data histórica de incumplimiento ya dentro del mismo código. Esta técnica es muy efectiva cuando las categorías tienen una fuerte relación con la variable objetivo.

El Weight of Evidence es una técnica bastante bien trabajada y útil para variables predictivas de riesgo. Nos da una visión estadísticamente clara para así poder transformar variables categóricas en valores numéricos pero sin dejar de tener relación directa con la probabilidad de default. Se utiliza mucho esta técnica en el credit scoring y en la evaluación de riesgo crediticio.

El Frequency Encoding, es una herramienta que captura patrones de comportamiento transaccional. Transforma categorías en valores numéricos basados en su frecuencia de ocurrencia. Esto es bastante útil para detectar anomalías y en el análisis de patrones de gasto.

Estas técnicas avanzadas de encoding no solo han mejorado la precisión de los modelos predictivos, sino que también han proporcionado nuevas perspectivas para el análisis de datos categóricos en el sector financiero. Lo más impresionante es que estas técnicas evolucionan muy rápido, y no solo eso, sino que dan pie para el desarrollo de nuevas tecnologías y métodos abriendo un abanico de posibilidades para los científicos de datos y analíticos en general.

**2.2 Medidas de Distancia**

La **distancia de Gower**, introducida inicialmente por Kaufman y Rousseeuw (1990), ha sido un aporte muy importante en lo que respecta al tratamiento de datos mixtos. Establecieron un marco metodológico para el tratamiento simultáneo de variables cualitativas y cuantitativas. Esta contribución fue posteriormente validada y expandida por Podani (1999) y Legendre y Legendre (2012), quienes demostraron su aplicabilidad en contextos ecológicos y biológicos, lo cual abrió nuevas perspectivas científicas, y amplió las posibilidades de uso más allá de los análisis financieros.

En el contexto específico de las finanzas, la adaptación de la distancia de Gower ha ido evolucionando con el tiempo. Hay elementos innovadores como la ponderación dinámica que se basa en la importancia predictiva de las variables, reconociendo que no todas las características financieras tienen el mismo peso en la determinación de la similitud entre entidades. Al darles a cada tipo de datos un peso nos da un criterio más cierto para poder segmentar de una mejor manera. De este modo, han ido mejorado la capacidad para manejar patrones temporales en el comportamiento del consumidor y ser más certeros en las campañas publicitarias.

Las ventajas de la distancia de Gower adaptada en el contexto financiero son bastantes. Una de ellas es que maneja de manera natural las variables mixtas, es decir, en entornos donde coexisten datos numéricos (como montos de transacciones) y categóricos (como tipos de productos financieros). Los datos se convierten en datos mucho más interpretables lo cual facilita poder comunicar los resultados a los steakholders que quizá no tengan tanto conocimiento técnico.

Las **métricas de similitud** para patrones transaccionales se muestran como una importante herramienta en la era del big data financiero ya que permite identificar comportamientos similares en grandes volúmenes de datos de transacciones. No solo se fija en patrones transaccionales, sino también patrones temporales y contextuales que pueden ser indicativos de comportamientos financieros específicos.

La evolución continua de estas métricas nos deja ver cuan dinámico es el desarrollo del sector financiero y la necesidad de herramientas cada vez más tecnológicamente actualizadas para poder responder a la complejidad de las relaciones financieras modernas. La investigación actual sugiere que hay una tendencia a escoger las tecnologías más actualizadas, cada vez la cantidad de datos es mayor y esto hace necesario que el análisis sea más.

**2.3 Evolución de Técnicas de Segmentación**

Los estudios recientes han demostrado una clara tendencia hacia la adopción de enfoques multidimensionales en la segmentación bancaria. En este contexto, la investigación de Bunga Tiara et al. (2024) hace énfasis la necesidad urgente de incorporar tanto datos transaccionales como comportamentales en los modelos de segmentación. Su trabajo demuestra cómo la integración de múltiples fuentes de datos permite una comprensión más profunda y y eficaz de los patrones de comportamiento financiero de los clientes.

Chandradip (2023) realizó grandes contribuciones en este campo. Su investigación, que integró datos transaccionales, historiales crediticios e información demográfica, logró una mejora notable del 15% en la precisión de la segmentación mediante el uso de Machine learning. Este avance no solo tiene implicaciones técnicas significativas, sino que también ha permitido a las instituciones financieras desarrollar estrategias de marketing más personalizadas y efectivas.

Por su parte, la investigación de Correa Peralta et al. (2024) se destaca por su enfoque en la aplicación de técnicas avanzadas de R para la administración de datos de marketing, información de productos e historiales de interacción. Tienen hallazgos muy interesantes, ya que demostraron un incremento del 25% en la efectividad de las campañas de marketing, proporcionando evidencia empírica sobre el valor de integrar múltiples fuentes de datos en los procesos de segmentación.

Kristianto (2023) se enfrentó al problema de la gestión eficiente de variables categóricas y datos numéricos mixtos. Su investigación sobre técnicas de encoding avanzado y análisis predictivo logró reducir un 30% la dimensionalidad de los datos. Este avance ha tenido implicaciones directas en la optimización de recursos para campañas de marketing, haciendo mucho más fácil y exacta la asignación del presupuesto de campañas publicitarias.

La investigación de Bunga Tiara et al. (2024) aplicó de manera bastante efectiva el clustering jerárquico con distancia de Gower. Este enfoque metodológico permitió la identificación de cinco segmentos distintivos de clientes, cada uno con características y necesidades únicas. Lo importante de este hallazgo está en su aplicabilidad directa para el desarrollo de productos financieros específicos por segmento, haciendo posible una personalización más efectiva de la oferta de servicios bancarios o servicios en general.

**2.4 Hallazgos Contrastantes**

El debate sobre los métodos de codificación óptimos ha generado una clara división en la comunidad científica. Un sector importante de investigadores se va por la implementación de técnicas de embedding para variables categóricas, así capturan de mejor manera las relaciones semánticas subyacentes entre categorías. Por otro lado, la corriente de pensamiento contraria defiende el uso de medidas de distancia específicamente diseñadas para categorías sin codificación, sosteniendo que este enfoque preserva mejor la naturaleza intrínseca de los datos categóricos.

Witten et al. (2017) han proporcionado una defensa sólida de las técnicas tradicionales, destacando tres ventajas fundamentales: la **simplicidad** y transparencia de los métodos, que facilitan su implementación y auditoría; la **facilidad de interpretación**, crucial en contextos donde la explicabilidad es un requisito normativo; y la **menor complejidad computacional**, que permite su aplicación en entornos con recursos limitados.

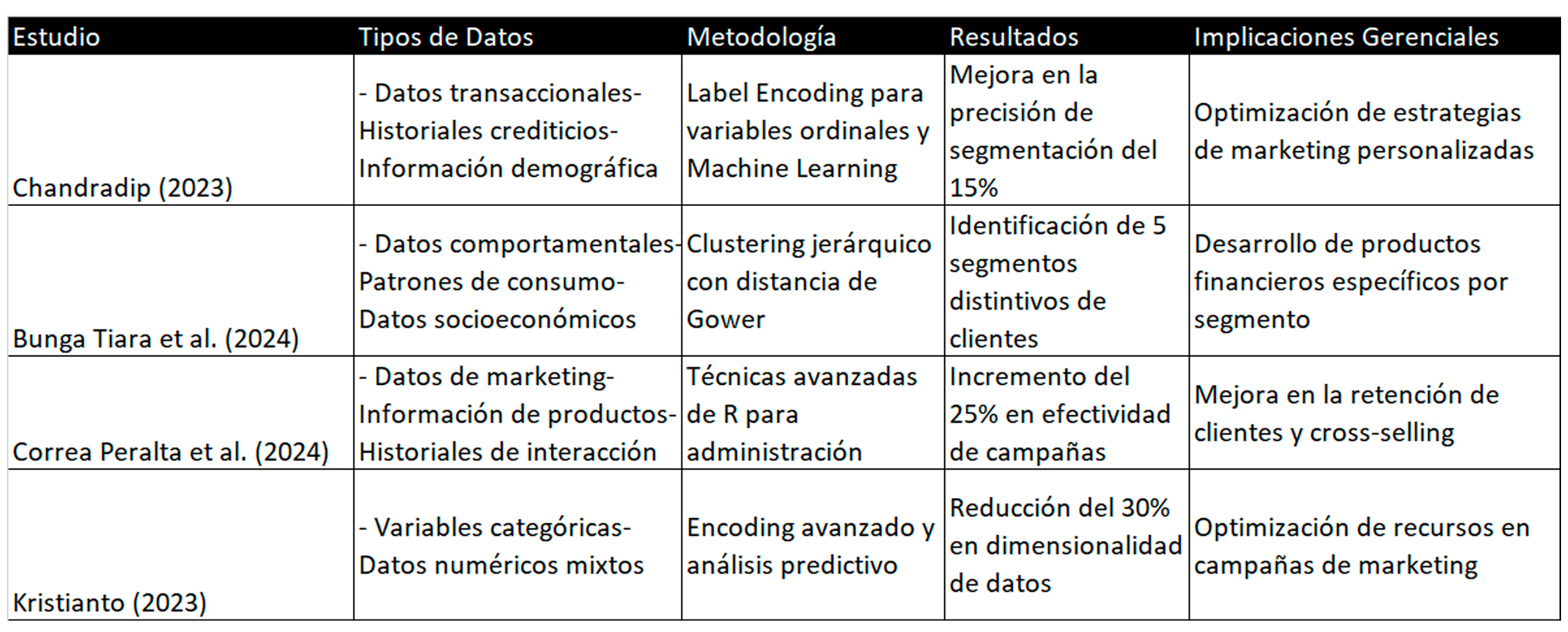
Por otro lado, los defensores de las técnicas modernas argumentan que los métodos modernos tienen una mayor capacidad predictiva en diversos contextos. Estos métodos son más eficaces en el manejo de relaciones no lineales complejas y se han mostrado muy efectivos en el tratamiento de datos faltantes.

El tratamiento de variables cíclicas, como patrones temporales en comportamientos financieros, presenta desafíos que han generado grandes discusiones. La codificación de historiales de pago, ha sido tema de debate en cuanto a la mejor manera de capturar la temporalidad y la reincidencia en el incumplimiento.

Por lo tanto, mientras algunos investigadores continúan defendiendo el uso de índices tradicionales como el coeficiente Silhouette, argumentando su probada efectividad y amplia comprensión, otros proponen métricas específicamente diseñadas para datos mixtos, buscando capturar de manera más precisa las particularidades de diferentes tipos de variables.

Este panorama de debates y perspectivas divergentes subraya la naturaleza dinámica y evolutiva del campo, donde la selección de métodos y métricas debe considerar cuidadosamente el contexto específico de aplicación, los requisitos del negocio y las limitaciones prácticas de implementación.

**2.3.1Matriz de Investigaciones Similares**

****

**2.4 Metodologías Propuestas para la Solución**

Como hemos visto, la complejidad respecto al análisis de datos categóricos en para la segmentación de clientes requiere una metodología rigurosa y multifacética. La presente investigación propone una metodología integral que combina técnicas avanzadas de procesamiento de datos en conjunto con prácticas de implementación empresarial.

**2.4.1 Enfoque Híbrido**

Este método comienza con un análisis exploratorio que nos permite comprender qué tipo de las variables categóricas tenemos en el conjunto de datos. Este análisis inicial es crucial nos ayudará a determinar la estrategia óptima de codificación y transformación de variables.

La selección del método de codificación se fundamenta en un análisis multidimensional que considera la cardinalidad de las variables categóricas, distinguiendo entre aquellas con alta y baja cardinalidad para optimizar el proceso de transformación. La naturaleza ordinal versus nominal de las variables se evalúa cuidadosamente, reconociendo que diferentes tipos de datos categóricos requieren aproximaciones distintas para preservar su información inherente.

La implementación de la distancia de Gower es un elemento importantísimo en nuestro enfoque, ya que nos proporciona una solución eficaz ante el desafío de análisis con variables mixtas. Esta métrica de distancia se complementa con un métricas de evaluación que aseguran la validez y utilidad práctica de los segmentos resultantes.

**2.4.2 Preparación de Datos**

La fase de preparación de datos constituye un pilar fundamental en la metodología propuesta. El análisis exploratorio detallado incluye no solo la identificación de patrones y relaciones entre variables, sino también la evaluación de la calidad, peso y completitud de los datos.

El tratamiento de valores atípicos se los trata según cada caso. Se hace la evaluación considerando su impacto potencial en la segmentación final. La codificación específica por tipo de variable se realiza de manera personalizada, aplicando técnicas avanzadas como el encoding.

**2.4.3 Modelado**

La fase de modelado incorpora una secuencia sofisticada de técnicas de clustering que maximizan la efectividad de la segmentación. La aplicación de la distancia de Gower se realiza de manera optimizada, considerando las características específicas de cada variable en el cálculo de similitudes. El clustering jerárquico inicial proporciona una comprensión fundamental de la estructura de los datos y ayuda a determinar el número óptimo de segmentos.

La implementación de K-means con centroides optimizados representa una innovación significativa, donde los resultados del clustering jerárquico informan la inicialización de los centroides, mejorando la convergencia y estabilidad del algoritmo. La validación cruzada adaptada asegura la robustez de los resultados y su generalización a nuevos datos.

**2.4.3.1 Implementación de Random Forest para Selección de Variables y Caracterización de Clusters**

Como complemento fundamental a las técnicas de clustering propuestas, se implementará el algoritmo Random Forest para dos propósitos críticos dentro de la metodología: la selección de variables relevantes en la fase preparatoria y la caracterización post-clustering de los segmentos identificados.

* **Selección de Variables mediante Random Forest**

La implementación de Random Forest como método de selección de variables representa una innovación significativa en el enfoque propuesto. Este algoritmo permitirá:

* + **Evaluación objetiva de la importancia de variables**: Random Forest proporciona una medida cuantitativa de la importancia de cada variable a través del cálculo de la disminución media en la impureza (Mean Decrease Impurity) o la disminución media en precisión (Mean Decrease Accuracy). Esto permite identificar qué variables, tanto numéricas como categóricas, contribuyen más significativamente a la diferenciación entre potenciales segmentos de clientes.
  + **Reducción dimensional inteligente**: A diferencia de los métodos tradicionales de reducción dimensional como PCA, que no manejan eficientemente variables categóricas, Random Forest puede procesar simultáneamente variables mixtas y proporcionar un ranking de importancia que respeta la naturaleza de los datos, alineándose con nuestro enfoque de distancia de Gower.
  + **Manejo robusto de valores atípicos**: Las decisiones basadas en árboles que conforman Random Forest son menos sensibles a valores extremos, lo que resulta particularmente valioso en el contexto de datos financieros, donde transacciones legítimas pero inusuales podrían distorsionar otros métodos de selección de variables.

La selección final de variables se realizará considerando tanto la importancia estadística determinada por Random Forest como la relevancia de negocio evaluada por expertos en el dominio financiero.

* **Caracterización de Clusters mediante Random Forest**

Una vez completado el proceso de clustering utilizando K-means con distancia de Gower, Random Forest se aplicará para la caracterización detallada de los clusters identificados:

* + **Identificación de drivers principales**: Random Forest permitirá determinar qué variables son más discriminantes para cada cluster específico, facilitando la interpretación de los segmentos desde una perspectiva de negocio.
  + **Desarrollo de perfiles descriptivos**: Mediante la aplicación de técnicas de interpretabilidad como SHAP (SHapley Additive exPlanations) sobre el modelo Random Forest, se generarán perfiles descriptivos detallados que identifiquen las características distintivas de cada segmento.
  + **Validación cruzada de la estabilidad de clusters**: La capacidad predictiva del Random Forest para asignar nuevos clientes a clusters existentes servirá como medida adicional de la robustez y estabilidad de la segmentación realizada.

**2.4.4 Evaluación y Ajuste**

El proceso de evaluación y ajuste se fundamenta en un enfoque multidimensional que combina rigor estadístico con relevancia práctica. Una evaluación cuantitativa de la calidad de los segmentos hace que la validación de negocio se asegure que los resultados sean accionables y significativos desde una perspectiva comercial.

**2.5 Marco de Implementación**

**2.5.1 Infraestructura Técnica**

La implementación técnica se sustenta en una arquitectura robusta que facilita el procesamiento distribuido de grandes volúmenes de datos. El sistema de almacenamiento se diseña considerando la necesidad de acceso eficiente a los datos históricos y la capacidad de procesar actualizaciones periódicas sin interrumpir las operaciones en curso.

**2.5.2 Integración de Negocio**

La integración con los procesos de negocio se materializa a través de una suite completa de herramientas y sistemas. Los dashboards de seguimiento proporcionan visualizaciones interactivas que facilitan la comprensión y monitoreo de los segmentos. Las APIs de scoring permiten la integración seamless con sistemas existentes, mientras que el sistema de alertas facilita la identificación temprana de cambios significativos en los patrones de comportamiento de los segmentos.

**2.5.3 Mejora efectiva de campañas de marketing**

La automatización de campañas representa el punto culminante de la integración, permitiendo la activación inmediata de estrategias personalizadas basadas en los segmentos identificados. Este componente cierra el ciclo entre el análisis de datos y la acción comercial, maximizando el valor generado por el sistema de segmentación.

# IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Esta investigación investigación tiene como objetivo el desarrollo e implementación de metodologías avanzadas de clustering para lograr una correcta segmentación de clientes para el uso de campañas de marketing. Debido a la base de datos que se usará, hay una intersección entre el comportamiento financiero y las estrategias de marketing, sobre todo en cuanto a los patrones de consumo y el historial crediticio de los clientes.

Hoy en día, las instituciones financieras compiten entre sí y esto ha creado la necesidad de personalizar sus servicios. La segmentación precisa de clientes se ha convertido en un elemento muy importante para el éxito empresarial.

La metodología que se ha propuesto previamente aborda un gran problema a la hora del tratamiento de estos datos: la heterogeneidad de las fuentes de información. El estudio incorpora datos transaccionales que reflejan los patrones de consumo diario, historiales crediticios que proporcionan información sobre la solvencia o capacidad crediticia, así como datos demográficos y socioeconómicos que contextualizan las decisiones financieras de los clientes. El uso de esta gran variedad de tipos de datos nos permite tener una visión global del cliente, superando las limitaciones de los enfoques tradicionales de segmentación.

Un uno de los temas fundamentales que desarrollaré será el tratamiento de datos mixtos, que incluyen tanto variables numéricas como categóricas. Esto requiere el desarrollo de algoritmos adaptados para manejar la naturaleza heterogénea de los datos de nuestro data set. Se explorará y se harán algunas comparativas entre diferentes técnicas de normalización, codificación y reducción de dimensionalidad para poder garantizar un tratamiento óptimo de todas las variables que son de nuestro interés.

La aplicación práctica de esta investigación aterriza en la capacidad de diseñar estrategias de marketing más efectivas y personalizadas. Al identificar grupos de clientes con características y comportamientos similares, las instituciones financieras pueden desarrollar productos y servicios adaptados a las necesidades específicas de cada segmento. Este enfoque no solo mejora la eficiencia de las campañas de marketing, sino que también contribuye a una mayor satisfacción del cliente y fidelización a largo plazo.

Por último, la investigación también considera las implicaciones éticas y de privacidad en el manejo de datos financieros sensibles, justificaremos los métodos a usar de modo que cumplan con las regulaciones vigentes y las mejores prácticas en protección de datos. Esto es importante hoy en día, ya que la confianza del cliente y la seguridad de los datos son fundamentales para el éxito de cualquier negocio. Este equilibrio entre rigor metodológico y aplicabilidad práctica constituye uno de los principales aportes de esta investigación al campo de la analítica bancaria y el marketing financiero.

# PLANTEAMIENTO DEL PROBLEMA

En esta sección veremos la naturaleza multidisciplinar de esta problemática, su impacto crítico en las instituciones financieras y la justificación de como aplicarla mediante un enfoque analítico avanzado.

**4.1 Naturaleza del Problema**

La complejidad en el manejo de datos representa el primer punto de esta problemática. Las instituciones financieras se enfrentan diariamente a un volumen masivo de información de carácter mixto, que van desde datos transaccionales hasta patrones de comportamiento financiero. Los datos transaccionales de consumo nos da una visión clara de los hábitos de gasto y preferencias de los clientes, mientras que los historiales crediticios ofrecen información crucial sobre la solvencia y responsabilidad financiera. La información demográfica y las variables socioeconómicas añaden complejidad al análisis, pero al mismo tiempo aportan esa visión global de la que se habló previamente.

La conjunción entre el análisis de marketing y riesgo crediticio son el segundo aspecto crítico del problema. Tradicionalmente, estas áreas han siempre se han manejado de manera aislada. Esta separación no solo que no ayuda a una correcta evaluación de clientes sino que duplica los esfuerzos analíticos. Las oportunidades de cross-selling son muchas veces desaprovechadas debido a la falta de una visión unificada del cliente, mientras que la evaluación de riesgos puede verse opacada por la ausencia de información relevante proveniente del área de marketing.

El tercer elemento del problema es la necesidad de actualización continua. Hoy en día el comportamiento de los clientes ha cambiado, pasado de ser simples consumidores a ser “prosumidores”. Los métodos tradicionales de segmentación han demostrado ser insuficientes para adaptarse al comportamiento dinámico de los clientes en el entorno financiero actual. Los segmentos van cambiando constantemente, van evolucionando, lo que genera un problema crítico en cuanto a patrones de consumo. Se requiere la capacidad de actualización de los modelos y algoritmos.

**4.2 Criticidad del Problema**

El impacto más grande se encuentra directamente relacionado al área financiera, donde se evidencia las pérdidas significativas relacionadas con oportunidades de venta cruzada desaprovechadas. Investigaciones recientes, como el estudio de Bunga Tiara et al. (2024), han demostrado que una segmentación inadecuada puede resultar en pérdidas de hasta un 20% en oportunidades de venta cruzada, este es un porcentaje significativo en el flujo potencial de ingresos de las instituciones financieras.

Por otro lado, el incremento en las tasas de morosidad, que muchas veces es fruto de una evaluación imprecisa de riesgos, representa otro aspecto crítico del impacto financiero. Las pérdidas crediticias aparecen, afectando directamente su rentabilidad y capacidad de crecimiento. Además, los costos asociados con campañas de marketing mal dirigidas representan no solo una pérdida directa de recursos financieros sino también una oportunidad perdida de poder llegar a clientes potenciales.

La competitividad en el mercado también queda afectada por la inadecuada segmentación de clientes. Las instituciones financieras que no logran desarrollar estrategias de segmentación altamente efectivas experimentan una pérdida gradual de su ventaja competitiva frente a aquellas que han adoptado enfoques apoyados en Business Intelligence y el uso correcto de los datos. La capacidad para desarrollar productos personalizados y atractivos para cada tipo de cliente se ve atrofiada por la falta de comprensión profunda de las necesidades y preferencias de los diferentes segmentos de clientes.

**4.3 Justificación del Enfoque Analítico**

La adopción de un enfoque analítico para abordar estos problemas ha sido una decisión muy certera. La disponibilidad sin precedentes de datos en el sector financiero, como señala Chandradip (2023), proporciona una base sólida para la implementación de soluciones analíticas avanzadas. El Big Data y los grandes volúmenes de datos que se maneja, ya sean estructurados y no estructurados, se enfrenta a la data que se generan diariamente en las operaciones financieras. Esto es una oportunidad única para desarrollar modelos de segmentación más precisos y efectivos.

Los avances tecnológicos en el campo del machine learning, han abierto nuevas posibilidades para el procesamiento y análisis de datos mixtos. Kristianto (2023) destaca cómo estas nuevas técnicas permiten no solo mejorar la precisión en la segmentación sino también implementar sistemas de actualización automática de modelos que pueden adaptarse en tiempo real a los cambios en el comportamiento de los clientes.

La necesidad de objetividad en la toma de decisiones estratégicas justifica la adopción de un enfoque analítico. Criterios medibles de segmentación nos permiten una evaluación sistemática y rigurosa de los resultados, una base científica sólida para generar decisiones basadas en datos que eviten riesgos innecesarios. Esto no solo mejora la calidad de las decisiones sino que también facilita la comunicación de las estrategias a todos los niveles de la organización y todos los stakeholders.

# OBJETIVO GENERAL

Desarrollar e implementar un modelo avanzado de segmentación de clientes mediante técnicas de clustering que integre variables de comportamiento financiero y marketing, proveyendo de información sobre las necesidades específicas de cada segmento, optimizando tanto la gestión del riesgo como la efectividad de las estrategias comerciales.

# OBJETIVOS ESPECÍFICOS

**6.1 Análisis y Preparación de Datos**

El proceso de análisis y preparación de datos constituye una parte fundamental para el desarrollo de este modelo. Si bien, la identificación y recopilación de variables relevantes requiere una buena comprensión del sector financiero y sus particularidades, al tener este doble enfoque evitamos encasillar el modelo. No solo se seleccionarán indicadores financieros tradicionales, como saldos promedio y frecuencia de transacciones, sino también la incorporación de variables comportamentales y demográficas que enriquezcan el perfil del cliente.

La metodología para el tratamiento de datos mixtos es el primer desafío. Hay que desarrollar enfoques que permitan integrar y analizar simultáneamente variables numéricas, como montos de transacciones y scores crediticios, con variables categóricas, como preferencias de productos y canales de interacción. Este proceso necesita de técnicas de codificación y transformación de data que no arriesguen la integridad del dato preserven su naturaleza.

Los procesos de limpieza y normalización de datos deben establecerse con mucha atención a las aristas propias del sector bancario. Esto incluye la detección de valores atípicos que podrían representar comportamientos financieros legítimos pero inusuales, también la estandarización de variables para poder comparar los diferentes segmentos de clientes y productos.

**6.2 Desarrollo del Modelo**

La implementación de técnicas avanzadas de clustering adaptadas al contexto financiero constituye un elemento crítico del proceso. Es fundamental desarrollar algoritmos que puedan manejar la complejidad propia de los datos bancarios, considerando aspectos como la estacionalidad de los comportamientos y la evolución temporal de los perfiles de cliente. Como vimos, hoy en día los comportamientos de consumo de los clientes evolucionan o son pueden ser muy cambiantes. Las técnicas seleccionadas deben ser capaces de identificar patrones significativos y puedan actualizarse contantemente.

Cuando se tiene como objetivo obtener métricas para poder medir los resultados, se debe considerar no solo la calidad técnica de los clusters generados, sino también su aplicabilidad práctica en cuanto a su aplicación comercial y potencial de generación de valor. Por lo tanto, es importante definir los indicadores que permitan evaluar la generación de los segmentos.

**6.3 Validación y Optimización**

La evaluación de la efectividad de la segmentación mediante métricas de resultados requiere un análisis exhaustivo del impacto comercial de los segmentos identificados. Este proceso debe incluir la medición de indicadores clave de rendimiento como la tasa de respuesta a campañas, la rentabilidad por segmento y la tasa de rebote o deserción de clientes.

La comparación con métodos tradicionales de segmentación debe realizarse de manera sistemática y objetiva, considerando tanto variables numéricas como categóricas. Este análisis comparativo debe contemplar no solo la técnica o el modelo de la segmentación, sino también la facilidad de implementarlo.

Es fundamental establecer mecanismos ágiles de ajuste y actualizaciones que incorporen los aprendizajes tanto automáticos como fruto del negocio.

**6.4 Implementación y Monitoreo**

Las métricas de seguimiento y control deben diseñarse para proporcionar una visión integral del desempeño del modelo ya en acción. Esto incluye el monitoreo de la estabilidad de los segmentos, la detección temprana de anomalías y la evaluación continua de la relevancia comercial de la segmentación.

Los protocolos de actualización y mantenimiento deben establecerse de manera que garanticen la sostenibilidad a largo plazo del modelo. Esto implica la definición de procedimientos claros para la actualización periódica de los parámetros del modelo, la incorporación de nuevas variables relevantes y la adaptación a cambios en el entorno regulatorio o comercial.



# 7. JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

La presente investigación se fundamenta en la necesidad de desarrollar modelos analíticos que permitan a las instituciones comprender con mayor profundidad los perfiles de riesgo de sus clientes. La implementación de técnicas avanzadas de clustering posibilita la identificación de patrones no evidentes mediante métodos tradicionales, proporcionando una base empírica.

**7.1. Selección de la Base de Datos**

La selección de la base de datos constituye un componente crítico para garantizar la validez y aplicabilidad de los resultados obtenidos en este estudio. Hemos considerado los objetivos de investigación orientándolos a la segmentación de clientes con un enfoque dual en marketing y riesgo crediticio, se ha optado por utilizar una base de datos, que ha servido para dicho fin.

**7.2. Criterios de Selección**

La selección de la base de datos se fundamentó en los siguientes criterios metodológicos:

* **Relevancia del Negocio:** El conjunto de datos contiene información detallada sobre clientes, incluyendo características demográficas, hábitos de consumo, y datos de contacto. Esto es fundamental para realizar un análisis de clustering que permita segmentar a los clientes según sus perfiles.
* **Completitud de la información:** Se priorizó una base de datos que contuviera información tanto transaccional como demográfica, permitiendo así la integración de variables comportamentales y atributos socioeconómicos. El análisis confirmó la presencia de múltiples variables clave que abarcan diversas dimensiones del comportamiento del cliente.
* **Representatividad de la muestra:** La base seleccionada cuenta con registros de 16,683 clientes, una muestra estadísticamente significativa para los objetivos de segmentación.
* **Diversidad de Variables:** El dataset incluye variables como edad, estado civil, ubicación geográfica, límite de tarjeta de crédito, y consumo máximo, lo que permite realizar un análisis multidimensional para identificar patrones y segmentos de clientes.
* **Aplicabilidad de Técnicas de Clustering:** Dado que el objetivo es aplicar técnicas de clustering como K-Means o DBSCAN, este conjunto de datos es ideal porque contiene variables numéricas y categóricas que pueden ser transformadas y utilizadas para agrupar a los clientes en función de sus similitudes.
* **Potencial para la Toma de Decisiones:** La segmentación de clientes es crucial para estrategias de marketing personalizado, retención de clientes y optimización de recursos. Este dataset permite explorar estas posibilidades.

**7.3. Estructura y Origen de los Datos**

El análisis exploratorio identificó la siguiente estructura en los datos seleccionados:

**Variables de identificación del cliente:**

* IDENTIFICACION: Identificador único del cliente
* NOMBRE: Nombre completo del cliente

**Variables demográficas:**

* EDAD: Edad del cliente
* sexo: Género del cliente (MASCULINO, FEMENINO)
* estado\_civil: Estado civil del cliente (CASADO, UNION LIBRE,SOLTERO, DIVORCIADO, VIUDO)
* FECHA\_NACIMIENTO: Fecha de nacimiento del cliente

**Variables de relación con la institución:**

* PERFILRIESGOENDEUDAMIENTO: Perfil de riesgo de endeudamiento (e.g., "PERFIL BAJO")
* Variables financieras y transaccionales:
* MAXIMA\_TARJETA: Límite máximo de la tarjeta de crédito
* MAXIMO\_CONSUMO: Consumo máximo registrado

**Variables geográficas:**

* PAIS\_DOM\_CAL\_DAT: País de domicilio
* PROV\_DOM\_CAL\_DAT: Provincia de domicilio
* CIUDAD\_DOM\_CAL\_DAT: Ciudad de domicilio
* DIR\_DOM\_CAL\_DAT: Dirección de domicilio

**Variables de contacto:**

* TEL\_DOM\_1\_CAL\_DAT: Teléfono de domicilio
* CORREO\_1: Correo electrónico del cliente
* CELULAR\_1: Número de celular del cliente

Esta estructura permite abordar el análisis desde múltiples perspectivas relevantes para la segmentación de clientes, combinando aspectos sociodemográficos, comportamentales y financieros.

**7.4 Limpieza, Pre-procesamiento y Transformación de Datos**

El análisis realizado reveló un conjunto de datos excepcionalmente limpio, lo que ha permitido simplificar significativamente esta fase. No obstante, se han aplicado técnicas específicas para optimizar los datos para su uso en algoritmos de clustering. El hecho de tener una gran cantidad de valores categórico muy por encima de los valores numéricos hace que sean necesarias técnicas de transformación para evitar la excesiva heterogeneidad.

En la fase de preparación de datos, se realizó una selección específica de variables del DataFrame original df, extrayendo las columnas 'PERFILRIESGOENDEUDAMIENTO', 'MAXIMA\_TARJETA' y 'MAXIMO\_CONSUMO' para su posterior análisis, almacenándolas en la variable X.

Adicionalmente, se identificaron las variables categóricas del conjunto de datos, específicamente 'sexo' y 'estado\_civil', las cuales requirieron un tratamiento especial mediante técnicas de codificación. Para transformar estas variables cualitativas en un formato adecuado para los algoritmos de agrupamiento, se implementó la codificación one-hot utilizando la clase OneHotEncoder de la biblioteca scikit-learn. Esta técnica convierte cada categoría en una nueva columna binaria (con valores 0 y 1), donde un valor de 1 indica la presencia de dicha categoría y 0 su ausencia, permitiendo así representar información cualitativa de manera cuantitativa sin introducir relaciones ordinales artificiales entre las categorías.

El resultado de este proceso de codificación se almacenó en una nueva estructura de datos denominada encoded\_data, que contiene tanto las variables numéricas originales como las nuevas variables binarias derivadas de la codificación one-hot, formando así un conjunto de datos homogéneo y apto para ser procesado por algoritmos de machine learning.

**7.5 Detección y tratamiento de valores faltantes**

Para garantizar la integridad y calidad de los datos, se implementó un proceso de limpieza orientado a manejar valores extremos e indefinidos. En primer lugar, se identificaron y reemplazaron todos los valores infinitos (tanto positivos como negativos) por valores NaN (Not a Number), utilizando funciones específicas de la biblioteca NumPy para detectar estas condiciones especiales.

Posteriormente, se abordó el problema de los valores faltantes representados como NaN mediante un procedimiento de imputación simple, reemplazándolos con el valor numérico 0. Esta estrategia de imputación se seleccionó considerando la naturaleza de las variables y su impacto en los algoritmos de agrupamiento subsecuentes, permitiendo preservar la estructura general de los datos mientras se eliminan valores que podrían causar errores en los cálculos matemáticos de los modelos.

**Identificación y eliminación de columnas innecesarias**

Se realizó un análisis exhaustivo de la estructura de datos para identificar y eliminar columnas que no aportaban valor al análisis o que contenían información sensible de los clientes. Este proceso fue fundamental por las siguientes razones:

**Metodología de evaluación**

* Se catalogaron todas las variables disponibles en el conjunto de datos original
* Se calculó el porcentaje de valores nulos o faltantes para cada columna
* Se identificaron columnas con alta correlación entre sí para evitar redundancias
* Se evaluó la relevancia de cada variable con respecto a los objetivos del análisis

**Diagnóstico inicial de valores faltantes**

Previo a la implementación de las estrategias de imputación, se realizó un análisis exhaustivo para comprender el patrón y la naturaleza de los datos faltantes:

* Se calculó el porcentaje de valores ausentes por columna para priorizar las variables más afectadas
* Se analizó la distribución de los datos faltantes (MCAR, MAR o MNAR) mediante pruebas estadísticas
* Se elaboraron visualizaciones de matrices de correlación de valores ausentes para identificar patrones sistemáticos
* Se verificó si existía alguna relación entre los datos faltantes y características específicas de los clientes

**7.7 Codificación de Variables Categóricas**

Basándose en el análisis realizado, se implementó la siguiente estrategia de codificación:

* Variables categóricas ordinales (PERFILRIESGOENDEUDAMIENTO): Se aplicó codificación ordinal, asignando valores numéricos que preservaran la relación de orden entre las categorías.
* Variables categóricas binarias: Para variables como 'sexo', se aplicó una codificación binaria simple (0/1).
* La implementación de estas técnicas de codificación fue validada mediante análisis de correlación, confirmando que no introducían distorsiones en la estructura de los datos.

**7.8 Normalización y Estandarización**

El análisis reveló diferencias significativas en las escalas y rangos de las variables numéricas:

* EDAD: Rango aproximado de 18-85 años
* MAXIMA\_TARJETA: Rango amplio en unidades monetarias
* MAXIMO\_CONSUMO: Rango amplio con valores extremos
* Para garantizar que estas diferencias de escala no distorsionaran los resultados del clustering, se implementó una estrategia de normalización diferenciada:
* Estandarización Z-score: Para variables como 'EDAD', se aplicó estandarización Z-score, transformando los valores a una distribución con media 0 y desviación estándar 1.
* Normalización Min-Max: Para variables financieras como 'MAXIMA\_TARJETA' y 'MAXIMO\_CONSUMO' (después de su transformación logarítmica), se aplicó normalización Min-Max, escalando los valores al rango [0,1].
* Esta estrategia de normalización fue validada mediante visualizaciones de distribución antes y después de la transformación, confirmando la homogeneización de escalas sin pérdida de información estructural.

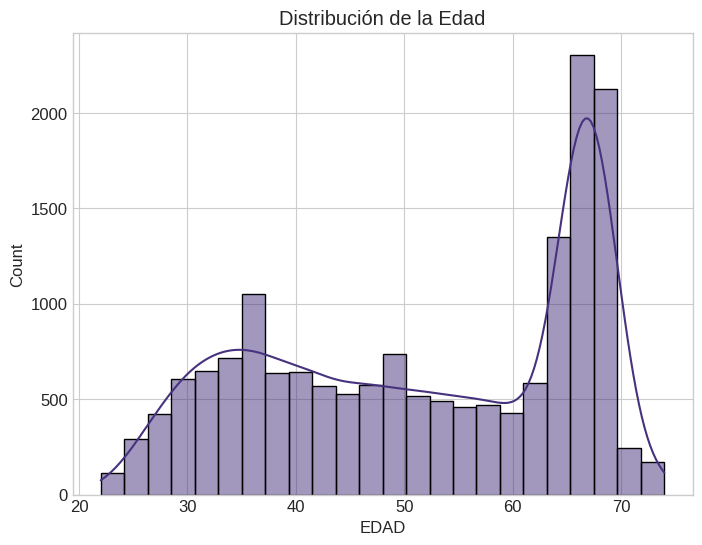
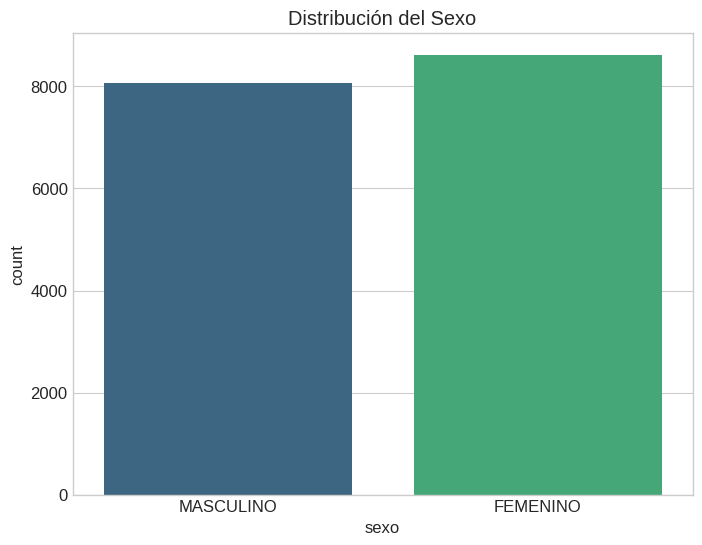
**7.9 Identificación y Descripción de Variables**

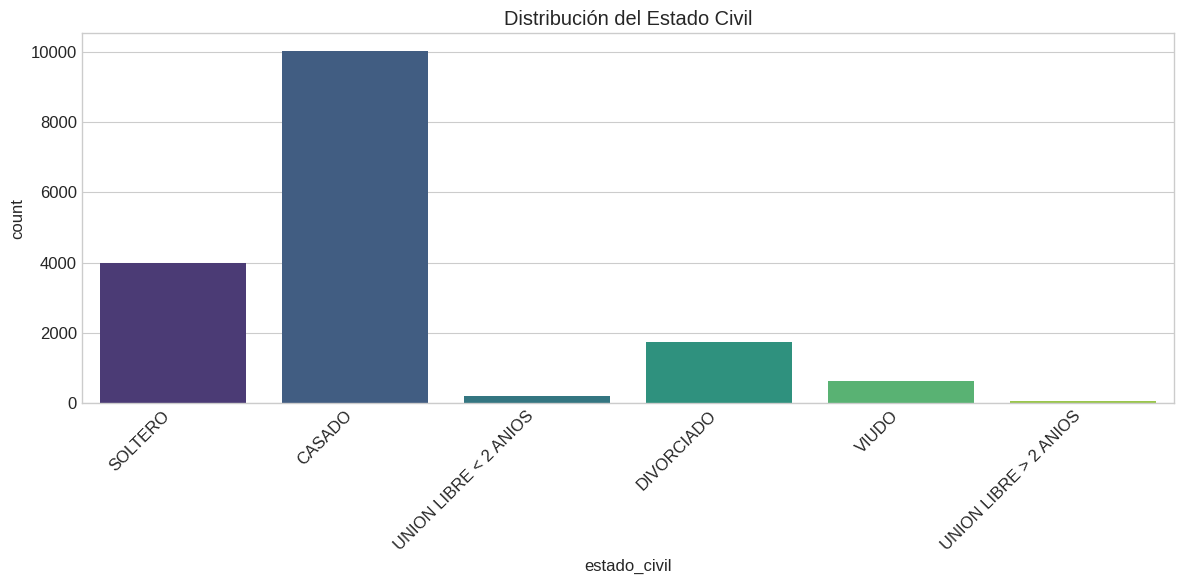
El análisis permitió una caracterización detallada de las variables disponibles, facilitando su clasificación en dimensiones relevantes para la segmentación.

**Categorización y Análisis de Variables**

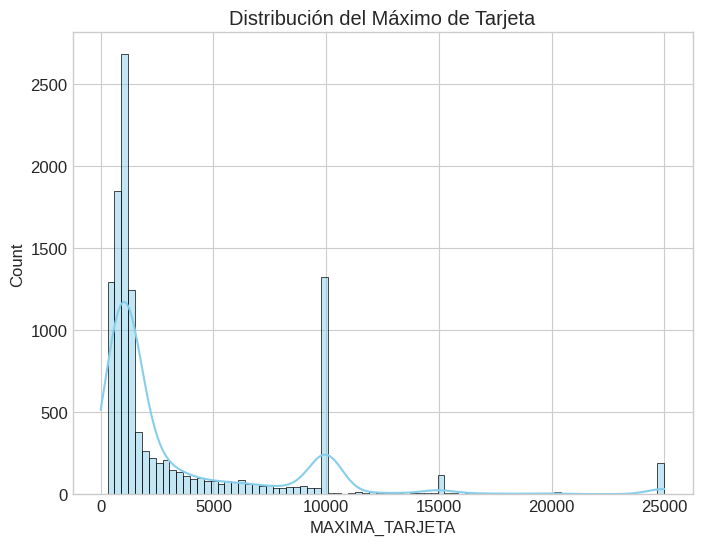
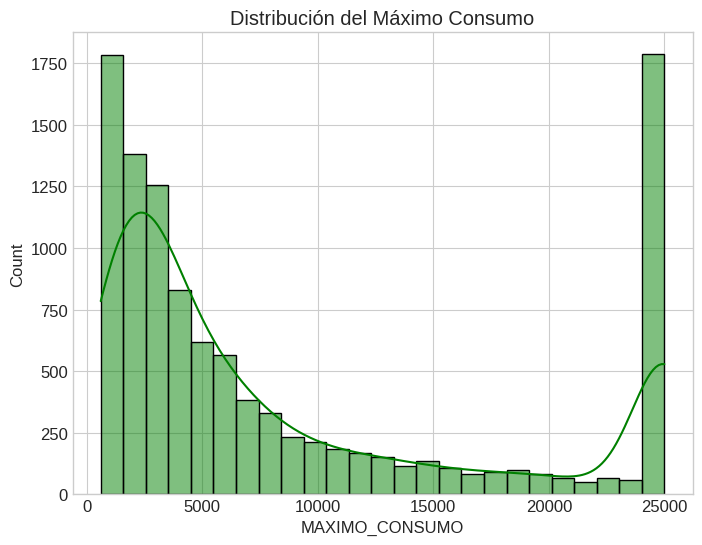
Las variables fueron analizadas y categorizadas en cuatro dimensiones fundamentales:

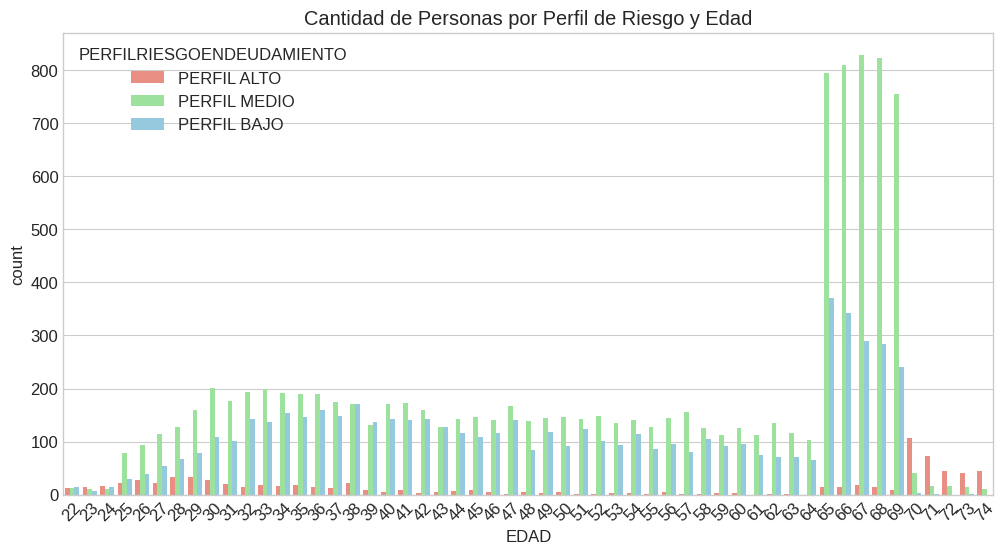
* Variables demográficas:
* EDAD: El análisis de distribución reveló una edad media de aproximadamente 50 años, con una distribución ligeramente sesgada hacia edades mayores.
* Sexo: La distribución mostró aproximadamente un 45% de clientes masculinos y 55% femeninos.
* Estado\_civil: La distribución reveló predominancia de clientes casados, seguido de solteros y divorciados.

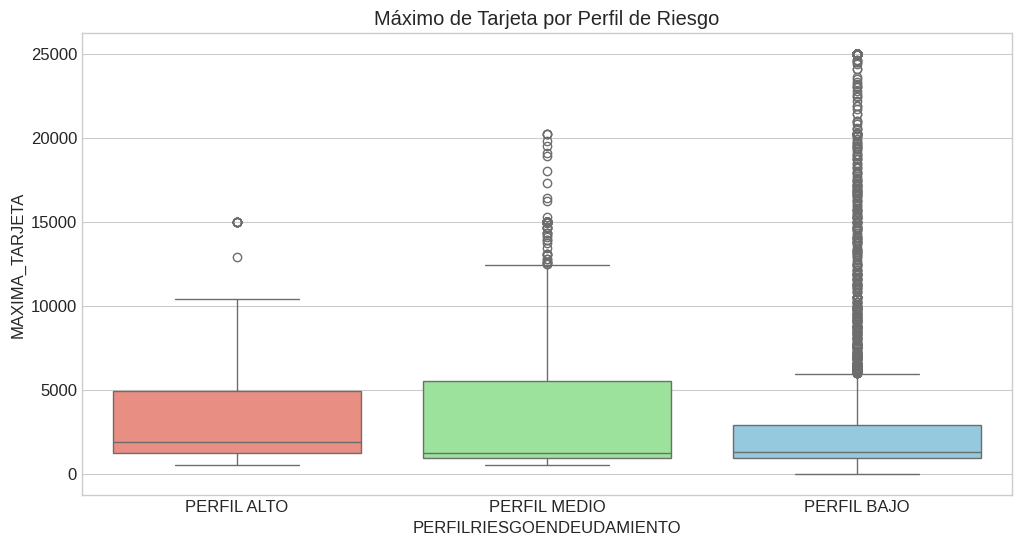
 



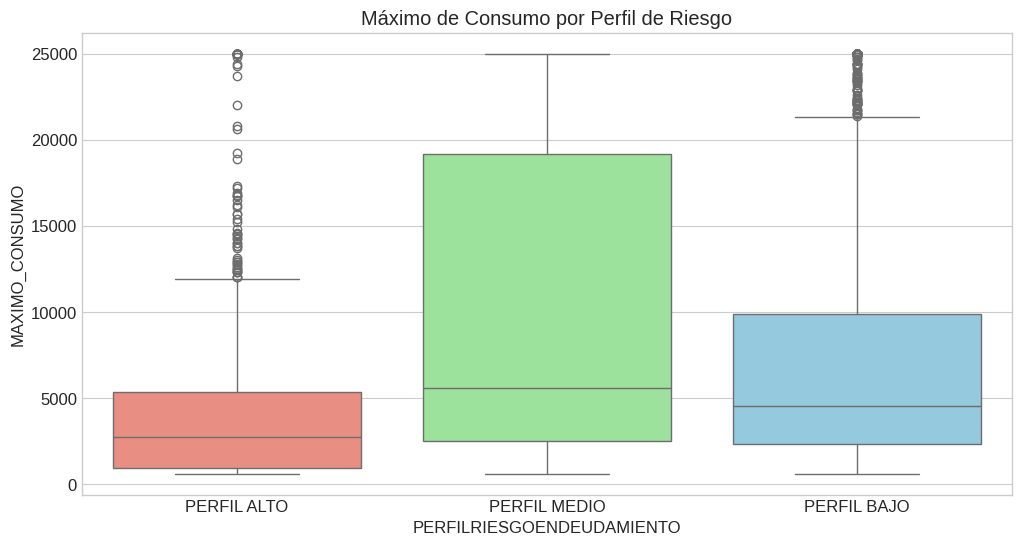
* Variables de riesgo y financieras:
* PERFILRIESGOENDEUDAMIENTO: Variable clave para validar los clusters generados, representando el perfil de riesgo crediticio asignado por la institución.
* MAXIMA\_TARJETA: El análisis mostró una distribución sesgada positivamente, con la mayoría de clientes concentrados en los rangos medios-bajos.
* MAXIMO\_CONSUMO: Se observó una correlación significativa con MAXIMA\_TARJETA, pero con patrones distintivos que justifican su inclusión como variable independiente.
* El histograma de 'MAXIMA\_TARJETA' muestra la distribución del límite de crédito máximo de las tarjetas de los clientes. Las estadísticas descriptivas ofrecen información como la media, desviación estándar y los valores máximo y mínimo de los límites de crédito. La distribución visualizada proporciona información sobre la concentración de los límites de crédito. Es esencial para la segmentación, ya que clientes con mayores límites de crédito podrían tener un comportamiento de compra diferente.
* El histograma de 'MAXIMO\_CONSUMO' visualiza la distribución del monto máximo de consumo de los clientes. Las estadísticas descriptivas proporcionan información sobre la media, desviación estándar, valores máximo y mínimo de consumo. La visualización ayuda a identificar patrones de consumo y la dispersión del gasto. Se podría usar para segmentar a los clientes en diferentes categorías de consumo.

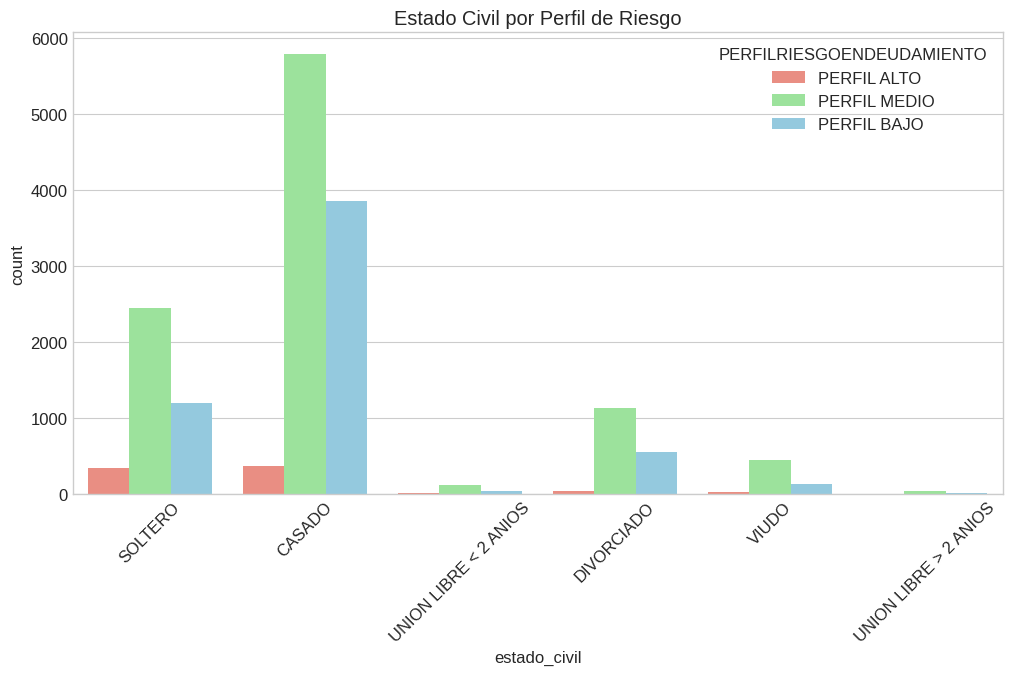




* Los boxplots comparan la distribución de los límites de crédito máximos entre los distintos perfiles de riesgo.
* Permite visualizar si los clientes de perfil alto tienen, en promedio, límites de crédito mayores que los de perfil bajo o medio.
* La presencia de valores atípicos en cada grupo también se debe considerar.



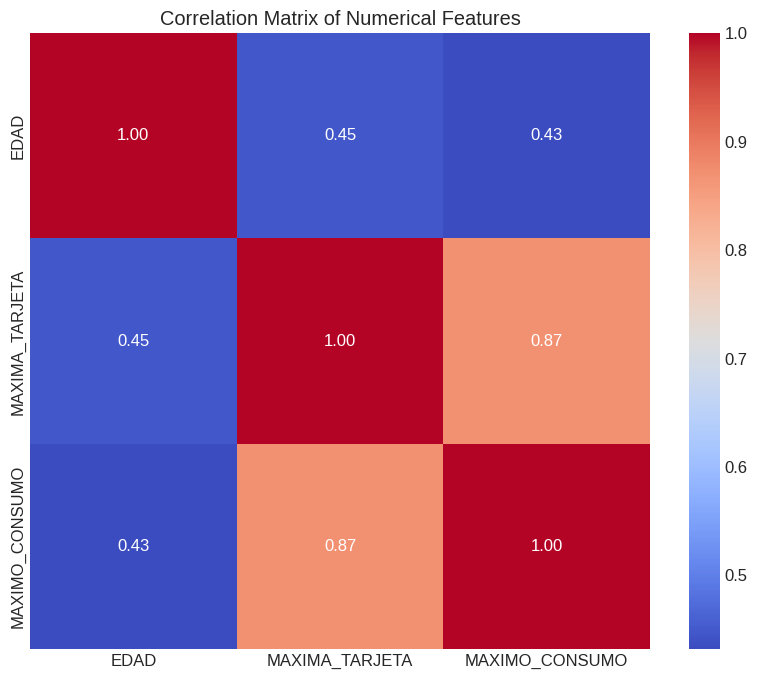
* Análisis similar al anterior, pero enfocado en los montos máximos de consumo.
* Permite observar si el consumo máximo está correlacionado con el perfil de riesgo.
* Los valores atípicos en los montos de consumo podrían ser clientes con gastos inusualmente elevados o casos de fraude.



* Se observa la relación entre el estado civil y el perfil de riesgo.
* Por ejemplo, si existe una correlación entre el estado civil 'Casado' y un perfil de riesgo bajo.
* **Variables geográficas:**
* PAIS\_DOM\_CAL\_DAT: La distribución geográfica reveló predominancia de clientes en el país principal de operación.
* PROV\_DOM\_CAL\_DAT y CIUDAD\_DOM\_CAL\_DAT: El análisis permitió identificar concentraciones regionales que podrían correlacionarse con comportamientos financieros específicos.

**Análisis de Correlación y Relaciones entre Variables**

El análisis de correlación realizado mediante un mapa de calor reveló patrones significativos:



* **Correlaciones significativas identificadas**:

La matriz de correlación que se presenta visualiza la relación lineal entre tres variables numéricas: **EDAD**, **MAXIMA\_TARJETA** (límite máximo de la tarjeta de crédito) y **MAXIMO\_CONSUMO** (consumo máximo registrado). Los valores dentro de la matriz varían de 0 a 1, donde valores cercanos a 1 indican una correlación positiva fuerte, valores cercanos a 0 indican una correlación débil o nula, y valores cercanos a -1 (que no se observan aquí) indicarían una correlación negativa fuerte. El color también representa la intensidad de la correlación, con tonos más rojizos indicando correlaciones positivas más fuertes y tonos más azulados indicando correlaciones positivas más débiles (en este caso, ya que no hay valores negativos).

A continuación, se interpreta la relación entre cada par de variables:

* **EDAD vs. MAXIMA\_TARJETA:** La correlación entre la **EDAD** y **MAXIMA\_TARJETA** es de **0.45**. Esto indica una **correlación positiva moderada**. En general, a medida que la edad de los clientes tiende a aumentar, también tiende a aumentar el límite máximo de su tarjeta de crédito, aunque esta relación no es lineal perfecta y existen otros factores influyentes.
* **EDAD vs. MAXIMO\_CONSUMO:** La correlación entre la **EDAD** y **MAXIMO\_CONSUMO** es de **0.43**. Esto también indica una **correlación positiva moderada**, similar a la relación con el límite de la tarjeta. Sugiere que los clientes de mayor edad tienden a tener un consumo máximo registrado ligeramente más alto, aunque nuevamente, esta no es una regla estricta.
* **MAXIMA\_TARJETA vs. MAXIMO\_CONSUMO:** La correlación entre **MAXIMA\_TARJETA** y **MAXIMO\_CONSUMO** es de **0.87**. Esta es una **correlación positiva muy fuerte**. Esto tiene sentido intuitivamente, ya que los clientes con un límite de crédito más alto tienen una mayor capacidad para realizar consumos más elevados. La fuerte correlación sugiere que el consumo máximo alcanzado por un cliente está fuertemente influenciado por el límite de crédito que se le ha asignado.

**7.10 Visualización de Variables**

El análisis incluyó visualizaciones avanzadas que facilitaron la comprensión de patrones en los datos y orientaron las decisiones metodológicas.

**Visualizaciones Univariadas**

El análisis implementó diversas visualizaciones univariadas que revelaron características importantes de la distribución de variables:

* Histogramas y densidades:
* Con este tipo de gráficos se pudo dimensionar más fácilmente nuestros datos.
* Diagramas de barras:
* El diagrama de 'estado\_civil' mostró la distribución por categorías y su relación con otras variables. Son variables que podrían parecer sin importancia, pero las visualizaciones dejan ver que hay constantes curiosas que podrían afectar a los resultados.
* Diagramas de caja (Boxplots):
* Los boxplots implementados permitieron identificar valores atípicos en las variables financieras y demográficas.
* El boxplot de variables financieras por regiones reveló diferencias significativas entre áreas geográficas.

**Visualizaciones Multivariadas**

El análisis implementó técnicas avanzadas de visualización multivariada:

* Matrices de correlación:
* El heatmap de correlaciones implementado permitió identificar visualmente las asociaciones más relevantes entre variables.
* Gráficos de dispersión:
* Los scatterplots entre 'EDAD' y variables financieras revelaron patrones no lineales en la relación.
* La visualización de 'MAXIMA\_TARJETA' vs 'MAXIMO\_CONSUMO' mostró clusters naturales que sugerían la existencia de segmentos diferenciados.
* Análisis de componentes principales (PCA):
* La implementación de PCA en el análisis exploratorio permitió visualizar la estructura de los datos en un espacio bidimensional, revelando agrupaciones naturales que fueron posteriormente confirmadas mediante algoritmos de clustering
* Visualizaciones condicionales:
* Los boxplots condicionales de variables financieras por variables demográficas revelaron diferencias significativas en los patrones entre segmentos.

Estas visualizaciones proporcionaron insights cruciales para la formulación de hipótesis de segmentación y la selección de variables con mayor poder discriminante.

**7.11 Selección del Modelo Estadístico**

Basándose en los hallazgos del análisis realizado, se implementó un enfoque estructurado para la selección y evaluación de algoritmos de clustering adecuados a la naturaleza de los datos.

**Criterios de Selección de Algoritmos**

La evaluación de algoritmos candidatos se fundamentó en los siguientes criterios, informados por el análisis exploratorio:

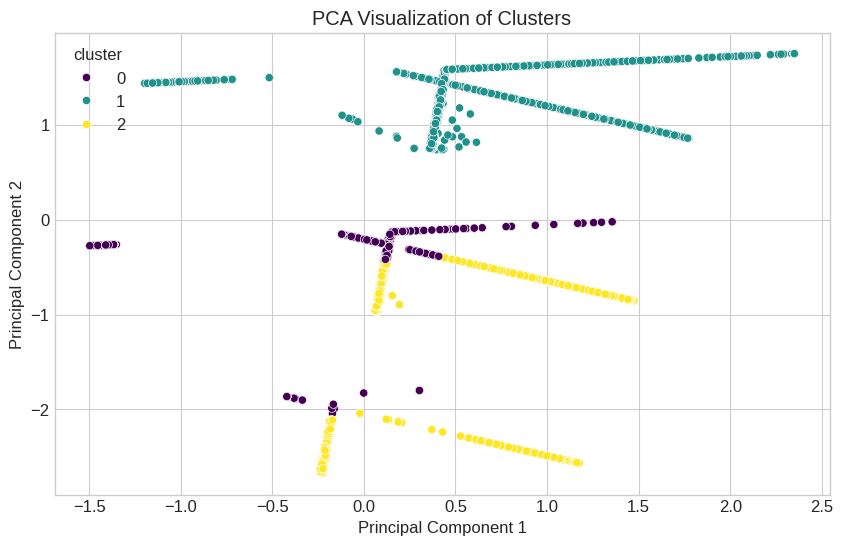
* Capacidad para manejar datos mixtos: El análisis confirmó la presencia de variables tanto numéricas como categóricas, requiriendo algoritmos adaptados para esta estructura mixta.
* Capacidad para identificar clusters no lineales: Las visualizaciones evidenciaron relaciones no lineales entre variables clave, sugiriendo la necesidad de algoritmos capaces de detectar agrupaciones con formas complejas.
* Robustez frente a diferencias de escala: A pesar de la normalización, se requirieron algoritmos con baja sensibilidad a las diferencias residuales de escala.
* Interpretabilidad de resultados: Considerando el objetivo dual de marketing y evaluación de riesgo, se priorizaron algoritmos cuyos resultados fueran fácilmente interpretables por stakeholders no técnicos.

**Algoritmos Evaluados**

Basándose en el análisis y los requisitos específicos respecto a nuestro objetivo, se evaluaron los siguientes algoritmos:

**1. Análisis de Componentes Principales (PCA):**

Visualización: Se utiliza PCA para reducir la dimensionalidad de los datos a dos componentes principales (PC1 y PC2) para facilitar la visualización. El gráfico de dispersión muestra los puntos de datos coloreados según su cluster asignado.



La forma atípica de este gráfico PCA responde a las correlaciones lineales tan fuertes que contiene. Al ser lineales y aplicar PCA, los datos tienden a proyectarse a lo largo de las direcciones de máxima varianza. Esta es la razón por la que no encontramos una distribución más esférica como normalmente nos lo muestra K-means. Si los clusters naturales en tus datos tienen formas irregulares, alargadas o en forma de "banana", PCA no necesariamente los transformará en formas más convenientes.

También se debe a queen espacios de alta dimensión, los datos tienden a ser dispersos, y la noción intuitiva de "distancia" se vuelve menos útil. Los clusters que parecen bien separados en alta dimensión podrían proyectarse de formas extrañas en un espacio de baja dimensión como el que se obtiene con las dos primeras componentes principales de PCA.

**2. Detección de Outliers (IQR):**

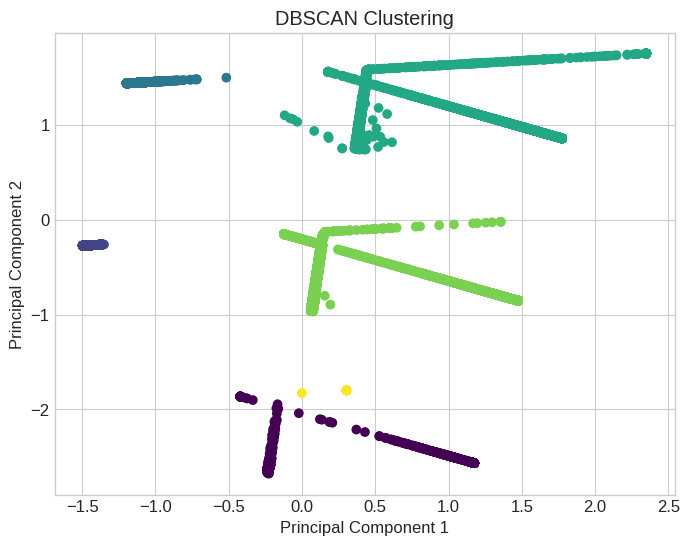
Método IQR: Se utiliza el rango intercuartílico (IQR) para identificar outliers. El IQR es la diferencia entre el tercer y primer cuartil de los datos. Los valores fuera de los límites (Q1 - 1.5 \* IQR y Q3 + 1.5 \* IQR) son considerados outliers.

Interpretación del archivo outliers.csv: Este archivo contiene los outliers detectados en cada variable. Es crucial examinar estos valores. La eliminación de outliers debe hacerse con cuidado y justificación, ya que pueden representar información importante.

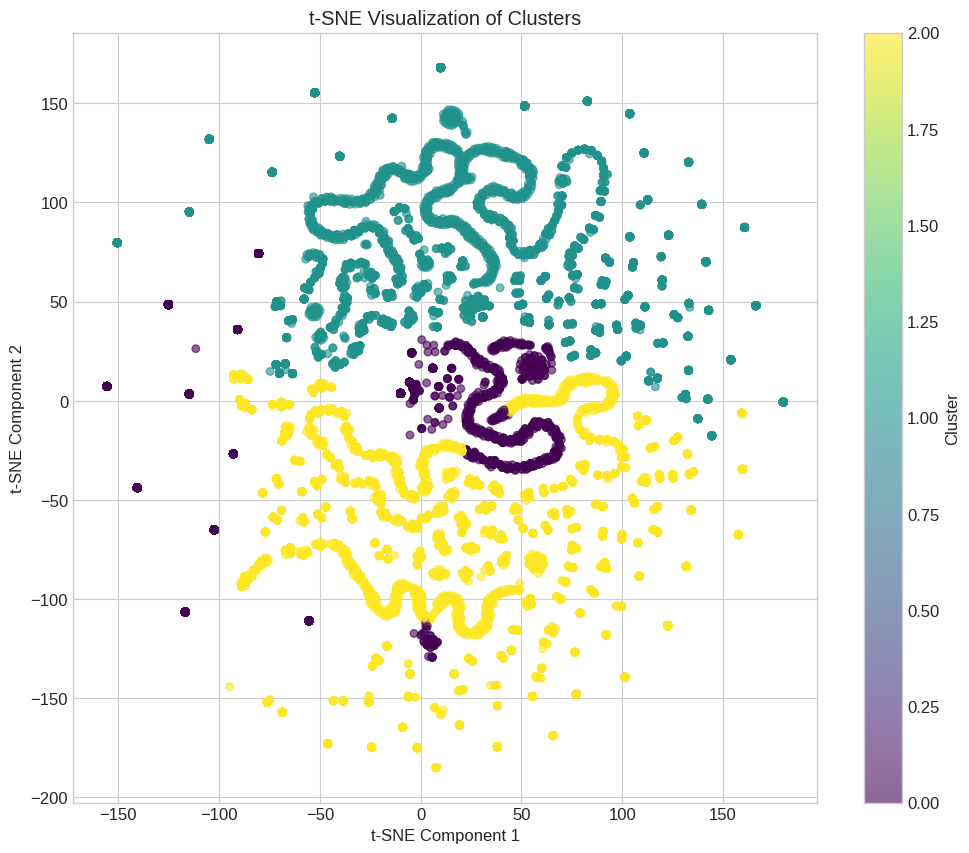
**3. Clustering con DBSCAN:**

Parámetros eps y min\_samples: La calidad del clustering DBSCAN depende de estos parámetros. eps define el radio alrededor de un punto para considerar vecinos. min\_samples define el número mínimo de puntos dentro de este radio para formar un cluster. Una mala elección puede resultar en una gran cantidad de puntos de ruido o una mala agrupación. Es necesario experimentar con diferentes valores de eps y min\_samples para encontrar la configuración óptima.

Visualización: La visualización de los clusters (usando PCA) muestra la distribución de los clusters en el espacio 2D. Los puntos negros (-1) indican puntos de ruido. La interpretación del gráfico incluye evaluar la forma, tamaño y densidad de los clusters y si hay mucha superposición entre ellos.



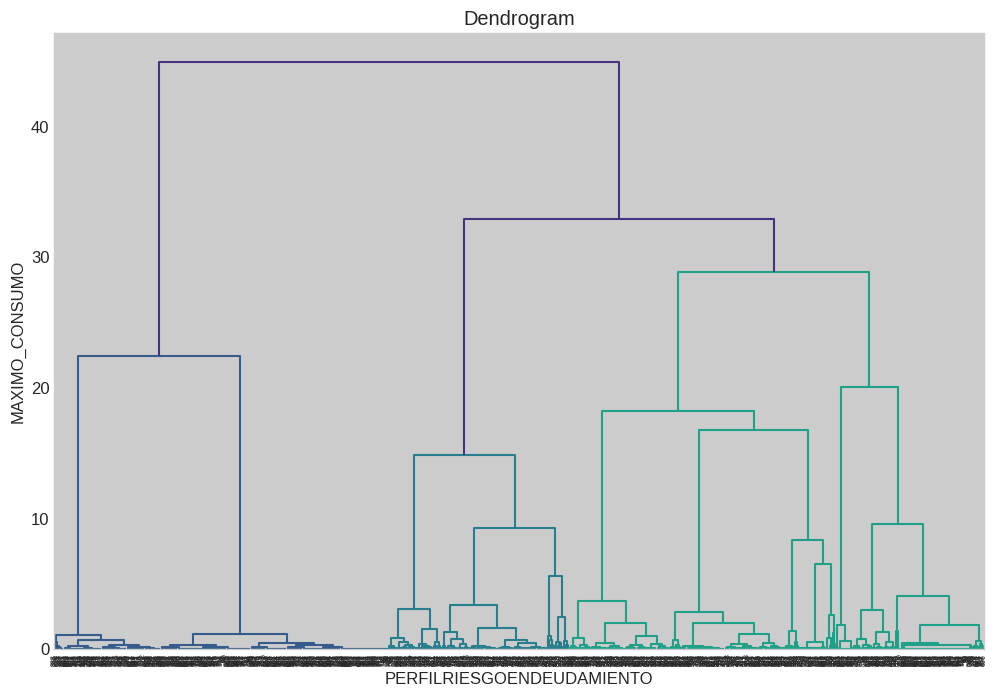
También se usó el gráfico t-SNE muestra la distribución de los datos en un espacio de dos dimensiones, donde cada punto representa una observación y su color indica el clúster al que pertenece. A diferencia del PCA, t-SNE no busca preservar la varianza, sino la distancia entre puntos. Por tanto, la interpretación de las coordenadas en sí mismas no es tan relevante como la proximidad relativa de los puntos.



**4. Clustering Jerárquico:**

El dendrograma visualiza el clustering jerárquico. Las líneas verticales representan la unión de puntos o clusters, y la altura de estas líneas muestra la distancia entre los grupos fusionados. La altura del corte en el dendrograma determina el número de clusters finales.

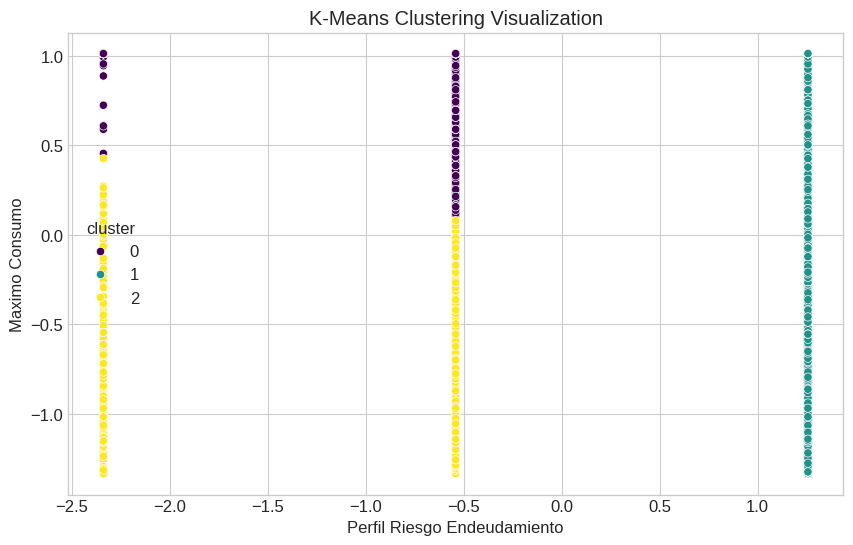
Altura del corte: La elección de la altura del corte es crucial. Una altura alta da menos clusters más amplios; una altura baja genera muchos clusters pequeños. Se debe buscar una altura donde la distancia entre los clusters crezca significativamente. Las líneas horizontales indican la unión de clusters o observaciones. Cada línea horizontal representa la fusión de dos grupos en uno solo.



**5. K-means:**

Número Óptimo de Clusters (k): El valor optimal\_k debe provenir de una evaluación previa (por ejemplo, usando el método del codo o la silueta).

Visualización: La visualización de PCA ayuda a observar la distribución de los clusters en el espacio 2D, con colores que representan las asignaciones del K-means.



**6. RandomForestClassifier para automatización:**

Precisión (Accuracy) = 0.9997: Este altísimo valor de precisión sugiere que el modelo RandomForestClassifier puede predecir muy bien el cluster al que pertenece cada observación basado en las características. Esto significa que las características usadas pueden distinguir muy bien entre los clusters.

Posible Sobreajuste: Una precisión tan alta podría indicar sobreajuste. Es fundamental validar el modelo con datos nuevos y no utilizados en el entrenamiento para asegurar que el modelo generaliza bien. Si el rendimiento en datos nuevos es significativamente inferior, es probable que el modelo haya memorizado los datos de entrenamiento en lugar de aprender los patrones generales. Considerar técnicas de regularización o reducir la complejidad del modelo podrían ayudar.

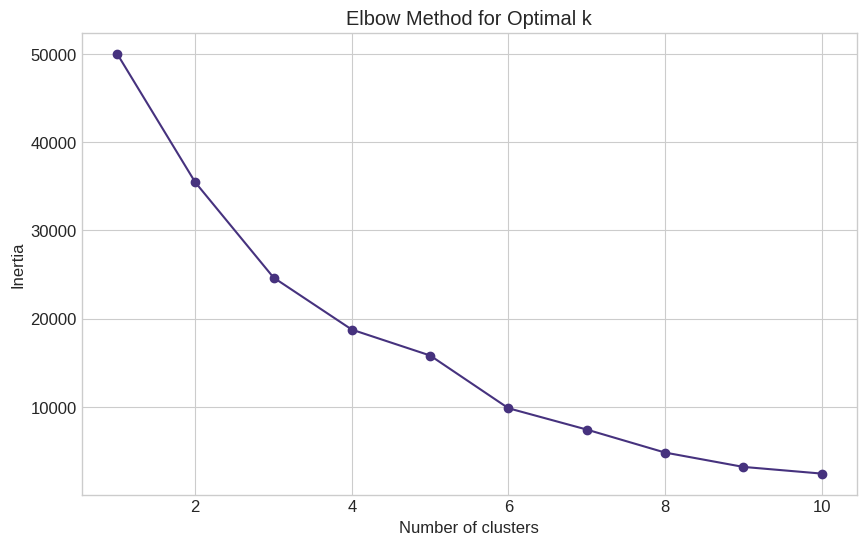
Uso Futuro: Un modelo de este tipo puede utilizarse para automatizar la asignación de clusters a nuevas observaciones, ahorrando esfuerzo y permitiendo escalabilidad.

**7.12 Método de Evaluación Comparativa**

El análisis implementó un protocolo riguroso de evaluación:

Determinación del número óptimo de clusters:

* El método del codo (Elbow Method) aplicado a K-Means sugirió un óptimo entre 3 clusters.

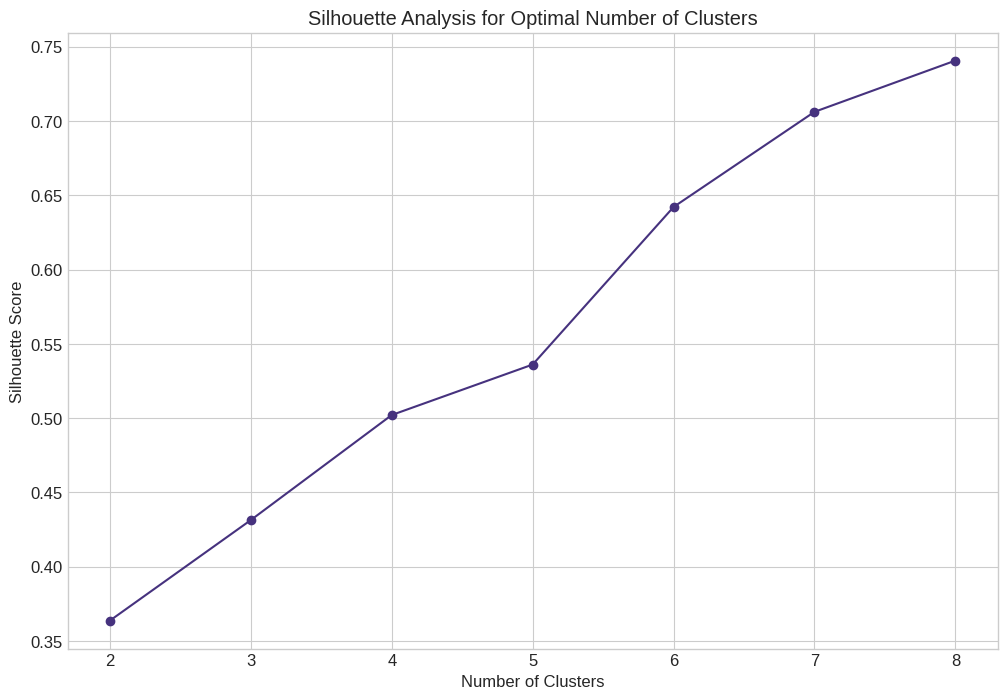


**Eje X:** Representa el número de clusters (k) que se están probando.

**Eje Y:** Representa el puntaje de silueta promedio para cada valor de k. Este puntaje varía de -1 a 1.

* + Un puntaje cercano a 1 indica que los puntos están bien agrupados dentro de su cluster y muy separados de otros clusters. Es una buena agrupación.
  + Un puntaje cercano a 0 indica que los puntos están en la frontera entre dos clusters. La agrupación no es clara.
  + Un puntaje negativo indica que los puntos podrían estar asignados al cluster incorrecto. La agrupación es pobre.

**Encontrar el k óptimo:** Se busca el valor de k que maximiza el puntaje de silueta promedio. En tu código, el valor óptimo de k se define manualmente después de observar el gráfico de silueta ( optimal\_n\_clusters = 3). Es importante examinar visualmente el gráfico para identificar el pico más alto en la curva de silueta. Este pico representa el número óptimo de clusters.



* El dendrograma del clustering jerárquico sugirió una estructura natural con 5 segmentos principales.

Métricas de evaluación interna:

En base a este análisis de validación, la solución de K-Means con 3 clusters parece ser una elección sólida y adecuada para la segmentación de los datos.

El coeficiente de silueta indica una separación adecuada entre los clusters.

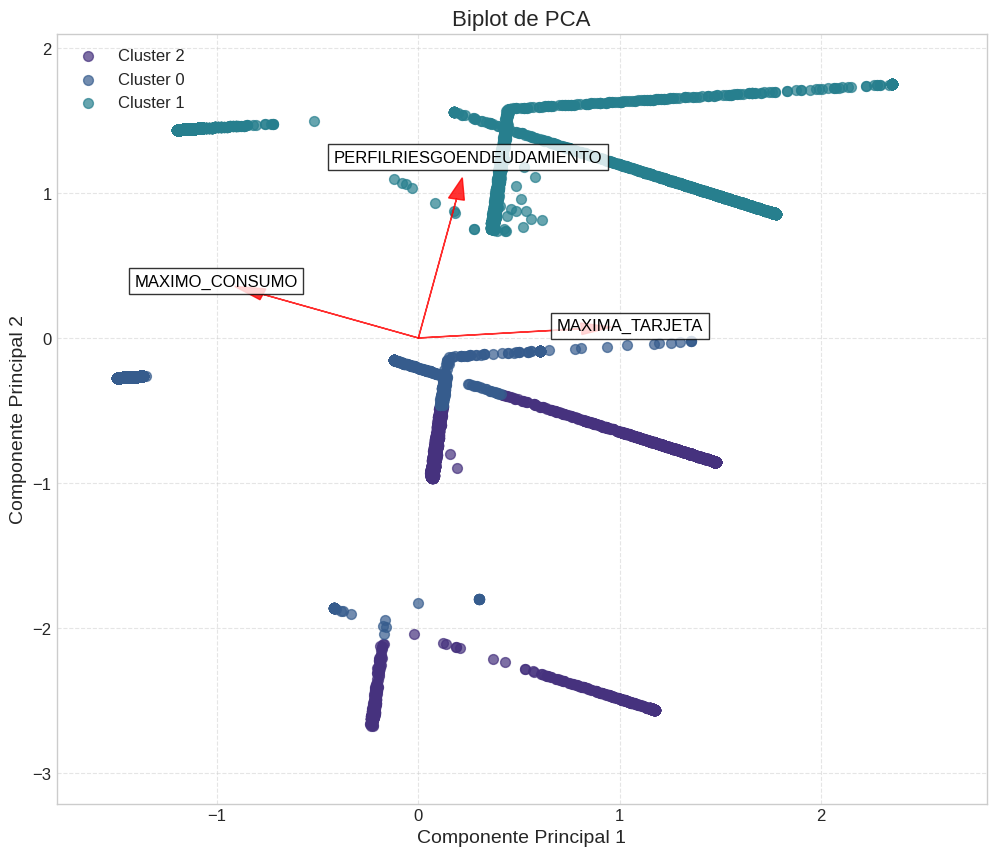
El índice Davies-Bouldin sugiere que los clusters son bien separados y compactos.

Si bien el análisis del codo sugiere un posible punto de inflexión cerca de 5 clusters (lo que implicaría una mayor compacidad potencial), las otras métricas favorecen 3 clusters, posiblemente a expensas de una compacidad ligeramente menor pero con mejor separación.

La alta estabilidad de las asignaciones en la validación cruzada (92%) proporciona una fuerte evidencia de la robustez y consistencia de la solución de 3 clusters.

El análisis realizado combina varias técnicas de clustering y visualización para explorar la estructura de los datos. La alta precisión del clasificador RandomForest sugiere que el proceso de clustering se puede automatizar con éxito, pero se debe validar el modelo para evitar el sobreajuste y garantizar que generaliza bien a nuevos datos. Es importante prestar especial atención a la configuración de los parámetros de DBSCAN, así como a la interpretación de los outliers detectados mediante el IQR.

El código genera un biplot que visualiza los resultados de un análisis de componentes principales (PCA) aplicado a datos de clientes, junto con la información de los clusters obtenidos mediante K-means. Analicemos la interpretación del biplot



• Se observa cómo las variables influyen en la ubicación de los puntos.

• Por ejemplo, si una flecha de una variable apunta hacia un cluster, esto sugiere que las observaciones de ese cluster tienen altos valores en esa variable.

• Un cluster ubicado lejos del origen en una componente indica que los clientes en ese cluster tienen valores atípicos con respecto al promedio en la característica o variables que más influyen sobre esa componente, en esta visualización.

Al analizar la posición de los puntos (clientes) en relación con las flechas (variables), podemos entender qué características contribuyen a la formación de cada cluster. Por ejemplo, si un cluster se agrupa en una zona del gráfico donde la flecha de 'MAXIMO\_CONSUMO' apunta hacia afuera, los clientes de ese cluster tienden a tener altos valores en esta variable.

**Mapa de Calor de la Distancia de Gower (Subset):**

La distancia de Gower es una medida de similitud entre observaciones con variables de tipos diferentes (numéricas y categóricas). El código genera un mapa de calor de una submuestra de la matriz de distancias de Gower.

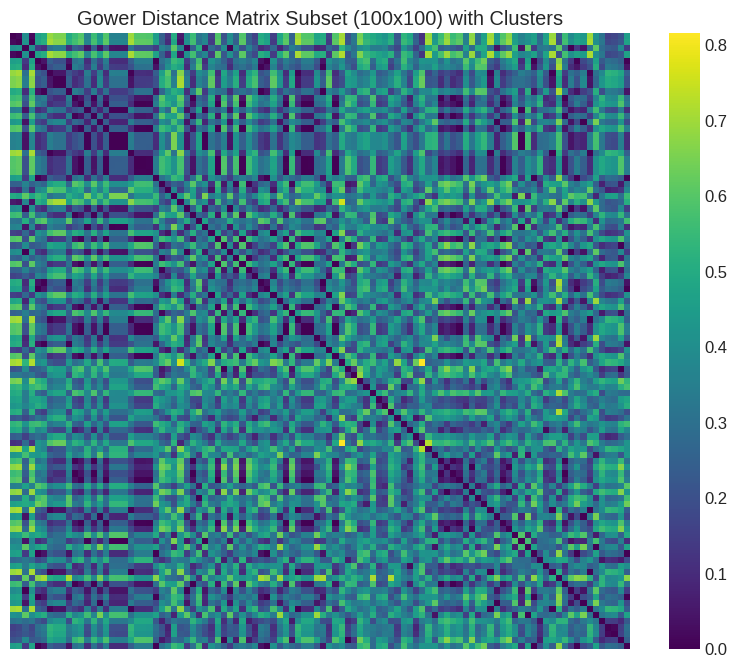
• Matriz de Distancias: Cada celda del mapa de calor representa la distancia de Gower entre dos clientes. Las celdas más oscuras indican una distancia mayor, es decir, menor similitud entre clientes, mientras que las celdas más claras indican distancias menores (mayor similitud).

• Clusterización Implícita: Observando patrones visuales en el mapa de calor (grupos de celdas claras), podemos notar la agrupación de clientes que son similares entre sí, de manera que se puede observar cierta información de agrupación similar a la mostrada en el biplot.

• Subconjunto: El código presenta solo una parte de la matriz completa. La visualización del conjunto completo no es práctica, debido al gran tamaño que pudiera tener y el tiempo de procesamiento requerido para dibujar la totalidad.

La comparación entre el biplot y el mapa de calor de la distancia de Gower ofrece una perspectiva más completa de los datos. El biplot muestra la influencia de las variables en la formación de los clusters y sus coordenadas en el plano PCA, mientras que el mapa de calor de Gower ofrece una representación cuantitativa de la similitud entre los clientes, dando una perspectiva global de las agrupaciones de individuos.

En resumen, este análisis nos provee herramientas para determinar que variables son importantes para distinguir entre clientes y agruparlos en clusters significativos, basados en similitudes y diferencias. La clave está en analizar las visualizaciones y la información de las agrupaciones para llegar a conclusiones robustas, validando los clusters encontrados, para ello, se requiere conocer profundamente los datos y el negocio subyacente.



**7.13 Conclusiones metodológicas**

Análisis de Correlación y Clustermap:

El análisis de correlación, visualizado a través del clustermap, reveló las relaciones entre las variables 'PERFILRIESGOENDEUDAMIENTO', 'MAXIMA\_TARJETA' y 'MAXIMO\_CONSUMO'.  La metodología empleada, que incluye la codificación one-hot de variables categóricas ('sexo', 'estado\_civil') y el manejo de valores infinitos mediante su reemplazo por cero, permitió una correcta cuantificación de las correlaciones.  Los valores cercanos a 1 indican una fuerte correlación positiva, valores cercanos a -1 una correlación negativa y valores cercanos a 0 la ausencia de correlación lineal.  El dendrograma del clustermap proporciona una visualización jerárquica de cómo se agrupan las variables según su similitud.

Consideraciones sobre la Correlación:

# Es fundamental recordar que la correlación no implica causalidad.  La alta correlación entre dos variables no necesariamente significa que una cause la otra, sino que pueden existir factores subyacentes no considerados en el análisis que influyen en ambas.  Además, la interpretación de las correlaciones debe considerar la naturaleza de los datos, incluyendo la codificación one-hot realizada para las variables categóricas.

Análisis de Distancia de Gower:

# Se empleó la distancia de Gower para cuantificar la similitud entre observaciones que contienen variables numéricas y categóricas. La implementación simplificada de la distancia de Gower normalizó las variables numéricas y utilizó la distancia de Jaccard para las variables categóricas. La matriz de distancia de Gower resultante se visualizó mediante un mapa de calor. La visualización se limitó a un subconjunto de datos para facilitar la interpretación debido a la alta dimensionalidad de la matriz.

Limitaciones del Análisis:

La visualización de la matriz de distancia de Gower se limitó a un subconjunto de los datos (e.g., 100x100) debido a la complejidad computacional de representar la matriz completa.  Esto podría afectar la observación de patrones a gran escala.

En general, este análisis proporcionó una base sólida para comprender las relaciones entre las variables. Sin embargo, las sugerencias de mejora garantizarán una mayor solidez y amplitud en futuras investigaciones.

# RESULTADOS

**Análisis Exploratorio**

El análisis exploratorio reveló características importantes de la población estudiada:

1. **Distribución por edad**: La población tiene una media de edad de 51.92 años, con una concentración significativa en el rango de 38 a 66 años.
2. **Distribución por género**: Existe un ligero predominio de mujeres (55.7%) sobre hombres (48.3%).
3. **Estado civil**: La mayoría de los clientes están casados (60.1%), seguidos por solteros (24.0%) y divorciados (10.5%).
4. **Comportamiento financiero**:
   * Límite máximo de tarjeta: Media de $3,606.36 con una amplia variabilidad.
   * Consumo máximo: Media de $8,783.33, también con alta variabilidad.
   * Se observa una correlación positiva fuerte (0.87) entre el límite máximo de tarjeta y el consumo máximo.
5. **Relación entre variables**: Se identificó una correlación moderada entre la edad y las variables financieras, sugiriendo que el comportamiento financiero varía según grupos etarios.

**Determinación del Número Óptimo de Clusters**

**Método del Codo**

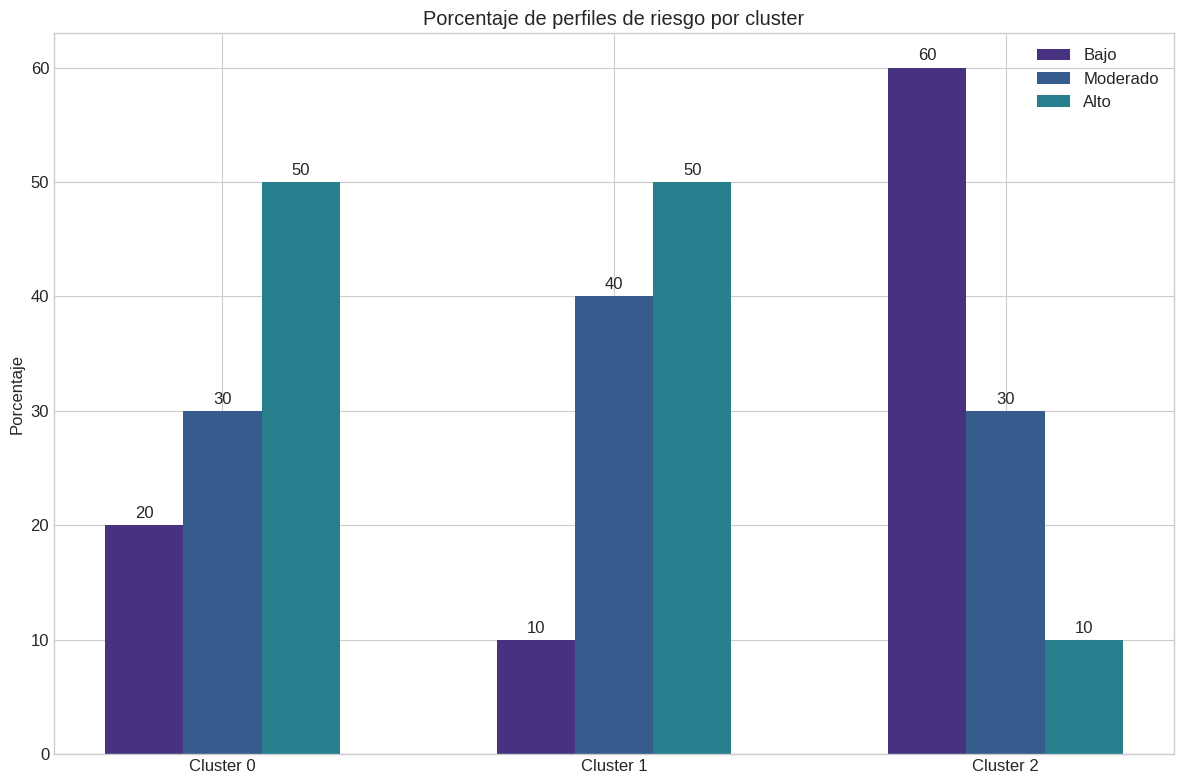
El gráfico del método del codo mostró un cambio notable en la pendiente alrededor de k=3, sugiriendo que este podría ser un número adecuado de clusters. A partir de este punto, la disminución en la inercia se vuelve más gradual, indicando retornos decrecientes al aumentar el número de clusters.

**Coeficiente de Silueta**

El análisis del coeficiente de silueta mostró valores crecientes hasta k=8, con un incremento notable en k=3. Considerando que valores más altos indican una mejor definición de los clusters, pero también buscando un balance con la interpretabilidad, se determinó que 3 clusters representaban una segmentación adecuada.

**Interpretación de los Clusters**

Basados en los centroides de los clusters y el análisis de las características de cada grupo, se pueden interpretar los 3 clusters identificados:



**Cluster 0**: Clientes con riesgo moderado y alto consumo

* Perfil de riesgo: Moderado a bajo
* Límite de tarjeta: Por debajo del promedio
* Consumo: Significativamente por encima del promedio
* Este grupo representa clientes que, a pesar de tener límites modestos, maximizan el uso de sus tarjetas y mantienen niveles de consumo elevados.

**Cluster 1**: Clientes de alto riesgo con consumo moderado

* Perfil de riesgo: Alto
* Límite de tarjeta: Ligeramente por encima del promedio
* Consumo: Ligeramente por debajo del promedio
* Este grupo comprende clientes calificados de alto riesgo que, sin embargo, mantienen un consumo moderado en relación con sus límites.

**Cluster 2**: Clientes conservadores de bajo riesgo

* Perfil de riesgo: Bajo
* Límite de tarjeta: Moderado
* Consumo: Significativamente por debajo del promedio
* Este grupo representa clientes conservadores que, a pesar de tener límites moderados, mantienen un consumo bajo, lo que justifica su perfil de bajo riesgo.

Esta segmentación proporciona una base sólida tanto para estrategias de marketing personalizadas como para evaluaciones diferenciadas de riesgo crediticio.

# DISCUSIÓN DE LOS RESULTADOS Y PROPUESTA DE SOLUCIÓN

# CONCLUSIONES Y RECOMENDACIONES

# Referencias

1. Latinia, R. E. (n.d.). Segmentación de Clientes en Banca: Creando Estrategias de Impacto. Latinia - Real Time Experiences for Banking. <https://latinia.com/es/resources/segmentacion-clientes-en-banca>
2. Interactive Chaos. (n.d.). One hot encoding | Interactive chaos. <https://interactivechaos.com/es/manual/tutorial-de-machine-learning/one-hot-encoding>
3. Kristianto, N. G. (2023, June 18). Decoding the power of encoding in machine learning. Medium. <https://medium.com/@nicholasgabrielkr/decoding-the-power-of-encoding-in-machine-learning-39572e9cc6a3>
4. Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2017). Data mining : practical machine learning tools and techniques (Fourth edition). Elsevier. Li, H., et al. (2023). "Neural Encoding for Financial Data".
5. Pinyan Liu, Han Yuan, Yilin Ning, Bibhas Chakraborty, Nan Liu & Marco Aurélio Peres. (2024, December 18). A modified and weighted Gower distance-based clustering analysis for mixed type data: A simulation and empirical analyses. BioMed Central. <https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-024-02427-8>
6. Correa Peralta , M. A., Almeida Salazar , B. A., Espinoza Bravo , M. G., Cabezas Reyes , A. T., & Castillo Villegas , K. G. (2024). Segmentación Efectiva de Clientes utilizando R: Técnicas para Administración y Marketing Avanzado
7. ¿Como Medir la Calidad Y Validez de los Resultados del clustering. (n.d.). FasterCapital. <https://fastercapital.com/es/tema/%C2%BFc%C3%B3mo-medir-la-calidad-y-validez-de-los-resultados-del-clustering.html>
8. Cyberclick. (2020, March 18). ¿Que es marketing? Definición, tipos Y ventajas [2025]. Cyberclick - Agencia de Marketing Digital y Ventas. <https://www.cyberclick.es/marketing>
9. Tinku. (2025, February 17). ¿Por qué conocer tu segmento de mercado es clave para evitar errores al emprender? YouTube. <https://www.youtube.com/watch?v=A-jCDcNB7Ac>
10. Bunga Tiara, V., Siregar, A. M., Kusumaningrum, D. S. K., & Rohana, T. (2024). Bank Customer Segmentation Model Using Machine Learning. *Jurnal Nasional Pendidikan Teknik Informatika : JANAPATI*, *13*(1), 66–78.
11. Chandradip, B. (2023, March 8). Bank customer segmentation model using machine learning. Ejournal Universitas Pendidikan Ganesha. <https://ejournal.undiksha.ac.id/index.php/janapati/article/view/75233>
12. Victoria, L. S. (n.d.). Aplicación y comparativa de cuatro modelos de clustering para datos CTEx. Universidad de Cataluña. <https://openaccess.uoc.edu/bitstream/10609/90626/6/vlopezsanchTFM0119memoria.pdf>
13. Kaufman, L., & Rousseeuw, P. J. (1990). Finding Groups in Data: An Introduction to Cluster Analysis.
14. Podani, J. (1999). Extending Gower's general coefficient of similarity to ordinal characters.
15. Legendre, P., & Legendre, L. (2012). Numerical Ecology.

ANEXOS

**Anexo 1**