



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

2. 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 stakeholders 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.

2.3.1 Matriz de Investigaciones Similares

Estudio	Tipos de Datos	Metodología	Resultados	Implicaciones Gerenciales
Chandradip (2023)	- Datos transaccionales- Historiales crediticios- Información demográfica	Label Encoding para variables ordinales y Machine Learning	Mejora en la precisión de segmentación del 15%	Optimización de estrategias de marketing personalizadas
Bunga Tiara et al. (2024)	- Datos comportamentales- Patrones de consumo- Datos socioeconómicos	Clustering jerárquico con distancia de Gower	Identificación de 5 segmentos distintivos de clientes	Desarrollo de productos financieros específicos por segmento
Correa Peralta et al. (2024)	- Datos de marketing- Información de productos- Historiales de interacción	Técnicas avanzadas de R para administración	Incremento del 25% en efectividad de campañas	Mejora en la retención de clientes y cross-selling
Kristianto (2023)	- Variables categóricas- Datos numéricos mixtos	Encoding avanzado y análisis predictivo	Reducción del 30% en dimensionalidad de datos	Optimización de recursos en campañas de marketing

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.

2.4.4 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.

3. IDENTIFICACIÓN DEL OBJETO DE ESTUDIO

Esta 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.

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.

4. 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.

5. 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.

6. 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.

JUSTIFICACIÓN Y APLICACIÓN DE LA METODOLOGÍA

RESULTADOS

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://bmcmmedresmethodol.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 CTEEx. 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