# Optimización de campañas de marketing en un centro de comercio mediante segmentación de clientes

Ian Lozano Ruiz

Asesor

Luis Angel Anillo Arrieta

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos Analítica

2024

## Dedicatoria

A mi familia, por ser mi pilar inquebrantable y mi mayor fuente de inspiración. A mis padres, por su amor incondicional, su apoyo inagotable y por enseñarme el valor del esfuerzo y la perseverancia. A mi hermano, por creer en mí en todo momento y por estar a mi lado en cada paso de este camino. Este logro es tan suyo como mío, pues sin ustedes no habría sido posible. Gracias por su paciencia, por su aliento y por cada palabra de ánimo cuando más lo necesitaba. Este trabajo es una expresión de todo lo que hemos construido juntos.

Con todo mi amor y gratitud.

**Agradecimientos**

Agradezco a la Universidad Nacional Abierta y a Distancia (UNAD) por brindarme la oportunidad de desarrollar mis estudios de especialización en ciencia de datos y analítica. Su enfoque innovador y flexible ha sido clave para mi crecimiento académico y profesional, permitiéndome avanzar en este fascinante campo.

De manera especial, extiendo mi gratitud al profesor Luis Ángel Anillo Arrieta, mi asesor, por su invaluable guía, paciencia y apoyo durante todo este proceso. Sus orientaciones han sido fundamentales para la culminación de este trabajo, y su compromiso con mi formación me ha impulsado a alcanzar mis metas.

A mis profesores, quienes me transmitieron su conocimiento con dedicación y pasión, les agradezco por los aprendizajes que han dejado huella en mi desarrollo académico y profesional. Cada lección fue un paso más en este camino.

Finalmente, a mis compañeros de estudios, quienes con su camaradería y espíritu colaborativo han hecho de este recorrido una experiencia más enriquecedora. Juntos hemos superado retos y compartido logros, y por ello, siempre les estaré agradecido.

**Resumen**

La ciencia de datos ofrece métodos avanzados para identificar oportunidades de mercadeo y segmentar clientes en la industria, a través del análisis de clúster, se pueden descubrir segmentos de clientes con comportamientos o características similares con el fin de dirigir campañas de marketing más efectivas al identificar grupos más susceptibles a ciertas promociones o productos, el análisis de canasta de compras revela patrones de compra y facilita la gestión de inventarios al predecir la demanda de productos de diferentes segmentos de clientes. Los modelos predictivos permiten anticipar necesidades futuras mejorando la retención y la percepción de los consumidores, contribuyendo a mejorar la satisfacción del usuario al asegurarse de que sus deseos son comprendidos y atendidos de manera más efectiva. De esta manera, la ciencia de datos proporciona insights valiosos que pueden transformar la manera en que un centro comercial interactúa con sus clientes, mejorando así la experiencia de usuario y aumentando la eficiencia operativa.

***Palabras claves*****:** Clúster, Marketing, Patrones, Retención.

## Abstract

Data science offers advanced methods to identify marketing opportunities and segment customers within the industry. Through cluster analysis, it is possible to discover customer segments with similar behaviors or characteristics in order to target marketing campaigns more effectively, identifying groups more susceptible to certain promotions or products. Basket analysis reveals purchase patterns and facilitates inventory management by predicting product demand across different customer segments. Predictive models allow for the anticipation of future needs, improving customer retention and perception, thus enhancing user satisfaction by ensuring that their desires are better understood and met. In summary, data science provides valuable insights that can transform how a shopping center interacts with its customers, thereby improving the user experience and increasing operational efficiency.


***Keywords*****:** Clúster, Marketing, Patterns, Retention.

# Tabla de Contenido

## Lista de Tablas

## Lista de Figuras

## Introducción

En la actualidad, la capacidad de las empresas para adaptarse rápidamente a las cambiantes demandas del mercado depende en gran medida de su habilidad para captar y analizar datos. La ciencia de datos ha emergido como una disciplina clave en este ámbito, ofreciendo herramientas avanzadas que permiten a los negocios tomar decisiones basadas en datos, identificar oportunidades de mercadeo, y segmentar clientes de manera efectiva. Este proyecto de grado tiene como objetivo principal optimizar el proceso de segmentación de clientes en un centro de comercio mediante el uso de algoritmos de Machine Learning, específicamente el algoritmo de clustering K-means.

El análisis de clúster permite descubrir segmentos de clientes con comportamientos o características similares, lo que habilita a los comerciantes para dirigir campañas de marketing más efectivas y personalizadas. Además, el análisis de patrones de compra facilita la gestión de inventarios al predecir la demanda de productos entre diferentes segmentos de clientes. Al integrar estas técnicas en las estrategias de mercadeo de un centro comercial, se busca mejorar la eficiencia operativa, aumentar la competitividad, y ofrecer a los clientes experiencias más satisfactorias y personalizadas.

Este trabajo se enfoca en abordar los desafíos actuales en la gestión de datos y segmentación de clientes, proporcionando un enfoque innovador para mejorar las campañas de marketing y la retención de clientes, todo ello apoyado en técnicas avanzadas de Machine Learning y análisis de datos.

## Planteamiento del Problema

El problema central identificado es la incapacidad de los empresarios y comerciantes para capturar y gestionar eficazmente los datos derivados del comportamiento y las preferencias de los clientes (Lee, Lee, Hsin, & Fang, 2024). Esta limitación surge, en gran medida, por la falta de herramientas analíticas avanzadas y la dificultad para integrar datos provenientes de diversas fuentes. Como consecuencia, los comerciantes no logran personalizar de manera efectiva la experiencia del cliente, lo que impacta negativamente en la competitividad y la eficiencia operativa de sus negocios (Rachman, Santoso, & Djajadi, 2021).

Este problema se manifiesta en campañas de marketing poco dirigidas y no relevantes, una gestión ineficiente de inventarios debido a predicciones inexactas de la demanda, y una comprensión insuficiente de las necesidades y deseos del cliente, lo que impide ofrecer una experiencia satisfactoria (Sun, Liu, & Gao, 2023). En un mercado saturado, la capacidad de personalizar la experiencia del cliente se ha vuelto un diferenciador clave para atraer y retener clientes, así como para optimizar recursos y aumentar la rentabilidad a largo plazo (Luo, Li, Fan, Wang, Koprinska, & Chen, 2023).

Además, la incapacidad para gestionar los datos de manera eficaz limita la capacidad de las empresas para adaptarse a las demandas cambiantes del mercado. Esta situación provoca una segmentación imprecisa de los clientes, resultando en campañas de marketing menos efectivas y una asignación ineficiente de recursos, lo que afecta directamente la competitividad y la rentabilidad (Yadegaridehkordi et al., 2021). Aunque las empresas han intentado abordar este problema mediante la adopción de tecnologías de análisis de datos y pronósticos, la integración de datos sigue siendo un desafío, lo que impide la optimización de estrategias de marketing y la mejora en la gestión de inventarios.

A pesar de los avances tecnológicos, aún quedan áreas inexploradas que requieren mayor atención, como la integración de datos de comportamiento en tiempo real y la adaptación dinámica de las estrategias de marketing (Rachman, Santoso, & Djajadi, 2021). La implementación de modelos de segmentación más precisos y personalizados podría ayudar a las empresas a superar estos desafíos, mejorando su capacidad de respuesta a las necesidades del mercado y optimizando tanto sus campañas de marketing como la gestión de inventarios.

## Justificación

La creciente competencia en el mercado actual obliga a las empresas a adoptar estrategias más eficientes y personalizadas para captar y retener clientes (Griva et al., 2024). En este contexto, la ciencia de datos ha demostrado ser una herramienta esencial para la optimización de campañas de marketing, permitiendo a las empresas identificar patrones de comportamiento de los clientes y segmentarlos de manera más precisa. La personalización de las campañas no solo mejora la efectividad de los esfuerzos de marketing, sino que también contribuye a una mejor gestión de inventarios y a la predicción de la demanda, lo que resulta en una mayor eficiencia operativa.

Este proyecto es relevante porque busca resolver el problema de la segmentación ineficiente de clientes, un desafío que impacta negativamente la rentabilidad y competitividad de los centros de comercio (Nilashi et al., 2021).. Al implementar técnicas de Machine Learning, como el clustering, este estudio proporciona una solución innovadora que no solo mejora la precisión en la segmentación, sino que también permite a las empresas personalizar las interacciones con sus clientes, optimizando los recursos destinados a las campañas de marketing y mejorando la experiencia del usuario (Sun et al., 2023).

La aplicación de modelos predictivos y el análisis en tiempo real de los datos de los clientes ofrecen una ventaja competitiva crucial, permitiendo a las empresas anticipar las necesidades de sus consumidores y adaptarse rápidamente a las fluctuaciones del mercado. En un entorno cada vez más saturado, donde los consumidores buscan experiencias personalizadas y relevantes, este proyecto tiene el potencial de transformar la manera en que los centros comerciales interactúan con sus clientes, generando un impacto positivo en la retención y satisfacción del cliente (Yadegaridehkordi et al., 2021).

# Objetivos

## Objetivo General

Optimizar el proceso de segmentación de clientes del marketing a través del machine learning en un centro de comercio.

## Objetivos Específicos

Identificar el modelo de Machine Learning óptimo para la segmentación en campañas de marketing mediante una revisión exhaustiva de la literatura académica y estudios de caso.

Desarrollar un modelo de Machine Learning que optimice la segmentación de clientes en campañas de marketing, ajustado a las características específicas de la base de datos existente.

Implementar y adaptar las campañas de marketing a cada segmento definido por el modelo de Machine Learning, ajustando el contenido y los canales de distribución para maximizar la efectividad de cada segmento.

## Marco Conceptual

El concepto de Gestión de Relaciones con el Cliente (CRM) representa un enfoque empresarial centrado en la optimización de las interacciones con los consumidores (Sun, Liu, & Gao., 2023). Este modelo se basa en el uso de herramientas tecnológicas que permiten gestionar eficientemente dichas relaciones, mejorando la experiencia del cliente y maximizando el valor a largo plazo que este puede generar para la empresa (Nguyen., 2021). La segmentación de clientes es uno de los procesos clave dentro del CRM. Este método permite clasificar a los consumidores en grupos homogéneos según sus características, necesidades o comportamientos compartidos, facilitando la focalización de recursos hacia los segmentos más lucrativos y aumentando la efectividad de las campañas de marketing.

Un aspecto esencial en este contexto es la personalización del marketing, que implica diseñar mensajes y seleccionar canales de comunicación específicos para cada segmento de clientes (Sun, Liu, & Gao., 2023). Esto permite a las empresas conectar más efectivamente con sus clientes, mejorando la relevancia de las campañas y aumentando las oportunidades de ventas y fidelización. Además, el Valor de Vida del Cliente (CLV), una métrica clave dentro del CRM, mide el valor total que un cliente genera para la empresa a lo largo de su relación con ella (Lee, Lee, Hsin, & Fang., 2024). El CLV es crucial para entender el impacto financiero de los clientes y guiar las estrategias de marketing hacia aquellos con mayor rentabilidad potencial.

## Marco Teórico

En el marco teórico de este proyecto, se utilizará el algoritmo de clustering K-means como técnica principal para la segmentación de clientes (Rachman, Santoso, & Djajadi., 2021). El clustering es una herramienta clave en ciencia de datos y marketing, ya que permite agrupar a los clientes en clústeres basados en características comunes, lo que facilita la personalización de estrategias de marketing y mejora la eficiencia en la asignación de recursos. El proceso comienza con la exploración y preprocesamiento de los datos, donde se verifica la integridad de los mismos y se codifican variables categóricas en formatos numéricos, asegurando que los datos sean adecuados para el algoritmo de clustering (Rachman, Santoso, & Djajadi., 2021).

El método de determinación del número de clústeres, como el "método del codo", ayuda a identificar cuántos clústeres son óptimos, equilibrando la complejidad del modelo con la cohesión dentro de los grupos formados. Este enfoque permite maximizar la precisión en la segmentación de clientes, garantizando que los segmentos formados sean homogéneos y relevantes para las estrategias de marketing (Mirfakhraei, Abdolvand, & Harandi., 2024).

En cuanto a la relación con las tendencias actuales, el uso de K-means está alineado con la creciente adopción de algoritmos de Machine Learning en el ámbito del marketing moderno. Estos modelos permiten a las empresas extraer información valiosa de grandes volúmenes de datos y automatizar la toma de decisiones clave, como la segmentación de clientes y la personalización de campañas. Esto no solo aumenta la eficacia del marketing y la retención de clientes, sino que también mejora el Valor de Vida del Cliente (CLV), ya que se pueden identificar y priorizar aquellos segmentos de mayor valor para la empresa (Lee, Lee, Hsin, & Fang., 2024).

## Metodología

El enfoque CRISP-DM (Cross-Industry Standard Process for Data Mining) es ampliamente utilizado en el ámbito de la minería de datos y el análisis predictivo (Mirantika, Rijanto, & Estiko., 2023). Aunque no está diseñado específicamente para la segmentación de clientes, su estructura es adaptable a este tipo de análisis, proporcionando una guía efectiva para abordar cada fase del proceso (Mirfakhraei, Abdolvand, & Harandi., 2024). La metodología comienza con una comprensión profunda de los objetivos del negocio y las necesidades específicas relacionadas con la segmentación de clientes. En esta fase inicial, es fundamental identificar y definir claramente el problema que se pretende resolver a través de la segmentación.

El siguiente paso es la comprensión de los datos disponibles. Aquí, se exploran y analizan las fuentes de información para identificar las características relevantes que se utilizarán en el proceso de segmentación (Mirantika, Rijanto, & Estiko., 2023). Este análisis inicial permite determinar qué atributos, como el comportamiento de compra o datos demográficos, son cruciales para agrupar a los clientes de manera efectiva.

En la fase de modelado, se seleccionan las técnicas de segmentación más adecuadas, como el clustering, para formar grupos homogéneos de clientes (Mirfakhraei, Abdolvand, & Harandi., 2024). En este punto, se aplican algoritmos de aprendizaje automático que ayudan a identificar patrones ocultos en los datos, creando los segmentos que serán utilizados en las estrategias de marketing.

Una vez creados los segmentos, es necesario realizar una evaluación del modelo. Se mide la calidad de los grupos generados utilizando métricas como el coeficiente de silueta o la inercia, que permiten evaluar la cohesión y separación de los clústeres formados. Esto garantiza que los segmentos sean útiles y diferenciados.

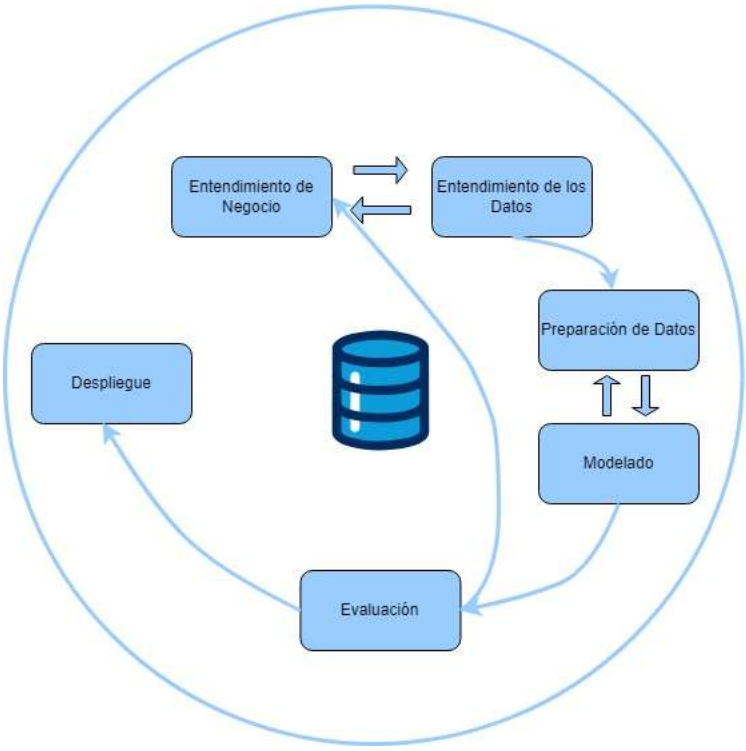
Con el modelo evaluado y validado, se pasa a la implementación. Los segmentos resultantes se aplican directamente en campañas de marketing personalizadas, ajustando las estrategias a las características particulares de cada grupo. Esto permite a las empresas adaptar sus esfuerzos de marketing de manera más precisa, dirigiendo mensajes y ofertas personalizadas a cada segmento de clientes.

El análisis de los resultados es esencial para evaluar el impacto de la segmentación en el rendimiento de la empresa. Los clientes pueden ser segmentados demográficamente, agrupándolos según variables como la edad, género, ingresos y nivel educativo, lo que ayuda a comprender mejor sus necesidades y preferencias. También se pueden utilizar criterios psicográficos, que se centran en las actitudes, intereses y comportamientos de los consumidores, ofreciendo una visión más profunda de sus motivaciones. Además, la segmentación conductual se enfoca en analizar la lealtad a la marca y el uso de productos o servicios en función del tiempo, lo que permite adaptar las estrategias de marketing de acuerdo con los patrones de interacción de los clientes.

El análisis de los resultados obtenidos facilita la evaluación del rendimiento de cada segmento en términos de ingresos y otros indicadores clave. A partir de estos datos, se ajustan las estrategias de marketing para maximizar el impacto de cada grupo. Al mismo tiempo, la información sobre los clústeres puede utilizarse para la personalización del marketing, diseñando mensajes específicos que respondan a las necesidades particulares de cada segmento y optimizando los recursos.

**Figura 1**

*Metodología CRIPS DM*




*Fuente.* Autoría Propia

Por último, este proceso también influye en el desarrollo de productos y servicios. Al comprender las preferencias y comportamientos de cada grupo de clientes, las empresas pueden diseñar productos que se ajusten mejor a sus expectativas. Asimismo, la segmentación permite una optimización de los recursos, ya que conocer en detalle a los clientes facilita la asignación más eficiente de los esfuerzos comerciales, dirigiéndolos hacia los segmentos con mayor potencial.

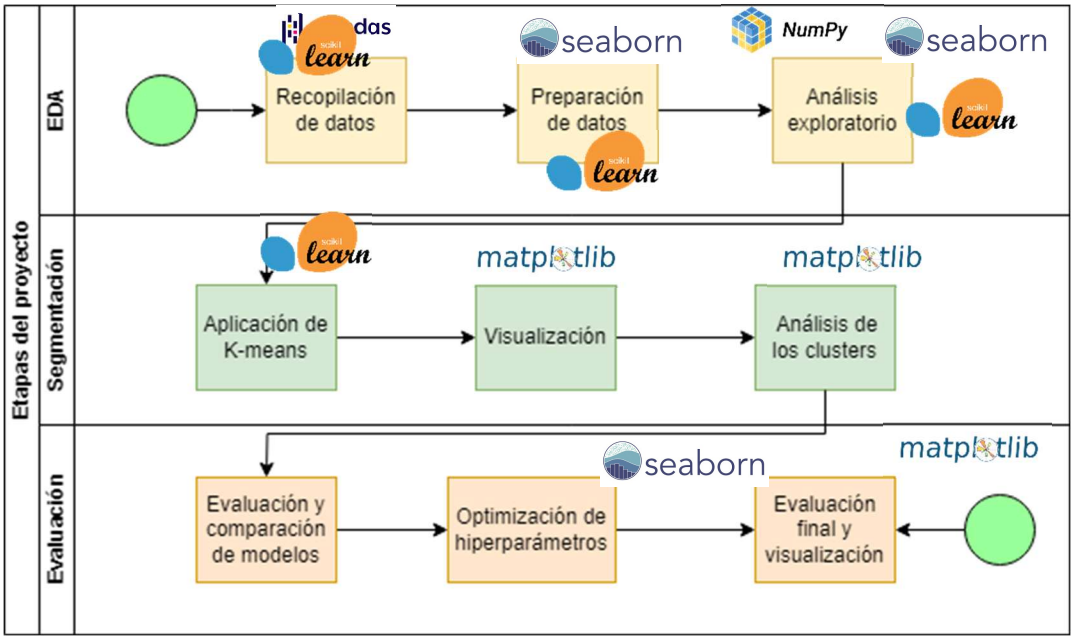
Finalmente, una estrategia bien estructurada puede mejorar la retención y fidelización de los clientes, enfocándose en aquellos grupos con mayor lealtad potencial y diseñando acciones específicas para fortalecer su relación con la marca.

A continuación presentaré un análisis detallado de un conjunto de datos de clientes, con el objetivo de obtener insights relevantes que apoyen las decisiones de marketing y optimización de estrategias comerciales. Se ha seguido un enfoque sistemático que incluye la carga de datos, su limpieza, el análisis exploratorio y la visualización de resultados, junto con recomendaciones basadas en los patrones identificados.

El proceso que has realizado es un flujo completo que involucra tanto algoritmos de aprendizaje no supervisado (como el algoritmo de clustering K-means) para segmentar a los clientes, como algoritmos de aprendizaje supervisado (como Bosque Aleatorio, Impulso de Gradiente, SVM y KNN) para evaluar y mejorar la capacidad de predicción de los clústeres obtenidos. Cada uno cumple un propósito distinto en el análisis de los datos, por lo que te explicaré paso a paso lo que se hizo y cómo encaja todo en el flujo de trabajo.

**Figura 2**

*Proceso de Modelado*




*Fuente.* Autoría Propia

**Preprocesamiento de Datos**

Estandarización de las variables numéricas: Antes de aplicar cualquier algoritmo de clustering, estandarizaste las columnas numéricas de los datos con StandardScaler. Esto es fundamental porque los algoritmos de clustering y muchos otros modelos supervisados son sensibles a las escalas de las variables (Sun, Liu, & Gao., 2023). El objetivo de este paso es asegurarse de que todas las variables tengan la misma escala.

**Segmentación con K-means (Aprendizaje No Supervisado)**

Aplicación del algoritmo K-means: En esta etapa, usaste el algoritmo de clustering K-means, que es un algoritmo de aprendizaje no supervisado. K-means intenta agrupar a los clientes en clústeres o segmentos basados en las similitudes entre las variables (en este caso, variables estandarizadas como ingresos, frecuencia de compra, cantidad de última compra, etc.).

Resultado: Cada cliente se asignó a uno de los 5 clústeres creados por K-means (Rachman, Santoso, & Djajadi., 2021). El objetivo de este paso es encontrar patrones o grupos naturales en los datos sin tener que usar etiquetas predefinidas.

**Reducción de Dimensionalidad y Visualización**

PCA (Análisis de Componentes Principales) y t-SNE: Utilizaste métodos como PCA y t-SNE para reducir la dimensionalidad de los datos y visualizar los clústeres en 2 dimensiones. Estos algoritmos te permiten ver gráficamente cómo los datos están agrupados en función de los clústeres obtenidos en la etapa de K-means (Mirfakhraei, Abdolvand, & Harandi., 2024).

Visualización: La visualización ayuda a interpretar mejor los grupos formados y si realmente tienen una estructura clara o no (Rachman, Santoso, & Djajadi., 2021). Sin embargo, en este punto solo estamos identificando clústeres sin saber qué significan en términos de predicción.

**Evaluación y Análisis de los Clústeres**

Descripción de clústeres: Luego de la segmentación, describiste las características promedio de cada clúster en función de las variables (como ingreso promedio, años de membresía, cantidad de última compra, etc.). Esto te ayudó a entender las diferencias entre los segmentos de clientes.

Insight: Cada clúster tiene su propio perfil. Por ejemplo, un clúster puede tener clientes con un alto gasto promedio pero con una menor frecuencia de compra, mientras que otro puede estar formado por clientes con ingresos medios pero una alta frecuencia de compra (Mirfakhraei, Abdolvand, & Harandi., 2024).

**Introducción de Algoritmos Supervisados**

Ahora viene la parte crucial que genera confusión: la introducción de algoritmos supervisados. Esto se hace por varios motivos

El clustering no supervisado simplemente agrupa a los clientes en función de las similitudes de sus características, pero no puede prever qué grupo pertenecerá un nuevo cliente que no ha sido visto antes. Aquí es donde entran los algoritmos supervisados (Sun, Liu, & Gao., 2023). Una vez que tienes los clústeres creados, puedes entrenar un modelo supervisado para predecir el clúster al que pertenecerá un nuevo cliente.

Proceso: Para lograr esto, conviertes la asignación de clústeres en una tarea de clasificación supervisada. Usas los clústeres generados por K-means como tu variable objetivo (target = df['Cluster']), y luego entrenas modelos supervisados (Bosque Aleatorio, Impulso de Gradiente, etc.) para clasificar nuevos clientes en uno de los clústeres (Rachman, Santoso, & Djajadi., 2021).

**Entrenamiento y Evaluación de Modelos Supervisados**

Entrenamiento: Entrenas cada modelo supervisado (Bosque Aleatorio, Impulso de Gradiente, SVM, KNN) con los datos de características como ingresos, edad, frecuencia de compra, etc., para predecir el clúster al que pertenecen los clientes.

Evaluación: Usas el conjunto de prueba para verificar qué tan bien cada modelo es capaz de clasificar correctamente a los clientes en los clústeres definidos. Generas informes de clasificación y matrices de confusión para ver qué tan precisos son los modelos.

Comparación: La idea es comparar los resultados de todos los modelos y ver cuál tiene mejor precisión, recall, f1-score, etc. Esto te permitirá elegir el mejor modelo para predecir los clústeres en clientes nuevos.

**Optimización del Modelo (Grid Search y selección de características)**

Optimización: En esta etapa, usas técnicas como Grid Search para buscar los mejores hiperparámetros de los modelos y así mejorar aún más su rendimiento (Jpung, & Kim., 2023).

Selección de características: También aplicas técnicas como RFE (Recursive Feature Elimination) para seleccionar las características más relevantes para el modelo de clasificación.

## Resultados:

En este análisis, se utilizó el algoritmo de clustering K-means para segmentar las observaciones del conjunto de datos en cinco grupos. Previamente, las variables numéricas se estandarizaron para garantizar una escala uniforme en el proceso de agrupamiento.

Los centros representan los valores medios de las variables para cada clúster en el espacio estandarizado. Esto permite caracterizar las diferencias entre los grupos formados.

Se determinó la cantidad de observaciones asignadas a cada clúster, indicando la distribución de los datos en los grupos identificados.

La tabla de centros de los clústeres describe las características promedio de las observaciones asignadas a cada clúster en el espacio estandarizado. Cada fila representa un clúster, y cada columna (Variable 1, Variable 2, etc.) indica el valor promedio de esa variable dentro del clúster. Los valores están en una escala estandarizada, lo que significa que tienen una media de 0 y una desviación estándar de 1.

**Tabla 1**

*Centros de los Clusters*

| Clúster | Variable 1 | Variable 2 | Variable 3 | Variable 4 | Variable 5 | Variable 6 |
|---|---|---|---|---|---|---|
| 0 | -0.8375 | 0.4034 | -0.7359 | 0.0546 | 0.1771 | 0.5714 |
| 1 | -0.4592 | 0.2039 | -0.167 | -0.4698 | -0.5588 | -0.7446 |
| 2 | 0.3069 | 0.4234 | 0.4054 | 0.1893 | -0.3457 | -0.6448 |
| 3 | 0.3185 | 0.1657 | -0.4026 | 0.2317 | 0.7323 | 0.2723 |
| 4 | -0.3533 | -0.8848 | 0.7339 | -0.4549 | -0.0359 | 0.5711 |

*Nota.* Esta tabla presenta las centros de los clusters. *Fuente.* Autoría Propia

Valores negativos: Indican que las observaciones dentro del clúster tienen, en promedio, un valor por debajo de la media global (en el conjunto de datos original) para esa variable. Por ejemplo, en el Clúster 0, la Variable 1 tiene un promedio de -0.8375, lo que significa que las observaciones asignadas a este clúster tienen valores significativamente inferiores a la media global en esta variable.

Valores positivos: Indican que las observaciones tienen un promedio por encima de la media global para esa variable. Por ejemplo, en el Clúster 3, la Variable 5 tiene un promedio de 0.7323, lo que sugiere que las observaciones en este grupo tienen valores más altos que el promedio global en esa variable.

Magnitud de los valores: Cuanto mayor es el valor (positivo o negativo), más diferente es el promedio del clúster respecto a la media global. Por ejemplo, en el Clúster 4, la Variable 2 tiene un promedio de -0.8848, lo que indica que las observaciones de este grupo están bastante por debajo de la media global en esa variable.

**Tabla 2**

*Tamaños de los Clusters*

| Clúster | Tamaño |
|---|---|
| 0 | 211 |
| 1 | 206 |
| 2 | 195 |
| 3 | 192 |
| 4 | 204 |

*Nota.* Esta tabla presenta las tamaños de los clusters. *Fuente.* Autoría Propia

Los resultados muestran que los datos están distribuidos de manera relativamente uniforme entre los cinco clústeres, con tamaños de clúster que varían ligeramente entre 192 y

211 observaciones. Esto sugiere que el algoritmo identificó grupos de características similares en las observaciones analizadas.

**Recomendaciones de Ventas Cruzadas y Adicionales**

Se realizó un análisis de las categorías preferidas por los clientes dentro de cada clúster para identificar oportunidades de ventas cruzadas. Esto permite entender las preferencias por grupo y diseñar estrategias personalizadas.

**Tabla 3**

*Recomendaciones de Ventas Cruzadas*

| Categoría Preferida | Clúster 0 | Clúster 1 | Clúster 2 | Clúster 3 | Clúster 4 |
|---|---|---|---|---|---|
| Clothing | 33 | 38 | 33 | 42 | 43 |
| Electronics | 39 | 49 | 40 | 41 | 45 |
| Groceries | 34 | 47 | 41 | 43 | 41 |
| Home & Garden | 33 | 41 | 41 | 43 | 40 |
| Sports | 39 | 38 | 43 | 34 | 39 |

*Nota.* Esta tabla presenta las recomendacions de ventas cruzadas. *Fuente.* Autoría Propia

La tabla de recomendaciones de ventas cruzadas y adicionales analiza las preferencias de categorías de productos por clúster, proporcionando información clave para diseñar estrategias de marketing personalizadas. Cada clúster representa un grupo de clientes con características similares, permitiendo identificar cuáles categorías tienen mayor demanda en cada segmento.

El Clúster 1 destaca por su fuerte preferencia hacia la categoría "Electronics", con 49 clientes, lo que lo convierte en un objetivo ideal para campañas relacionadas con productos

tecnológicos. Este segmento podría responder favorablemente a estrategias de ventas cruzadas que incluyan dispositivos complementarios o servicios asociados, como garantías extendidas o accesorios.

En el Clúster 4, las categorías "Clothing" y "Electronics" muestran una ligera predominancia, con 43 y 45 clientes, respectivamente. Esto sugiere que los clientes de este grupo tienen intereses diversificados, pero con una inclinación hacia moda y tecnología. Este clúster podría beneficiarse de promociones que combinen ambas categorías, como descuentos por la compra de ropa junto con dispositivos portátiles.

Los Clústeres 0, 2 y 3 presentan una distribución más equilibrada en las categorías, lo que sugiere preferencias menos marcadas. Por ejemplo, el Clúster 2 tiene 41 clientes en "Groceries" y "Home & Garden", lo que indica que las estrategias de ventas cruzadas podrían enfocarse en la promoción de productos relacionados con el hogar. Asimismo, el Clúster 3 tiene 43 clientes en "Sports", lo que lo posiciona como un grupo relevante para artículos deportivos y servicios relacionados, como membresías de gimnasios.

En general, los resultados reflejan que las preferencias varían significativamente entre los clústeres, lo que refuerza la importancia de diseñar estrategias personalizadas para cada segmento. Estas recomendaciones permiten no solo satisfacer mejor las necesidades de los clientes, sino también incrementar las oportunidades de ingresos mediante ventas adicionales y cruzadas bien dirigidas.

**Eficiencia de Gasto**

En el análisis, se calculó la eficiencia de gasto (spending efficiency) como la relación entre el puntaje de gasto (spending score) y los ingresos del cliente (income). Este indicador proporciona información sobre qué tan eficiente es un cliente al convertir sus ingresos en puntuaciones de

gasto, lo que puede ser útil para segmentar clientes y diseñar estrategias de marketing o gestión financiera.

**Tabla 4**

*Eficiencia de Gasto*

| Puntaje de Gasto | Ingresos | Eficiencia de Gasto |
|---|---|---|
| 90 | 99342 | 0.000906 |
| 60 | 78852 | 0.000761 |
| 30 | 126573 | 0.000237 |
| 74 | 47099 | 0.001571 |
| 21 | 140621 | 0.000149 |

*Nota.* Esta tabla presenta las eficiencia de gasto. *Fuente.* Autoría Propia

La tabla presenta tres columnas principales: el Puntaje de Gasto, los Ingresos y la Eficiencia de Gasto de cinco clientes. El Puntaje de Gasto representa una métrica asociada al consumo del cliente, mientras que los Ingresos reflejan su capacidad económica. La Eficiencia de Gasto es una proporción calculada dividiendo el puntaje de gasto entre los ingresos, lo que permite evaluar qué tan eficientemente un cliente utiliza sus recursos para consumir.

En los resultados, se observa que los clientes con ingresos más bajos, como aquel con un ingreso de 47,099 y un puntaje de gasto de 74, tienen una eficiencia de gasto más alta (0.001571), indicando que destinan una mayor proporción de sus ingresos al consumo. Por otro lado, los clientes con ingresos más altos, como aquel con un ingreso de 140,621 y un puntaje de gasto de 21, muestran una eficiencia de gasto menor (0.000149), sugiriendo que, a pesar de contar con mayores recursos, consumen proporcionalmente menos en relación con sus ingresos.

Estos patrones permiten identificar tendencias significativas: los clientes con ingresos más limitados tienden a consumir de manera más intensa, mientras que aquellos con ingresos más elevados presentan un comportamiento más moderado en términos de eficiencia de gasto. Este análisis ofrece una base sólida para segmentar a los clientes y diseñar estrategias personalizadas, enfocadas en aumentar la eficiencia de gasto de los clientes con ingresos altos o fortalecer la relación con los clientes de ingresos bajos a través de ofertas que maximicen su valor percibido.

Los valores obtenidos reflejan diferencias significativas en la eficiencia de gasto entre los clientes. Por ejemplo, en los primeros cinco casos analizados.

Un cliente con un ingreso de 99,342 y un puntaje de gasto de 90 tiene una eficiencia de gasto de 0.000906, lo que indica que este cliente convierte menos del 0.1% de sus ingresos en puntajes de gasto.

Otro cliente, con ingresos más bajos (47,099) pero un puntaje de gasto de 74, tiene una eficiencia de gasto más alta de 0.001571. Esto sugiere que los clientes con ingresos más bajos tienden a tener una eficiencia de gasto mayor, ya que destinan proporcionalmente más de sus ingresos al consumo medido en el puntaje.

**Implicaciones**

Este análisis de eficiencia de gasto puede ayudar a identificar segmentos de clientes según su comportamiento financiero.

Clientes con alta eficiencia de gasto: Son más propensos a destinar una mayor proporción de sus ingresos al consumo. Pueden ser un buen objetivo para campañas de fidelización o promociones personalizadas.

Clientes con baja eficiencia de gasto: Representan una oportunidad para incentivar el gasto mediante estrategias de marketing que subrayen el valor percibido o los beneficios de productos y servicios específicos.

Estos resultados pueden ser utilizados para mejorar la segmentación y optimizar estrategias de ventas o servicios financieros dirigidos a los distintos perfiles de clientes.

**Modelos Supervisados**

En este análisis, se evaluó el desempeño de cuatro modelos de clasificación supervisada: Bosque Aleatorio, Impulso de Gradiente, Support Vector Machine (SVM) y K-Nearest Neighbors (KNN). Estos modelos se aplicaron a un conjunto de datos multiclase, y su rendimiento se midió utilizando métricas como precisión, recall, F1-score y accuracy global. Adicionalmente, generamos informes de clasificación y matrices de confusión para identificar la capacidad de cada modelo en la predicción de las diferentes clases. Este enfoque nos permitió comparar la eficacia de los modelos, identificar áreas de mejora y seleccionar el más adecuado para el problema planteado, considerando tanto el balance general entre las métricas como el comportamiento en clases específicas.

**Tabla 5**

*Métricas Bosque Aleatorio*

| Clase | Precisión | Sensibilidad (Recall) | Puntaje F1 | Soporte |
|---|---|---|---|---|
| 0 | 0.77 | 0.75 | 0.76 | 45 |
| 1 | 0.72 | 0.72 | 0.72 | 43 |
| 2 | 0.88 | 0.89 | 0.88 | 75 |
| 3 | 0.75 | 0.83 | 0.79 | 37 |
| Promedio | 0.8 | 0.8 | 0.8 | 200 |

*Nota.* Esta tabla presenta las métricas del modelo Bosque aleatorio. *Fuente.* Autoría Propia

El modelo Bosque aleatorio muestra un rendimiento equilibrado con una precisión promedio ponderada del 80%. Destaca en la clase 2 con un F1-score de 0.88, lo que indica que es muy efectivo para predecir esta clase. Sin embargo, el desempeño en las clases 0 y 1 es algo inferior, lo que sugiere posibles oportunidades de mejora en estas categorías.

**Tabla 6**

*Métricas Impulso de Gradiente*

| Clase | Precisión | Sensibilidad (Recall) | Puntaje F1 | Soporte |
|---|---|---|---|---|
| 0 | 0.73 | 0.71 | 0.72 | 45 |
| 1 | 0.72 | 0.72 | 0.72 | 43 |
| 2 | 0.86 | 0.89 | 0.87 | 75 |
| 3 | 0.8 | 0.73 | 0.76 | 37 |
| Promedio ponderado | 0.8 | 0.8 | 0.8 | 200 |

*Nota.* Esta tabla presenta las métricas del modelo Impulso de Gradiente. *Fuente.* Autoría Propia

El modelo Impulso de Gradiente tiene un rendimiento similar al Bosque aleatorio, con un promedio ponderado del 80%. Similar al Bosque Aleatorio, el mejor desempeño está en la clase 2, mientras que las clases 0 y 1 son menos precisas. La clase 3 también muestra un recall más bajo, lo que implica que algunas instancias de esta clase no están siendo correctamente identificadas.

**Tabla 7**

*Métricas SVM*

| Clase | Precisión | Sensibilidad (Recall) | Puntaje F1 | Soporte |
|---|---|---|---|---|
| 0 | 0.42 | 0.78 | 0.55 | 45 |
| 1 | 0.62 | 0.7 | 0.66 | 43 |
| 2 | 0.83 | 0.86 | 0.85 | 75 |
| 3 | 0.33 | 0.08 | 0.13 | 37 |
| Promedio ponderado | 0.63 | 0.65 | 0.63 | 200 |

*Nota.* Esta tabla presenta las métricas del modelo SVM. *Fuente.* Autoría Propia

El modelo SVM tiene el rendimiento más bajo entre los cuatro modelos evaluados, con un promedio ponderado del 63%. Aunque el modelo logra un buen desempeño en la clase 2, tiene dificultades significativas en la clase 3, con un F1-score de solo 0.13, lo que indica un bajo reconocimiento de instancias de esta clase.

**Tabla 8**

*Métricas KNN*

| Clase | Precisión | Sensibilidad (Recall) | Puntaje F1 | Soporte |
|---|---|---|---|---|
| 0 | 0.56 | 0.49 | 0.52 | 45 |
| 1 | 0.44 | 0.44 | 0.44 | 43 |
| 2 | 0.8 | 0.84 | 0.82 | 75 |
| 3 | 0.4 | 0.38 | 0.39 | 37 |
| Promedio ponderado | 0.6 | 0.6 | 0.6 | 200 |

*Nota.* Esta tabla presenta las métricas del modelo KNN. *Fuente.* Autoría Propia

El modelo KNN presenta un rendimiento general bajo, con un promedio ponderado del 60%. Aunque tiene un desempeño decente en la clase 2, las clases 0, 1 y 3 muestran métricas significativamente más bajas, especialmente en la precisión y el recall.

En general, el modelo Bosque aleatorio y el de Impulso de Gradiente se destacan como los más equilibrados y eficaces, con un promedio ponderado del 80% en todas las métricas.

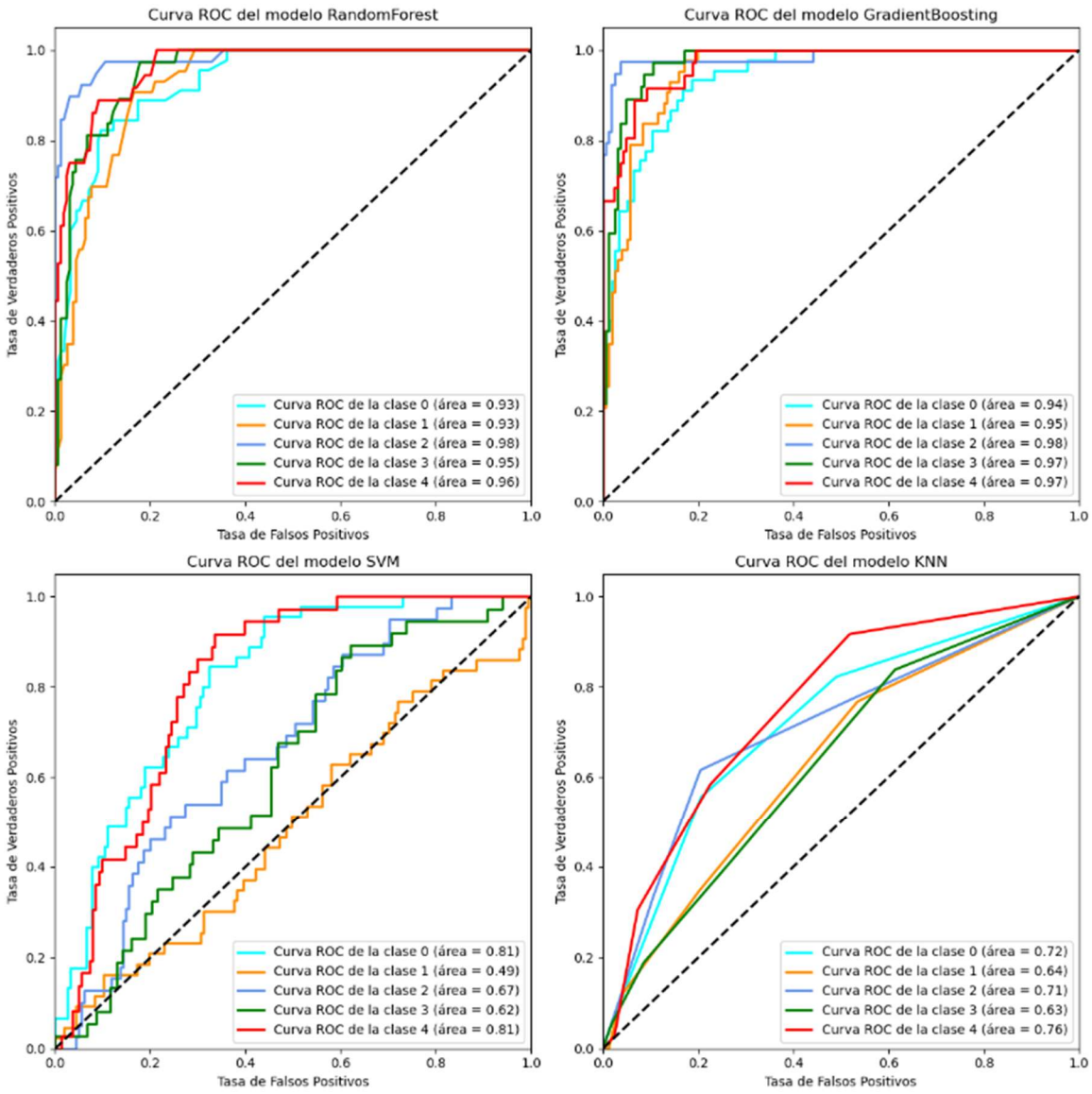
Por otro lado, SVM y KNN tienen un desempeño inferior, con promedios ponderados del 63% y 60%, respectivamente. Estos modelos tienen dificultades significativas en la identificación de ciertas clases, como la clase 3 en ambos casos.

Con base en estos resultados, Bosque Aleatorio y Impulso de Gradiente son los modelos más recomendados, especialmente si el enfoque está en obtener un balance entre precisión y recall en todas las clases. Sin embargo, se ajustaran los hiperparámetros para mejorar el rendimiento en las categorias.

Con el fin de seleccionar el mejor modelo evaluaremos las métricas ROC (Curva Característica Operativa del Receptor) y AUC (Área Bajo la Curva) de los modelos previamente analizados.

**Figura 3**

*Métricas de los Modelos Supervisados Implementados*




*Fuente.* Autoría Propia.

Este gráfico muestra las curvas ROC (Curva Característica Operativa del Receptor) para cuatro modelos de clasificación diferentes: Bosque Aleatorio, Impulso de Gradiente, SVM y KNN. A continuación, te explico cómo interpretar los resultados para cada uno de los modelos:

El modelo de Bosque Aleatorio ha demostrado un rendimiento sobresaliente, con valores de AUC entre 0.93 y 0.98 para las distintas clases. Sus curvas ROC se acercan significativamente a la esquina superior izquierda, lo que indica una gran capacidad para discriminar entre clases. Esto sugiere que el modelo puede realizar predicciones precisas con un margen de error bajo, siendo una opción altamente confiable en problemas de clasificación multiclase.

El Impulso de Gradiente también muestra un desempeño excelente, con AUC entre 0.94 y 0.97. Aunque sus métricas son similares a las de Bosque Aleatorio, presenta una ligera mejora en algunas clases, en particular en la clase 4. La estabilidad y precisión de este modelo lo convierten en una opción ideal para la tarea de clasificación, destacándose por su capacidad de ajuste y manejo eficiente de patrones complejos en los datos.

El modelo SVM (Support Vector Machine) muestra una capacidad de discriminación variable, con AUC que van desde 0.49 en la clase 1 hasta 0.81 en la clase 0. Las curvas ROC no están tan cerca de la esquina superior izquierda en comparación con los modelos anteriores, lo que indica un rendimiento inconsistente. Esto sugiere que SVM tiene dificultades para separar correctamente ciertas clases, lo que lo convierte en una opción menos robusta para este problema.

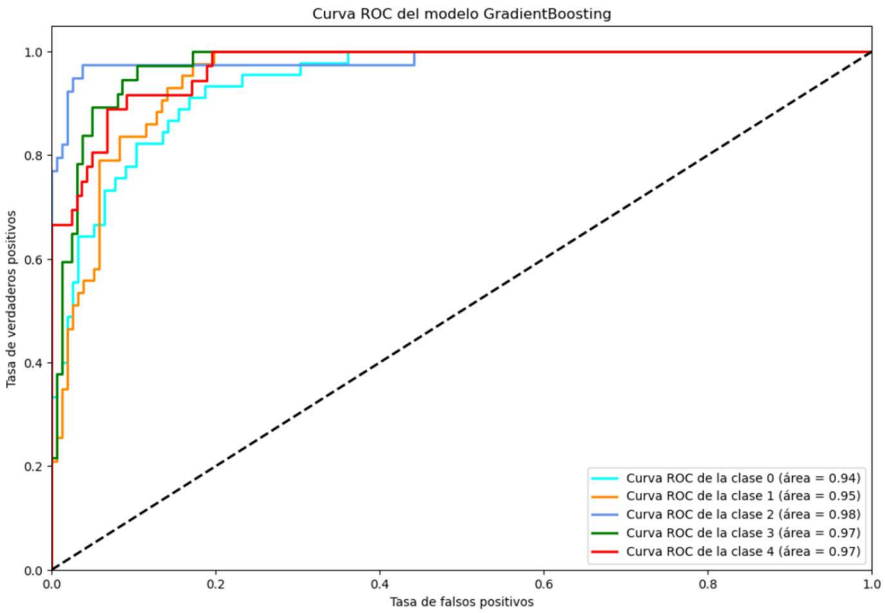
Por otro lado, el modelo KNN (K-Nearest Neighbors) presenta el peor desempeño, con AUC entre 0.64 y 0.72. Sus curvas ROC están más cercanas a la diagonal de no discriminación, lo que indica que el modelo no logra separar adecuadamente las clases. Debido a su limitada

capacidad de discriminación, KNN no es la mejor opción en este caso, especialmente cuando se requiere una clasificación precisa.

En conclusión, Impulso de Gradiente es el modelo más robusto, con Bosque Aleatorio como una alternativa igualmente efectiva. Ambos ofrecen una gran capacidad de clasificación y alta precisión. SVM tiene un rendimiento intermedio y su efectividad depende de la clase específica, mientras que KNN es el modelo con el peor desempeño, lo que lo hace poco adecuado para este problema.

**Figura 4**

*Curva ROC modelo Gradiente Boosting*




*Fuente.* Autoría Propia.

Como podemos ver en la figura 4, gráfica de ROC correspondiente al modelo de Impulos de Gradiente sin optimización, se destaca un excelente rendimiento en términos del área bajo la curva (AUC). Las clases 2, 3 y 4 presentan los valores más altos, con AUC de 0.98, 0.97 y 0.97 respectivamente, indicando un desempeño notable en la clasificación de estas categorías. Las

clases 0 y 1 también muestran un rendimiento sólido, con AUC de 0.94 y 0.95. Estos resultados reflejan una alta capacidad del modelo para discriminar entre las diferentes clases, como se puede observar en la gráfica presentada.

El modelo sin optimización ya muestra un rendimiento excelente. Las AUC para todas las clases son bastante altas, lo que indica una gran capacidad de discriminación. El modelo tiene un buen balance entre la detección de verdaderos positivos y la minimización de falsos positivos en cada clase.

**Optimización Modelo Seleccionado:**

La optimización de hiperparámetros es un paso fundamental en el desarrollo de modelos de machine learning, ya que permite ajustar configuraciones clave para maximizar su desempeño. En este análisis, utilizamos esta técnica para mejorar el modelo de Impulso de Gradiente, el cual mostró las mejores métricas generales entre las opciones evaluadas. Este modelo combina múltiples árboles de decisión, ajustándolos de manera iterativa para minimizar el error y mejorar su capacidad predictiva. La optimización de sus hiperparámetros, como el número de estimadores, la profundidad de los árboles y la tasa de aprendizaje, permitió encontrar un balance óptimo entre sesgo y varianza, logrando un rendimiento robusto en todas las clases del problema. Este enfoque demuestra la importancia de ajustar cuidadosamente los parámetros del modelo para obtener resultados más precisos y confiables.

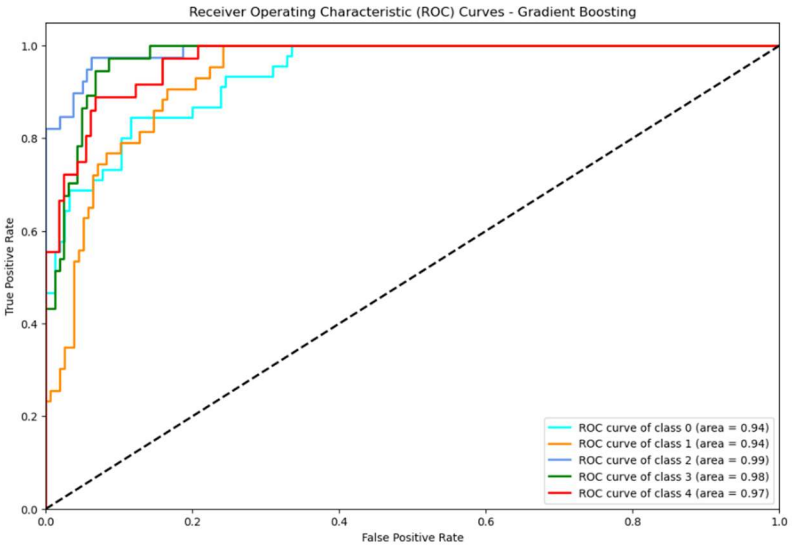
**Análisis de Impulso de Gradiente Optimizado (con Grid Search):**

El análisis del modelo de Impulso de Gradiente optimizado mediante Grid Search, representado en la segunda gráfica de ROC, muestra un desempeño notable en términos de AUC. La Clase 2 alcanza el valor más alto con una AUC de 0.99, seguida por la Clase 3 con 0.98 y la Clase 4 con 0.97, evidenciando una mejora en la capacidad del modelo para clasificar

correctamente estas categorías. Por su parte, las Clases 0 y 1 obtienen una AUC de 0.94 cada una, manteniendo un rendimiento sólido pero sin cambios significativos respecto al modelo sin optimización. En conjunto, los resultados reflejan una mejora general en la discriminación de las clases, especialmente en las categorías con AUC más altas, como se puede apreciar en la gráfica presentada.

**Figura 5**

*Curva ROC Modelo Gradiente Boosting Optimizado*




*Fuente.* Autoría Propia

Después de la optimización con Grid Search, las AUC no cambian significativamente en algunas clases, pero se observa una mejora leve en la clase 2, que alcanza un AUC de 0.99, lo cual es excepcional. Esto significa que el modelo optimizado tiene una mejor capacidad para discriminar correctamente entre las clases, especialmente en las clases con mayor complejidad. M[etricas del modelo Impulso de graadiente optimizado:

**Tabla 9**

*Métricas Impulso de Gradiente Optimizado*

| Clase | Precisión | Sensibilidad (Recall) | Puntaje F1 | Soporte |
|---|---|---|---|---|
| 0 | 0.8333 | 0.6667 | 0.7407 | 45 |
| 1 | 0.7111 | 0.7442 | 0.7273 | 43 |
| 2 | 0.9167 | 0.8462 | 0.88 | 39 |
| 3 | 0.7674 | 0.8919 | 0.825 | 37 |
| 4 | 0.75 | 0.8333 | 0.7895 | 36 |

*Nota.* Esta tabla presenta las métricas del modelo. *Fuente.* Autoría Propia

El modelo optimizado de Impulso de Gradiente presenta métricas detalladas para cada clase, lo que permite analizar su desempeño en términos de precisión, sensibilidad (recall) y puntaje F1. Los resultados muestran que el modelo tiene su mejor desempeño en la clase 2, con una precisión de 0.9167, una sensibilidad de 0.8462 y un puntaje F1 de 0.88. Este rendimiento destaca significativamente frente a las demás clases, indicando que el modelo predice con alta eficacia las observaciones de esta categoría. Por otro lado, el desempeño más bajo se encuentra en la clase 0, con un puntaje F1 de 0.7407, debido a una sensibilidad más reducida (0.6667).

En comparación con las métricas generales del modelo de Impulso de Gradiente sin optimización (precisión promedio de 0.7957, sensibilidad de 0.7964 y puntaje F1 de 0.7925), se observa que el modelo optimizado mejora el equilibrio entre las clases, logrando métricas más altas en clases específicas como la clase 3 (F1: 0.8250) y la clase 4 (F1: 0.7895). Además, la precisión generalizada entre las clases optimizadas resulta más consistente que en el modelo sin optimización.

Estos resultados reflejan que la optimización de hiperparámetros tuvo un impacto positivo en la capacidad del modelo para manejar mejor las diferencias entre las clases, reforzando su utilidad en escenarios donde es crucial una alta precisión en múltiples categorías. Esto hace del modelo optimizado una opción superior en comparación con su versión inicial.

**Comparación General:**

En términos de AUC (Área Bajo la Curva ROC), tanto el modelo estándar como el optimizado presentan un excelente rendimiento. Sin embargo, el modelo optimizado muestra una ligera mejora en algunas clases, lo que indica que la optimización ha permitido un ajuste más preciso del modelo. Esto sugiere que el modelo optimizado tiene una mejor capacidad para diferenciar entre clases, lo cual es crucial en problemas donde la clasificación precisa es fundamental.

Respecto a precisión, recall y F1-score, el modelo optimizado también presenta una leve mejora en comparación con el estándar. Estas métricas reflejan la calidad de las predicciones, asegurando que el modelo no solo maximiza la cantidad de aciertos, sino que también mantiene un equilibrio entre la detección de clases positivas y la reducción de falsos positivos. Esto indica que el modelo optimizado es más confiable en términos de predicción general y minimización de errores en la clasificación.

En cuanto a recomendaciones, aunque el modelo estándar de Impulso de Gradiente ya ofrece un rendimiento excelente, la optimización con Grid Search permite obtener un mejor ajuste para clases específicas. Esto puede ser particularmente valioso en aplicaciones donde la correcta discriminación entre ciertas clases es crítica. Sin embargo, si el costo computacional es un factor importante, el modelo sin optimización podría ser suficiente para muchas aplicaciones.

Si el objetivo es obtener el mejor desempeño posible, la optimización con Grid Search es la mejor opción, incluso si las mejoras son marginales.

La optimización de campañas de marketing mediante un enfoque mixto de modelos no supervisados, como el clustering, y modelos supervisados puede mejorar significativamente la toma de decisiones estratégicas. Los modelos de clustering permiten segmentar clientes en grupos con patrones de comportamiento similares, mientras que los modelos supervisados pueden predecir la probabilidad de respuesta a una campaña específica. Al combinar ambos enfoques, es posible diseñar estrategias más personalizadas y eficientes, optimizando los recursos y maximizando la efectividad de las campañas.

**Segmentación del Mercado con Clustering no Supervisado (K-means)**

El primer paso en la optimización de campañas de marketing es la segmentación del mercado, utilizando un algoritmo no supervisado como K-means. Este método permite agrupar a los clientes según características similares, como patrones de compra, comportamiento de gasto, frecuencia de compra y variables demográficas. Al identificar estos segmentos, las estrategias de marketing pueden ser más precisas y efectivas, maximizando el impacto de las campañas.

Para garantizar que el algoritmo K-means funcione correctamente, es fundamental realizar la estandarización de las características numéricas. Dado que K-means es sensible a la escala de los datos, la normalización asegura que todas las variables contribuyan de manera equitativa en la formación de los clusters, evitando que variables con rangos más grandes dominen la segmentación.

Una vez estandarizados los datos, se aplica el algoritmo K-means para dividir a los clientes en varios clusters. En este caso, se utilizan cinco clusters, lo que permite identificar diferentes perfiles de clientes con base en sus hábitos y comportamientos. La correcta elección

del número de clusters es clave para garantizar una segmentación representativa y útil en la toma de decisiones estratégicas.

Finalmente, se analizan los centros de los clusters para entender las diferencias clave entre los grupos de clientes. Aspectos como nivel de ingresos, patrones de compra y frecuencia de gasto pueden proporcionar información valiosa sobre qué estrategias de marketing serán más efectivas para cada segmento, permitiendo una personalización más precisa de las campañas.

**Insights Clave Obtenidos del Clustering:**

Segmentación Prioritaria por Precisión y Sensibilidad Alta, la clase 2 es el segmento más consistente y bien identificado por el modelo, con una precisión de 91.67% y un puntaje F1 de 88.00%. Esto sugiere que los clientes en esta categoría tienen características claras que el modelo reconoce fácilmente. Por lo tanto, se pueden diseñar campañas dirigidas exclusivamente a este grupo, optimizando la personalización de productos y servicios.

Estrategias de Mejora para Clases con Baja Sensibilidad, la clase 0 tiene el puntaje F1 más bajo (74.07%) debido a su sensibilidad limitada (66.67%), lo que indica que algunos clientes de esta clase no están siendo reconocidos adecuadamente. Esto sugiere la necesidad de mejorar la comprensión del comportamiento de este segmento mediante análisis adicionales, como encuestas o estudios cualitativos, para identificar patrones subrepresentados y ajustar estrategias de captación.

Campañas Equilibradas para Clases con Alto Recall, la clase 3 destaca con una alta sensibilidad (89.19%) y un puntaje F1 de 82.50%. Esto indica que el modelo identifica con gran precisión a los clientes en esta categoría. Una estrategia efectiva podría ser la implementación de campañas de retención que maximicen la lealtad del cliente en este segmento, como descuentos personalizados o programas de fidelización.

Potencial de Crecimiento en la Clase 1, aunque la clase 1 tiene métricas moderadas (F1 de 72.73%), presenta una sensibilidad aceptable (74.42%). Este segmento podría beneficiarse de campañas que aumenten su compromiso o actividad, como incentivos para compras recurrentes o promociones por referidos.

Enfoque Diversificado para la Clase 4, la clase 4 muestra un desempeño balanceado (F1 de 78.95%), lo que sugiere que es un segmento confiable para campañas amplias que combinen adquisición y retención. Se pueden implementar estrategias como lanzamientos de nuevos productos o servicios que capten su atención y mantengan su interés.

Personalización Basada en Métricas por Clase, el uso de las métricas detalladas por clase permite crear campañas personalizadas que maximicen la efectividad en cada segmento. Por ejemplo, se pueden dirigir recursos hacia las clases con mayor precisión y recall, mientras se diseñan estrategias específicas para mejorar la identificación y conversión de las clases con menor desempeño.

En general, estos insights permiten priorizar esfuerzos de marketing, optimizar la asignación de recursos y diseñar estrategias diferenciadas basadas en el comportamiento de los clientes, mejorando tanto la captación como la retención en diferentes segmentos del mercado.

**Conclusiones del Análisis por Clúster**

Clase 0, los clientes de este clúster tienen una edad promedio de 49.84 años, con ingresos promedio de 63,322. Su puntaje de gasto es 52.26, indicando un comportamiento de consumo moderado en relación con sus ingresos. La frecuencia máxima de compra (50) sugiere una actividad constante. Este grupo puede beneficiarse de estrategias de engagement como promociones recurrentes o programas de fidelización para aumentar su gasto promedio.

Clase 1, la edad promedio de los clientes en este clúster es 46.85 años, y sus ingresos son 82,787, superiores a la media general. El puntaje de gasto promedio es 49.32, indicando un comportamiento de consumo conservador. Sin embargo, la frecuencia máxima de compra de 50 sugiere un interés en las compras recurrentes. Las estrategias podrían enfocarse en potenciar la experiencia del cliente para aumentar el gasto promedio.

Clase 2, este clúster destaca por tener los ingresos más altos, con un promedio de 110,416. Sin embargo, su puntaje de gasto promedio (56.16) no es el más bajo, aunque sugiere que los clientes priorizan compras selectivas. El monto promedio de la última compra (230.19) indica una preferencia por adquisiciones menos costosas. Campañas que destaquen el valor de productos premium y la personalización pueden incentivar un mayor gasto en este segmento.

Clase 3, los clientes de este grupo tienen una edad promedio de 41.29 años y un ingreso promedio de 74,600. Aunque su puntaje de gasto (57.39) y eficiencia de gasto (0.000889) son altos en comparación con otros clústeres, el monto promedio de sus últimas compras (353.80) es moderado. Este grupo podría responder bien a estrategias de retención y programas de recompensas exclusivas para fortalecer la lealtad a largo plazo.

Clase 4, este clúster está compuesto por clientes con la edad promedio más baja (30.48 años) pero con ingresos promedio relativamente altos (113,613). Su puntaje de gasto es el más bajo (37.51), reflejando un comportamiento conservador. La frecuencia máxima de compra (50) sugiere que participan activamente en el mercado, pero con compras de bajo valor. Estrategias enfocadas en descuentos y promociones accesibles pueden ser efectivas para capturar este segmento.

**Optimización Basada en Clustering:**

Campañas personalizadas: Cada cluster puede ser el objetivo de una campaña personalizada. Por ejemplo, los clientes de alto ingreso con baja frecuencia de compra pueden ser estimulados con promociones especiales que incentiven compras más frecuentes.

Estrategia de ventas cruzadas: A través de los análisis de preferencia de categorías (categorías favoritas de cada cluster), podemos generar recomendaciones personalizadas de productos para ventas cruzadas (por ejemplo, si un grupo compra mayoritariamente electrónicos, se les pueden recomendar accesorios relacionados).

**Predicción y Clasificación con Modelos Supervisados**

Una vez que los clientes han sido segmentados, los modelos supervisados pueden ayudar a predecir el comportamiento de los nuevos clientes o clasificar a los clientes dentro de los clusters previamente definidos. Aquí es donde entran en juego algoritmos como Bosque Aleatorio, Impulso de Gradiente, SVM, y KNN.

Evaluación de la efectividad, métricas de clasificación (precisión, recall, F1-score, etc.) proporcionan una evaluación sobre qué tan bien el modelo supervisado está prediciendo a los clientes dentro de los clusters correctos. Por ejemplo, un buen valor de precisión significa que las recomendaciones de productos están siendo dirigidas al público correcto.

Curvas ROC y AUC: Estas métricas ayudan a determinar la capacidad del modelo para distinguir entre clientes de diferentes clusters o para prever qué clientes responderán mejor a las campañas.

Para optimizar las campañas de marketing, es fundamental implementar estrategias de optimización continua que permitan ajustar de forma dinámica los modelos utilizados. En este sentido, se recomienda emplear técnicas como Grid Search o Random Search para identificar los

hiperparámetros que maximicen el rendimiento de los modelos supervisados. Asimismo, es crucial realizar un análisis de importancia de características, lo que ayudará a determinar cuáles variables, como la frecuencia de compra o el nivel de ingresos, son más relevantes para clasificar correctamente a los clientes.

Por otro lado, la integración de datos adicionales resulta vital para enriquecer el análisis y mejorar la precisión de los resultados. Incorporar información de comportamiento digital, como las visitas al sitio web o las interacciones en redes sociales, puede afinar tanto la creación de clusters como las predicciones. Además, considerar variables temporales, por ejemplo, la frecuencia de compra a lo largo del tiempo, permite identificar patrones estacionales y ajustar las campañas en función de dichos ciclos.

En conclusión, el enfoque combinado de segmentar clientes mediante K-means clustering y clasificar nuevos clientes con modelos supervisados ofrece una estrategia robusta para optimizar las campañas de marketing. El uso del clustering facilita la detección de patrones ocultos en los datos, permitiendo el diseño de campañas personalizadas y más efectivas, mientras que los modelos supervisados ayudan a predecir el comportamiento futuro, maximizando la conversión y la retención. La optimización continua de estos procesos garantiza campañas cada vez más precisas, eficaces y adaptadas a las necesidades específicas de cada grupo de clientes, lo que se traduce en un mayor retorno sobre la inversión (ROI) y una mejora en la satisfacción del cliente.

## Conclusiones

El proceso de segmentación mediante Machine Learning, utilizando algoritmos de clustering como K-meas, permitió no solo identificar grupos de clientes basados en características clave como ingresos, frecuencia de compra y preferencias de productos, sino también adquirir aprendizajes valiosos sobre cómo optimizar estas técnicas para obtener resultados más efectivos. Durante el análisis, se evidenció la importancia de seleccionar características significativas que reflejen patrones reales de comportamiento, lo que garantizó que los clusters generados fueran útiles para personalizar campañas de marketing y tomar decisiones estratégicas.

Un aspecto clave aprendido fue cómo interpretar los resultados del clustering para traducirlos en insights accionables. Por ejemplo, identificar segmentos con alta lealtad pero menores ingresos ayudó a diseñar estrategias para fortalecer su relación con la marca, mientras que los grupos de ingresos altos y baja frecuencia de compra representaron oportunidades para aumentar la recurrencia. Este enfoque mejoró significativamente la precisión en la identificación de patrones de compra, demostrando que una segmentación adecuada puede superar los problemas asociados con metodologías tradicionales.

En general, el proceso reafirmó que aplicar Machine Learning en la segmentación no solo facilita la personalización de campañas, sino que también optimiza el retorno de inversión al identificar grupos con características específicas y necesidades particulares.

La evaluación de modelos supervisados como Impulso de Gradiente y Bosque Aleatorio ha demostrado ser una solución altamente eficaz para la clasificación de nuevos clientes en los clústeres definidos previamente mediante técnicas de clustering. Ambos modelos exhiben un desempeño sobresaliente, evidenciado por métricas clave como precisión, recall y AUC, lo que

confirma su robustez en términos de capacidad predictiva. Esta capacidad permite clasificar con exactitud a los clientes en segmentos específicos, optimizando la adaptabilidad de las estrategias comerciales.

El uso de estos modelos no solo facilita la incorporación eficiente de nuevos clientes, sino que también aporta flexibilidad al proceso de segmentación. La clasificación dinámica de nuevos datos permite que los clústeres se mantengan relevantes y representativos de los cambios en las características del mercado. Esto se traduce en la posibilidad de ajustar campañas de marketing y estrategias de ventas en tiempo real, utilizando información actualizada para maximizar el impacto.

Además, el enfoque supervisado permite reducir la dependencia de análisis manual, agilizando la toma de decisiones al automatizar la asignación de clientes a los segmentos adecuados. Por ejemplo, un cliente nuevo puede ser clasificado en un clúster de alta frecuencia de compra y bajo ingreso, lo que activa inmediatamente estrategias específicas como descuentos personalizados o promociones exclusivas, asegurando una respuesta rápida a sus necesidades.

En resumen, la integración de modelos supervisados como Impulso de Gradiente y Bosque Aleatorio refuerza la capacidad de segmentación dinámica, garantizando una mayor precisión y eficiencia en la adaptación de campañas de marketing. Este enfoque aprovecha al máximo los datos disponibles, asegurando que las estrategias comerciales estén alineadas con las características y comportamientos de los clientes en tiempo real.

Impacto en la personalización de estrategias de marketing La capacidad de adaptar mensajes y ofertas según las características de cada clúster ha optimizado la eficiencia de las campañas de marketing. Este enfoque no solo mejora la relevancia de las campañas, sino que también impacta positivamente en la retención y lealtad de los clientes, ya que estos reciben contenido alineado a sus necesidades y preferencias. La personalización permite a las empresas ofrecer una experiencia de cliente enriquecida y con mayor potencial de fidelización, optimizando el uso de recursos comerciales y mejorando la competitividad en un mercado saturado.

## Recomendaciones

La segmentación puede mejorar significativamente al integrar datos de patrones temporales y bases de datos externas obtenidas de redes sociales o adquiridas de empresas que venden información de mercado. Incorporar variables como interacciones digitales, estacionalidad y comportamiento de compra permite identificar cambios dinámicos en las preferencias de los clientes. Además, el acceso a datos externos enriquece los análisis con información relevante como intereses específicos, tendencias de mercado y comportamientos observados fuera del entorno interno de la empresa.

Esta integración de bases de datos disponibles en el mercado no solo optimiza la precisión de la segmentación, sino que también facilita la personalización de campañas de marketing adaptadas a datos actualizados y patrones emergentes. Al emplear esta combinación, se pueden ajustar estrategias en tiempo real para responder a las fluctuaciones del mercado, logrando una conexión más efectiva con los clientes y un mayor retorno sobre la inversión.

Optimización continua de modelos predictivos, a medida que se recolectan más datos, es fundamental realizar ajustes periódicos en los modelos supervisados para garantizar su precisión y relevancia en un entorno dinámico. La aplicación de técnicas como Grid Search, Random Search o enfoques más avanzados como la optimización bayesiana permite recalibrar los hiperparámetros de los modelos, optimizando su desempeño en función de los cambios observados en los datos. Esto es particularmente importante para adaptarse a tendencias emergentes en el mercado, patrones de comportamiento de clientes y cambios en las características demográficas o de consumo.

Este proceso de ajuste no solo mejora la capacidad predictiva de los modelos, sino que también asegura su capacidad para integrar nuevos datos de manera eficiente. Por ejemplo, a

medida que los clientes adoptan nuevos hábitos de compra o interactúan con diferentes canales digitales, un modelo ajustado periódicamente puede incorporar estas variaciones y clasificar con mayor precisión a nuevos clientes en los segmentos correspondientes. Este enfoque proactivo permite que las estrategias de marketing sigan siendo efectivas y alineadas con las necesidades del mercado, maximizando el impacto de las decisiones basadas en datos y el retorno de la inversión en la analítica predictiva.

Desarrollo de campañas de retención específicas Dado que algunos segmentos muestran una alta frecuencia de compra pero menor ingreso promedio, se recomienda diseñar campañas de retención que se enfoquen en estos grupos para incrementar la lealtad. Ofrecer beneficios o descuentos especiales puede fortalecer la relación con estos clientes, aumentando su valor a largo plazo para la empresa.

Las empresas deben priorizar la construcción de una base de datos robusta como un pilar fundamental para el desarrollo de estrategias basadas en datos, especialmente aquellas que aún no cuentan con un nivel de madurez analítica adecuado. Una base de datos bien estructurada y rica en información permite no solo optimizar las decisiones operativas y estratégicas, sino también sentar las bases para implementar una cultura del dato y un sólido gobierno de datos dentro de la organización. Estos elementos son esenciales para garantizar la calidad, seguridad y disponibilidad de los datos, lo que a su vez fomenta la confianza en su uso y promueve la toma de decisiones fundamentadas.

En paralelo, la fidelización de clientes debe abordarse desde una perspectiva psicológica, comprendiendo sus necesidades, comportamientos y emociones para construir relaciones sólidas y de largo plazo. Esto requiere métodos efectivos de captura de información, como encuestas, análisis de interacciones digitales y programas de lealtad, que permitan recopilar datos relevantes

sobre preferencias y hábitos. Al integrar estos enfoques con una base de datos robusta, las empresas pueden diseñar estrategias personalizadas que no solo fortalezcan la relación con el cliente, sino que también optimicen los procesos de recolección y análisis de información, asegurando una ventaja competitiva en el mercado.

## Referencias Bibliográficas

Alsayat, A. (2023). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Computing & Applications*, 35(6), pp. 4701-4722. Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca | Neural Computing and Applications (springer.com)

Bratina, D; Faganel, A. (2023). Using Supervised Machine Learning Methods For Rfm Segmentation: A Casino Direct Marketing Communication Case. *Market-Trziste*, 35(1), pp. 7-22. Using Supervised Machine Learning Methods for RFM Segmentation: A Casino Direct Marketing Communication Case (srce.hr)

Chavhan, S; Dharmik, RC; Jain, S; Kamble, K. (2022). RFM Analysis For Customer Segmentation Using Machine Learning: A Survey Of A Decade Of Research. *3C Tic*, 11(2), pp. 166-173. art-13-3c-tic-ed41-vol11-n2-RFM-analysis-for-customer-segmentation-using-machine-learning-a-survey-of-a-decade-of-research.pdf (3ciencias.com)

Geiler, L; Affeldt, S; Nadif, M. (2022). An effective strategy for churn prediction and customer profiling. *Data & Knowledge Engineering*, 142, 102100. An effective strategy for churn prediction and customer profiling - ScienceDirect

Griva, A; Zampou, E; Stavrou, V; Papakiriakopoulos, D; Doukidis, G. (2024). A Two-Stage Business Analytics Approach To Perform Behavioural And Geographic Customer Segmentation Using E-Commerce Delivery Data. *Journal Of Decision Systems, 33(1)*, 1-29. Full article: A two-stage business analytics approach to perform behavioural and geographic customer segmentation using e-commerce delivery data (tandfonline.com)

Joung, J; Kim, H. (2023). Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews. *International Journal Of Information Management*, 70, 102641. Interpretable machine learning-based approach for customer segmentation for new product development from online product reviews - ScienceDirect

Lee, NT; Lee, HC; Hsin, JS; Fang, SH. (2024). Prediction of Customer Behavior Changing via a Hybrid Approach. *IEEE Open Journal Of The Computer Society, 5*, 27-38. Prediction of Customer Behavior Changing via a Hybrid Approach | IEEE Journals & Magazine | IEEE Xplore

Luo, L; Li, B; Fan, XH; Wang, Y; Koprinska, I; Chen, F. (2023). Dynamic customer segmentation via hierarchical fragmentation-coagulation processes. *Machine Learning, 112(1)*, 281-310. Dynamic customer segmentation via hierarchical fragmentation-coagulation processes | Machine Learning (springer.com)

Nakao, J; Nishi, T. (2022). A bilevel production planning using machine learning-based customer modeling. *Journal Of Advanced Mechanical Design Systems And Manufacturing*, 16(4), 21-00393. A bilevel production planning using machine learning-based customer modeling (jst.go.jp)

Nguyen, SP. (2021). Deep customer segmentation with applications to a Vietnamese supermarkets' data. *Soft Computing*, 25(12), pp. 7785-7793. Deep customer segmentation with applications to a Vietnamese supermarkets' data | Soft Computing (springer.com)

Nilashi, M; Ahmadi, H; Arji, G; Alsalem, KO; Samad, S; Ghabban, F; Alzahrani, AO; Ahani, A; Alarood, AA. (2021). Big social data and customer decision making in vegetarian

restaurants: A combined machine learning method. *Journal Of Retailing and Consumer Services, 62*, 102630. Big social data and customer decision making in vegetarian restaurants: A combined machine learning method - ScienceDirect

Nilashi, M; Samad, S; Minaei-Bidgoli, B; Ghabban, F; Supriyanto, E. (2021). Online Reviews Analysis for Customer Segmentation through Dimensionality Reduction and Deep Learning Techniques. *Arabian Journal For Science And Engineering*, 46(9), pp. 8697-8709. Online Reviews Analysis for Customer Segmentation through Dimensionality Reduction and Deep Learning Techniques | Arabian Journal for Science and Engineering (springer.com)

Rachman, FP; Santoso, H; Djajadi, A. (2021). Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation. *International Journal Of Advanced Computer Science And Applications, 12(10)*, 218-227. (PDF) Machine Learning Mini Batch K-means and Business Intelligence Utilization for Credit Card Customer Segmentation (researchgate.net)

Rogic, S; Kascelan, L. (2021). Class Balancing in Customer Segments Classification Using Support Vector Machine Rule Extraction and Ensemble Learning. *Computer Science And Information Systems*, 18(3), pp. 893-925. DOISerbia - Class balancing in customer segments classification using support vector machine rule extraction and ensemble learning - Rogić, Sunčica; Kašćelan, Ljiljana (nb.rs)

Salminen, J; Mustak, M; Sufyan, M; Jansen, BJ. (2023). How can algorithms help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation. *Journal Of Marketing Analytics*, 11(4), pp. 677-692. How can algorithms

help in segmenting users and customers? A systematic review and research agenda for algorithmic customer segmentation | Journal of Marketing Analytics (springer.com)

Sun, YC; Liu, HY; Gao, Y. (2023). Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model. *Heliyon, 9(2)*, e13384. Research on customer lifetime value based on machine learning algorithms and customer relationship management analysis model: Heliyon (cell.com)

Ullah, A; Mohmand, MI; Hussain, H; Johar, S; Khan, I; Ahmad, S; Mahmoud, HA; Huda, S. (2023). Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time. *Sensors*, 23(6), 3180. Sensors | Free Full-Text | Customer Analysis Using Machine Learning-Based Classification Algorithms for Effective Segmentation Using Recency, Frequency, Monetary, and Time (mdpi.com)

Wu, SL; Yau, WC; Ong, TS; Chong, SC. (2021). Integrated Churn Prediction and Customer Segmentation Framework for Telco Business. *IEEE Access*, 9, pp. 62118-62136. Integrated Churn Prediction and Customer Segmentation Framework for Telco Business | IEEE Journals & Magazine | IEEE Xplore

Yadegaridehkordi, E; Nilashi, M; Nasir, MHNB; Momtazi, S; Samad, S; Supriyanto, E; Ghabban, F. (2021). Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques. *Technology In Society, 65*, 101528. Customers segmentation in eco-friendly hotels using multi-criteria and machine learning techniques - ScienceDirect

Yuan, YX; Dehghanpour, K; Bu, FK; Wang, ZY. (2020). A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand. *IEEE Transactions*

*On Power Systems*, 35(5), pp. 4026-4035. A Data-Driven Customer Segmentation Strategy Based on Contribution to System Peak Demand | IEEE Journals & Magazine | IEEE Xplore

Mirantika, Nita; Rijanto, Estiko. (2023). Comparative Analysis of K-Means and K-Menoids Algorithms in Determinig Customer Segmentation Using RFM Model. *Journal of Engineering Science and Technology*, (18), 2340-2351. COMPARATIVE ANALYSIS OF K-MEANS AND K-MEDOIDS ALGORITHMS IN DETERMINING CUSTOMER SEGMENTATION USING RFM MODEL-Web of Science Core Collection (unad.edu.co)

Mirfakhraei, S; Abdolvand, N; Harandi, SR. (2024). The RFMRv Model for Customer Segmentation Based on the Referral Value. *Univ Tehran*, (17), pp. 455-473. The RFMRv Model for Customer Segmentation Based on the Referral Value (ut.ac.ir)