OPTIMIZACIÓN DE CAMPAÑAS DE MARKETING MEDIANTE TÉCNICAS DE CLASIFICACIÓN

Ángela Mellado Salinas

Grado en Estadística Empresarial

Tutor/a: María Asunción Martínez Mayoral

Universidad Miguel Hernández

Facultad de Ciencias Sociales y Jurídicas de Elche

Curso académico 2025-2026

Índice de contenidos

1. Resumen

Este Trabajo Fin de Grado tiene como objetivo optimizar las campañas de marketing mediante el uso de técnicas de clasificación aplicadas a la base de datos *Marketing Campaign* disponible en Kaggle. A partir de información sobre el perfil sociodemográfico y los hábitos de consumo de 2.240 clientes, se pretende identificar los factores que influyen en la aceptación de las campañas y desarrollar modelos predictivos que ayuden a mejorar la segmentación y asignación de recursos publicitarios. Para ello, se aplican y comparan tres modelos de clasificación: regresión logística, Random Forest y Gradient Boosting. Los resultados muestran que el gasto en vinos, seguido de carnes y productos de oro, son factores clave para explicar la respuesta positiva de los clientes, mientras que ciertas variables sociodemográficas, como número de hijos, tienen menor relevancia. Entre los modelos evaluados, el Gradient Boosting ofrece el mejor equilibrio entre exactitud y capacidad discriminativa. En conjunto, el estudio demuestra que el aprendizaje automático permite identificar patrones de comportamiento más precisos que el análisis exploratorio, facilitando la implementación de campañas segmentadas, personalizadas y más efectivas.

2. Palabras clave

- Campañas de marketing.
- Aprendizaje automático.
- Modelos de clasificación.
- Regresión logística.
- Random Forest.
- Gradient Boosting.

3. Contexto

En un mundo cada vez más digitalizado y competitivo, las campañas de marketing desempeñan un papel crucial en el éxito de las empresas. Estas estrategias no solo permiten atraer clientes y aumentar las ventas, sino que también ayudan a construir y consolidar la imagen de marca, mejorar la relación con los consumidores y generar ventajas competitivas en el mercado.

Las empresas invierten sumas considerables en marketing para captar la atención de los consumidores. Según datos de eMarketer (2023), el gasto global en publicidad superó los 600.000 millones de dólares en 2023 y se espera que continúe en aumento. Empresas como Amazon y Coca-Cola destinan miles de millones de dólares a campañas publicitarias anuales, lo que demuestra la relevancia de este sector. Solo Amazon invirtió más de $21 mil millones en publicidad y promoción en 2024, consolidándose como el mayor anunciante del mundo,

mientras que Coca-Cola mantiene su fuerte compromiso, enfocando ahora más del 65% de su presupuesto en canales digitales (Marketers by Adlatina, 2025; Ekos, 2024). No obstante, no todas las campañas tienen el mismo impacto; algunas logran una rentabilidad significativa, mientras que otras fracasan en alcanzar sus objetivos. De ahí la importancia de estudiar los factores que determinan el éxito de una campaña publicitaria.

Un ejemplo destacado de una campaña publicitaria exitosa es la iniciativa "Red Bull Stratos" de 2012. En esta campaña, Red Bull patrocinó al paracaidista Felix Baumgartner para que realizara un salto en caída libre desde la estratosfera, a una altura de aproximadamente 39 kilómetros. Este evento no solo capturó la atención mundial, sino que también estableció nuevos estándares en el marketing experiencial. El presupuesto de la campaña se estimó en más de 30 millones de dólares, según análisis de Marketing Ideas 101 (2024), lo que representa una inversión significativa dentro de las estrategias publicitarias de la marca. Los resultados fueron impresionantes: durante el evento, aproximadamente 8 millones de personas lo vieron en vivo, y Red Bull experimentó un aumento del 7 % en sus ventas en EE. UU., lo que se tradujo en un valor agregado estimado de 1.6 mil millones de dólares para la marca (Marketing Ideas 101, 2024). Este caso demuestra cómo una campaña de marketing innovadora y bien ejecutada puede generar un impacto significativo tanto en la percepción de la marca como en sus resultados financieros.

El impacto de las estrategias de marketing se puede plantear en distintos aspectos del negocio. Una campaña bien diseñada puede aumentar la participación de mercado, fidelizar clientes y mejorar la reputación de la empresa. Últimamente, con la evolución de la tecnología y el auge de las redes sociales, las estrategias han cambiado radicalmente. Hoy en día, el marketing basado en datos permite personalizar las campañas según el perfil del consumidor, optimizando los recursos. Empresas como Netflix y Spotify utilizan algoritmos de aprendizaje automático para analizar el comportamiento de los usuarios y ofrecer recomendaciones personalizadas (iArtificial, 2025), lo que ha revolucionado la forma en que interactúan con sus audiencias.

Además, las campañas de marketing no solo benefician a las empresas, sino también a los consumidores. A través de estrategias dirigidas y segmentadas, los clientes reciben información relevante sobre productos y servicios que se ajustan a sus necesidades y preferencias. De este modo, las empresas pueden maximizar sus ventas mientras que los consumidores encuentran más fácilmente lo que buscan.

En este TFG analizamos la base de datos *Marketing Campaign*, disponible en Kaggle (https://www.kaggle.com/datasets/rodsaldanha/arketing-campaign), con el objetivo de utilizar el comportamiento y patrones de consumo de los clientes para explicar la efectividad de las campañas de marketing realizadas y con estos resultados planificar futuras estrategias de publicidad.

Varios investigadores han desarrollado enfoques distintos para analizar la respuesta del cliente ante estrategias de marketing o para aplicar técnicas de análisis de datos que pueden ser útiles en este contexto. Por ejemplo, Javanmardi et al (2025) aplicaron técnicas de análisis cuantitativo y Grey System Theory (basada en el estudio de sistemas inciertos con información parcialmente conocida) para identificar los factores que influyen en la aceptación de la realidad virtual en el turismo, destacando la importancia de los factores demográficos y psicológicos en la adopción de nuevas tecnologías.

En un estudio reciente, Venkateswaran (2025) analizó el impacto de las herramientas de análisis predictivo impulsadas por inteligencia artificial en el marketing digital. El autor utilizó un enfoque cuantitativo para evaluar cómo estas herramientas pueden predecir el comportamiento del cliente y mejorar la personalización de las campañas, optimizando la toma de decisiones estratégicas. El estudio destaca la capacidad de los modelos predictivos de IA para identificar patrones en los datos del consumidor, mejorar la segmentación y aumentar la efectividad de las estrategias de marketing digital.

Por otro lado, Usman et al (2024) analizaron la detección de ataques en el tráfico web utilizando métodos de ensamblaje (basados en la combinación de modelos) y selección de características. Aplicaron modelos de aprendizaje automático, incluyendo Random Forest y Extreme Gradient Boosting, para identificar patrones de ataques en un conjunto de datos simulado, y consiguieron mejoras significativas en la precisión de la detección.

Asimismo, Othman et al (2023) analizaron el comportamiento del consumidor utilizando datos de redes sociales y técnicas de aprendizaje profundo. El estudio aplicó modelos avanzados de predicción para evaluar cómo los consumidores responden a diferentes estrategias de marketing digital, destacando la capacidad del aprendizaje profundo para identificar patrones de comportamiento y mejorar la efectividad de las campañas, sin enfocarse en segmentación por edad o ingresos específicos.

Finalmente, Ullah et al (2022) aplicaron técnicas de aprendizaje automático para analizar señales EEG de los consumidores y predecir sus preferencias hacia productos en línea. Utilizaron clasificadores como redes neuronales artificiales, máquinas de soporte vectorial, regresión logística, árboles de decisión y k-means, logrando una precisión de hasta 81.23% en la predicción de preferencias por producto. El estudio demuestra cómo la combinación de análisis de señales neurológicas y machine learning puede mejorar la comprensión de la respuesta del consumidor y optimizar estrategias de marketing digital.

Todos estos estudios ofrecen una base sólida de ideas y métodos que pueden aprovecharse en este trabajo para identificar qué factores influyen realmente en el éxito de una campaña de marketing. A su vez, sirven de inspiración para aplicar técnicas que mejoren la toma de decisiones, hagan las campañas más efectivas y logren una conexión más profunda con las necesidades de los clientes.

4. Objetivos

Al trabajar sobre la base de datos *Marketing Campaign*, en Kaggle. se busca en este trabajo, desarrollar un modelo predictivo que permita optimizar la asignación de recursos en futuras campañas de marketing, aumentando la tasa de respuesta y los ingresos percibidos, y reduciendo los costes operativos, para conseguir una mayor tasa de éxito en la respuesta de los clientes a dichas campañas, y en consecuencia a las compras que realicen.

A partir del objetivo general, se plantean los siguientes objetivos específicos, que guiarán el desarrollo del trabajo:

1. Describir y comparar el comportamiento de los clientes en la aceptación de alguna de las 5 campañas de marketing implementadas por la empresa cuyos datos utilizaremos en nuestro trabajo fin de grado.
2. Desarrollar modelos de clasificación válidos para diferenciar a los clientes en función de su comportamiento en la aceptación de alguna de las campañas de marketing lanzadas.
3. Identificar las variables relevantes en la clasificación y caracterizar los diferentes tipos de clientes según su respuesta a las campañas publicitarias, en función de sus tendencias y hábitos de compra (preferencia de productos, volumen y coste de compras, frecuencia, descuentos, plataforma de compra (web, físico, catálogo), así como de otros factores socioeconómicos como la edad, estado civil, ingresos, educación, nº hijos y adolescentes en la familia.
4. Ajustar y comparar varios modelos de clasificación para diferenciar a los clientes que aceptan de los que no aceptan alguna campaña de marketing. Encontrar el modelo más preciso en la clasificación de los clientes para utilizarlo como herramienta de predicción y con él poder planificar y testar diferentes estrategias de marketing.

5. Información disponible

5.1 Base de datos

Los datos que utilizamos en este trabajo, denominados *Marketing Campaign* en Kaggle, fueron recopilados en 2014 por Rodolfo Saldanha, del SAS Institute, y posteriormente publicados en Kaggle. La información proviene de encuestas a clientes y registros internos de la empresa, abarcando comportamientos de compra, datos del núcleo familiar y respuestas a diversas campañas de marketing. La información fue obtenida con el objetivo de predecir qué clientes responderán a una oferta de producto o servicio. La recopilación se llevó a cabo siguiendo la metodología descrita en *Analítica de negocios con SAS Enterprise Guide y SAS Enterprise Miner* (Instituto SAS, 2014). El análisis de los datos se estructura en tres fases principales: análisis exploratorio de datos, modelización mediante técnicas de aprendizaje automático y descripción de los recursos utilizados, lo que permite comprender la estructura de

la información, construir modelos predictivos precisos y asegurar la reproducibilidad y transparencia del proceso analítico.

Los datos se encuentran disponibles en un solo archivo con 2240 registros de clientes y 29 variables. No se requiere integración de otras fuentes de información.

5.2 Variables disponibles

Las variables disponibles y que utilizaremos en este trabajo, son las siguientes:

- Año de nacimiento: Año de nacimiento del cliente (fecha numérica, rango de variación entre 1893 y 1996).
- Educación: Nivel de educación del cliente (categórica, con niveles: “Básico”, “Ciclo 2n”, “Graduación”, “Master”, “PhD”).
- Marital Status: Estado civil (categórica, con niveles: “Soltero”, “En pareja”, “Casado”, “Divorciado”, “Viudo”).
- Income: Ingresos anuales del hogar del cliente (numérica, rango de variación entre $3000 y $1500000).
- Kids Home: Número de niños en el hogar (numérica, rango de variación entre 0 y 5).
- Teen Home: Número de adolescentes en el hogar (numérica, rango de variación entre 0 y 5).
- Mnt Wines, Mnt Fruits, Mnt Meat Products, Mnt Fish Products, Mnt Sweet Products, Mnt Gold Prods: Cantidad en dólares, gastada en distintas categorías de productos en los últimos dos años: vino, fruta, carne, pescado, dulces y joyas (numéricas, rango de variación entre 0 y 1000).
- Num deals Purchases: Número de compras realizadas con descuento (numéricas, rango de variación entre 0 y 50).
- Num Web Purchases: Número de compras realizadas por la web (numéricas, rango de variación entre 0 y 20).
- Num Catalog Purchases: Número de compras a través de catálogos (numéricas, rango de variación entre 0 y 15).
- Num Store Purchases: Número de compras en tiendas físicas (numéricas, rango de variación entre 0 y 40).
- Accepted Cpm1, Cpm2, Cpm3, Cpm4 Cpm5: Si el cliente aceptó o no, la oferta de cada una de las cinco campañas de marketing lanzadas (dicotómica con niveles 1=Yes, 0=No).
- Dt _Customer: Fecha de registro del cliente en la empresa (fecha numérica, rango de variación entre 2012 y 2014).
- Num WebVisit Month: Número de visitas al sitio web en el último mes (numéricas, rango de variación entre 0 y 20).

Entre las variables disponibles en la base de datos y que no se utilizaron en este trabajo se encuentra el ID, que corresponde al identificador único del cliente, que no aporta valor analítico. La variable Recency, que mide los días desde la última compra, se decidió excluirla porque el estudio se centró en patrones globales de compra y no en la inmediatez de la última transacción. Por otro lado, Complain, que indica si el cliente presentó una reclamación en los últimos dos años, fue descartada debido a su baja frecuencia en la muestra. Finalmente, las variables Z Cost Contact y Z Revenue no se usaron al no estar especificado su significado y además ser constantes en todos los registros, y la variable Response, que señala si el cliente aceptó la última campaña, se excluyó para evitar redundancia, dado que el análisis se centró en las cinco campañas anteriores (Cmp1 a Cmp5).

5.3 Procesado de los datos

Como resultado del análisis exploratorio, se identifican las siguientes carencias en los datos disponibles, y se tratan en consecuencia para corregirlas.

Identificación de datos faltantes y tratamiento

Se ha detectado la presencia de valores faltantes en la variable Income, con un total de 24 registros sin información. Dado que la base de datos cuenta con 2.240 observaciones, estos valores representan un porcentaje reducido del total y por lo tanto, se ha decidido eliminar esos 24 registros.

Transformaciones de variables y recodificaciones

Para posibilitar un mejor tratamiento de la información disponible, se ha procedido a transformar las siguientes variables:

Marital Status (Estado civil), que inicialmente consta de 5 categorías, se ha recodificado en dos únicas categorías que son:

- “Solo” , con las categorías Soltero, Divorciado y Viudo.
- “En pareja” las categorías Casado y En pareja.

Education (Nivel educativo), que constaba de cinco categorías, se ha recodificado en tres categorías:

- Educación básica, que corresponde a la categoría original “Básico”.
- Educación media, que agrupa las categorías “Ciclo 2n” y “Graduación”.
- Educación avanzada, que agrupa las categorías “Máster” y “PhD”.

En la variable Income (Ingreso anual), se han creado tres categorías para diferenciar clientes con distinta capacidad adquisitiva, que son:

- Bajo: con ingresos inferiores a $30.000 anuales.
- Medio: ingresos entre $30.000 y $60.000 anuales.
- Alto: ingresos superiores a $60.000 anuales.

Creación de nuevas variables

- Acepta_campañas : Se crea como variable objetivo, indica si el cliente ha aceptado alguna campaña de las cinco disponibles (1), o no ha aceptado ninguna (0). Se obtiene pues, a partir de las variables Accepted Cpm1, Cpm2, Cpm3, Cpm4 y Cpm5.
- Total_hijos: Obtenida de la suma de la cantidad de niños (kids home) y adolescentes en casa (teens home).
- Antigüedad: Que indica los días que el cliente lleva registrado en la empresa, calculada a partir de la fecha de registro (Dt_Customer) con respecto al 1 de septiembre de 2014, que es la fecha de recogida de los datos .

5.4 Variables disponibles y objetivos de investigación

Para alcanzar los objetivos planteados, es fundamental identificar cuál es la variable objetivo y cuáles utilizaremos como variables predictoras para construir modelos de clasificación precisos.

La variable objetivo es Acepta_campañas, que indica si un cliente aceptó al menos una de las campañas de marketing implementadas. Esta variable permite clasificar a los clientes según su respuesta global a las campañas, en alguno de los lanzamientos.

Las variables predictoras incluyen tanto características sociodemográficas como patrones de comportamiento de compra de los clientes. Entre las variables sociodemográficas utilizaremos: Año de nacimiento, Estado civil (Marital Status), Educación, Ingreso anual (Income) y Total_hijos, que representa el número total de menores en el hogar.

En cuanto a las variables de comportamiento de compra, se consideran las cantidades gastadas en distintas categorías de productos en los últimos dos años (MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds) y la frecuencia de compras a través de distintos canales (NumDealsPurchases, NumWebPurchases, NumCatalogPurchases y NumStorePurchases), así como el número de visitas al sitio web en el último mes (NumWebVisitMonth) y la antigüedad de los clientes en la empresa, .

Todas estas variables predictoras se utilizan para desarrollar modelos de clasificación que permitan diferenciar a los clientes que aceptan al menos una campaña de marketing de los que

no, identificar los factores más relevantes y optimizar la asignación de recursos en futuras campañas.

6. Metodología

La metodología empleada en este trabajo, tras definir los objetivos, se estructura en dos puntos principales: el análisis exploratorio de datos y la fase de modelización mediante técnicas de aprendizaje automático. En esta misma sección incluiremos información sobre los recursos de software y hardware utilizados. Este planteamiento permite, en primer lugar, comprender las características y distribución de los datos; en segundo lugar, construir modelos predictivos adecuados para los objetivos planteados; y finalmente, garantizar la reproducibilidad y transparencia del análisis a través de la especificación de las herramientas utilizadas.

6.1 Análisis exploratorio

El análisis exploratorio es una fase inicial clave que ayuda a comprender el conjunto de datos, evaluar patrones y relaciones, identificar variables predictoras potenciales y sentar las bases para la modelización supervisada. Este enfoque asegura que las decisiones estratégicas posteriores y los modelos predictivos estén fundamentados en una comprensión sólida y coherente de los datos.

En el contexto del comportamiento de clientes frente a campañas de marketing, este análisis permite identificar variables relevantes que podrían influir en la respuesta a las campañas publicitarias.

Una parte fundamental del análisis exploratorio consiste en visualizar la distribución de la variable objetivo, que en este caso corresponde a la aceptación o no de campañas de marketing. Este procedimiento ayuda a detectar posibles desequilibrios entre clases y a orientar la definición de estrategias de segmentación. Para examinar la relación entre la variable objetivo y las características sociodemográficas o de consumo de los clientes, se emplean gráficos como boxplots (para las variables numéricas) y gráficos de barras (para las variables categóricas). Estos gráficos permiten detectar asociaciones y facilitan la identificación de factores y covariables que podrían ser útiles como predictores en los modelos a ajustar. Además, permiten generar hipótesis sobre el comportamiento del cliente para plantear modelos de predicción/clasificación mejorados.

6.2 Modelos de clasificación

En el ámbito del análisis de datos y el aprendizaje automático, los modelos de clasificación constituyen una de las herramientas más utilizadas para la toma de decisiones. Su objetivo principal es asignar una categoría de clasificación en la variable respuesta para cada observación, en función de un conjunto de variables explicativas, y con ella conseguir una buena predicción, útil para predecir comportamientos futuros en otros clientes.

La clasificación se aplica en una gran variedad de contextos: detección de fraudes en transacciones bancarias, diagnóstico médico a partir de datos clínicos, segmentación de clientes en marketing, entre otros. La capacidad de estos modelos para transformar datos históricos en conocimiento útil los convierte en una pieza fundamental dentro del proceso de apoyo a la toma de decisiones.

En este estudio se trabaja con tres modelos de clasificación supervisada aplicados a la predicción de la aceptación de ofertas comerciales por parte de los clientes. Se explicarán en esta sección sus características principales, el procedimiento de ajuste, las métricas utilizadas para evaluar su desempeño, y con las que se podrá comparar su efectividad, y cómo obtener información sobre las predictoras más relevantes para ayudar a la toma de decisiones estratégicas en marketing. Se presentarán también las ideas clave sobre validación cruzada y curvas de aprendizaje, para concluir sobre la robustez del ajuste. Concretamente, se analizan tres enfoques: la Regresión Logística Binaria, el Random Forest y el Gradient Boosting. La regresión logística se utiliza por su simplicidad, robustez e interpretabilidad a la hora de identificar qué variables predicen la respuesta, mientras que los modelos de Random Forest y Gradient Boosting, permiten capturar relaciones no lineales y mejorar la precisión en la clasificación y generalización del modelo.

En términos generales, cada modelo requiere un preprocesamiento adecuado de los datos, una división entre entrenamiento y test, una fase de optimización de hiperparámetros y una evaluación mediante métricas estándar y técnicas de validación. Posteriormente, se identifican las variables que mayor impacto tienen en la clasificación.

6.2.1 Modelo logístico binario

La regresión logística binaria es un tipo de análisis de regresión utilizado para predecir el resultado de una variable categórica dicotómica (con solo dos valores posibles, como *sí/no* o *1/0*) en función de un conjunto de variables explicativas, sean numéricas o categóricas. Este modelo se enmarca dentro de la familia de los modelos lineales generalizados (GLM), diseñados para extender la regresión lineal a casos donde la variable dependiente no sigue una distribución normal, como es el caso de este tipo de respuestas, donde la variable tiene una distribución Bernoulli.

El objetivo de este modelo es pues estimar la probabilidad de que ocurra un evento (éxito) frente a que no ocurra (fracaso). En este contexto, el éxito será la aceptación de alguna campaña, y el fracaso la no aceptación de ninguna, y la probabilidad de interés, p, será la probabilidad de aceptar alguna campaña.

La denominación *logística* proviene de la función logística, definida como:

$$f(t) = \frac{1}{1 + e^{-t}},$$

que transforma un parámetro definido en la recta real, en una probabilidad comprendida entre 0 y 1. Esta característica la hace especialmente adecuada para modelar datos binarios a través de un predictor lineal que puede tomar valores en toda la recta real.

El modelo logit estima la relación entre las variables independientes X y el logaritmo de la razón de probabilidades, conocido como *logit*:

$$\operatorname{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

El ajuste de los parámetros se realiza mediante el método de máxima verosimilitud, y el resultado del modelo siempre se expresa en forma de probabilidad. Para conseguir una clasificación a partir de la estimación de estas probabilidades, esto es, distinguir entre un cliente que acepta o no una campaña, es necesario definir un umbral de corte. Habitualmente se adopta el valor de 0.5, de modo que si la probabilidad estimada es mayor a este umbral, la predicción será "1", esto es, se clasificará como que acepta alguna campaña (éxito), y en caso contrario, como "0" o fracaso.

La regresión logística binaria es uno de los modelos más empleados en Estadística y aprendizaje automático por su simplicidad, robustez e interpretabilidad, lo que la hace especialmente útil en contextos como medicina, marketing o ciencias sociales, aunque puede ser limitada frente a métodos más flexibles como *Random Forests* o *Gradient Boosting* cuando existen relaciones no lineales complejas.

6.2.2 Random forest

El Random Forest es un algoritmo de aprendizaje automático de tipo ensamble, basado en la predicción conjunta con muchos árboles de decisión ajustados sobre los datos. Su objetivo es mejorar la precisión y la capacidad de generalización mediante la construcción de

múltiples árboles y la combinación de sus predicciones. Se aplica tanto en problemas de clasificación (variables categóricas) como en regresión (variables continuas).

El nombre de "bosque aleatorio" proviene de la idea de crear un conjunto de árboles de decisión que se entrenan de manera independiente utilizando diferentes subconjuntos de datos y variables elegidos aleatoriamente en cada ocasión. Esta aleatoriedad introduce diversidad entre los árboles, lo que reduce el riesgo de sobreajuste característico de un único árbol de decisión.

En el caso de clasificación, cada árbol predice una clasificación y el modelo selecciona la clase más votada. En regresión, el resultado se obtiene como el promedio de las predicciones individuales de cada árbol.

El modelo se basa en dos ideas clave que incrementan la robustez del modelo, frente a los árboles de decisión:

1. Bagging (Bootstrap Aggregating): se generan distintas muestras aleatorias con reemplazo del conjunto de datos original, de modo que cada árbol se entrena con un subconjunto distinto, capturando así la diversidad existente en los datos, para generalizar mejor cuando se mezclen todas las predicciones.

2. Selección aleatoria de variables: en cada división del árbol, en lugar de considerar todas las variables predictoras, se evalúa un subconjunto aleatorio, para reducir los problemas de multicolinealidad entre predictores y también la correlación entre árboles.

Para ajustar un bosque aleatorio hemos de decidir sobre el número de árboles en el bosque, la profundidad máxima de los árboles y el número de variables consideradas en cada división. Utilizaremos valores estándar para optimizar estos hiperparámetros, adaptados a los datos que nos ocupan, y utilizando búsqueda en grid del óptimo.

Una ventaja fundamental del Random Forest es que ofrece una medida de importancia de las variables, lo que permite identificar cuáles son más influyentes en la clasificación. Además, su rendimiento suele ser muy alto incluso sin necesidad de una gran optimización en los hiperparámetros.

El Random Forest es ampliamente utilizado por su robustez, precisión e interpretabilidad relativa, siendo especialmente útil en contextos como medicina, finanzas, marketing y análisis de datos sociales. Sin embargo, puede resultar más costoso en términos computacionales que modelos más simples y, en casos de relaciones muy complejas, puede ser superado por métodos más avanzados como los basados en Boosting.

6.2.3 Gradient boosting

El Gradient Boosting es un algoritmo de aprendizaje automático de tipo ensamble basado también en árboles de decisión. Su objetivo es mejorar la precisión y la capacidad de generalización construyendo secuencialmente múltiples árboles, donde cada nuevo árbol corrige los errores residuales de los anteriores. Se aplica tanto en problemas de clasificación (variables categóricas) como en regresión (variables continuas).

El nombre "Gradient Boosting" proviene de la combinación de:

- boosting (fortalecer un modelo débil mediante impulsos, aprovechando la secuencia de modelos ajustados previamente): cada árbol se entrena para corregir los errores de los árboles anteriores, enfocándose en las observaciones más difíciles de predecir.
- y la idea de gradiente, ya que el algoritmo optimiza una función de pérdida usando técnicas basadas en el gradiente descendiente; se minimiza de forma iterativa una función de pérdida mediante el cálculo de los gradientes, ajustando así las predicciones paso a paso (Wikipedia, 2024).

En clasificación, el modelo Gradient Boosting combina secuencialmente los árboles construidos, donde cada uno corrige los errores residuales del anterior. A diferencia del Random Forest, estos árboles no votan ni sus predicciones se promedian; en su lugar, cada árbol aporta una pequeña corrección al modelo acumulado. La predicción final se obtiene sumando todas estas contribuciones y transformando el resultado mediante una función logística para obtener una probabilidad. Finalmente, la clase asignada es aquella cuya probabilidad estimada es mayor. Este funcionamiento se corresponde con la formulación estándar del Gradient Boosting descrita en Wikipedia (2024).

Los hiperparámetros principales a optimizar en esta técnica incluyen el número de árboles, la profundidad máxima de los árboles, la tasa de aprendizaje (learning rate) y la fracción de datos utilizada en cada iteración. La tasa de aprendizaje controla el tamaño de la contribución de cada árbol ajustado al modelo final: valores más bajos hacen que cada árbol tenga un impacto menor, lo que generalmente mejora la capacidad de generalización y reduce el riesgo de sobreajuste, aunque requiere entrenar más árboles para alcanzar un buen desempeño.

Una ventaja clave del Gradient Boosting es su elevada precisión y su flexibilidad para abordar distintos tipos de problemas y funciones de pérdida. Además, ofrece mecanismos de control del sobreajuste mediante la regularización (selección de variables) y el ajuste de sus hiperparámetros, lo que permite optimizar el desempeño del modelo según las características del conjunto de datos.

El Gradient Boosting es ampliamente utilizado por su rendimiento superior en contextos como finanzas, medicina, marketing y competiciones de ciencia de datos. Sin embargo, suele ser más sensible al sobreajuste y más costoso computacionalmente que modelos como Random

Forest, requiriendo una selección cuidadosa de los valores a utilizar para optimizar los hiperparámetros.

6.2.4 Procedimiento de ajuste

El procedimiento de ajuste de los tres modelos, Regresión Logística Binaria, Random Forest y Gradient Boosting, sigue una misma estructura metodológica. Primero, se realiza el preprocesamiento de los datos, que incluye la estandarización de las variables predictoras numéricas y la codificación de las categóricas.

A continuación, el conjunto de datos se divide en entrenamiento y test, respetando la estratificación derivada del desequilibrio entre los clientes que aceptan alguna campaña y los que no aceptan ninguna. El modelo aprende a partir del conjunto de entrenamiento, mientras que el conjunto test se utiliza para evaluar su capacidad predictiva sobre datos no vistos, proporcionando así una estimación más realista de su rendimiento futuro.

Posteriormente, se procede con la optimización de los hiperparámetros que definen el comportamiento de cada algoritmo. Esta fase es fundamental para mejorar la precisión del modelo y su capacidad de generalización, y se realiza mediante técnicas como la validación cruzada.

El rendimiento de cada modelo se evalúa mediante métricas estándar de clasificación, que permiten valorar la calidad de las predicciones y comparar la eficacia entre modelos. Para reforzar la evaluación, se emplea validación cruzada, con el fin de reducir la dependencia de una única partición de los datos y asegurar la robustez del ajuste. Asimismo, se utilizan curvas de aprendizaje para analizar la evolución del rendimiento conforme aumenta el tamaño de la muestra de entrenamiento, lo que ayuda a detectar problemas de sobreajuste o infraajuste y a valorar si el conjunto de datos disponible es suficiente.

Finalmente, una vez validado el modelo ajustado, se identifican las variables explicativas más relevantes en la clasificación. Este análisis proporciona una visión adicional sobre los factores que influyen en la respuesta de los clientes, ofreciendo información útil para la toma de decisiones estratégicas en marketing.

6.2.5 Evaluación y métricas

En los modelos de clasificación, el informe de resultados suele recoger un conjunto de métricas fundamentales que permiten evaluar el rendimiento y la calidad de las predicciones.

Entre las más utilizadas se encuentran la exactitud (accuracy), la precisión (precision), la sensibilidad o recuerdo (recall), el F1-score y el área bajo la curva ROC o AUC.

La exactitud (accuracy) mide el porcentaje total de predicciones correctas respecto al número total de observaciones. Es una métrica global que ofrece una visión general del rendimiento, aunque puede resultar poco representativa en problemas con clases desequilibradas, como es el caso en nuestros datos.

La precisión (precision) indica el porcentaje de aciertos entre las observaciones clasificadas como positivas por el modelo, esto es, en nuestro caso el porcentaje de aciertos entre todos los clientes que han sido clasificados como sensibles a las campañas de marketing . Una precisión alta implica que el modelo comete pocos falsos positivos, es decir, que rara vez predice un caso positivo cuando en realidad no lo es.

La sensibilidad o recall mide la capacidad del modelo para identificar correctamente los casos de la clase positiva, es decir, en nuestro caso identificaría la tasa de aciertos en la clasificación de los clientes que aceptaron alguna campaña de marketing.. Un valor alto de recall significa que el modelo comete pocos falsos negativos. Puesto que nos interesa la clasificación correcta de los que son sensibles a las campañas de marketing, nos interesa especialmente esta métrica.

La medida F1 (f1-score) combina precisión y recall en un único indicador a través de su media armónica. Se utiliza especialmente en contextos donde es necesario lograr un equilibrio entre ambas métricas, evitando que el modelo se centre únicamente en una de ellas. Dado su carácter promedio, utilizaremos esta métrica para comparar los modelos.

Finalmente, el AUC (Area Under the Curve) evalúa la capacidad discriminativa del modelo, es decir, su habilidad para distinguir entre las clases positiva y negativa. Representa la probabilidad de que el modelo, ante un cliente sensible a las campañas y otro que no lo es, clasifique mejor al sensible que al no sensible. Un valor de AUC cercano a 1 refleja un modelo con alto poder de discriminación, que siempre asignará a un cliente sensible al marketing una probabilidad de ser sensible mayor que a uno que no haya respondido a ninguna campaña de marketing.

En conjunto, estas métricas permiten analizar desde diferentes perspectivas el rendimiento de un modelo de clasificación, proporcionando una visión más completa y detallada de su comportamiento predictivo.

Se recomienda completar la información sobre las métricas en los manuales de Google para Machine Learning.

6.2.6 Validación (curva de aprendizaje y cross-validation score)

En el desarrollo de modelos de aprendizaje automático, es fundamental evaluar no solo el rendimiento final, sino también cómo el modelo se comporta durante el proceso de entrenamiento. Para ello se utilizan herramientas como la curva de aprendizaje (learning curve) y la validación cruzada (cross-validation). En conjunto, la curva de aprendizaje y la validación cruzada son herramientas esenciales para evaluar la capacidad de generalización de un modelo, guiar la selección de parámetros y detectar problemas de ajuste, contribuyendo así a desarrollar modelos predictivos más fiables y sólidos.

La curva de aprendizaje es una representación gráfica que muestra el rendimiento del modelo en función de la cantidad de datos de entrenamiento utilizados. Por lo general, se trazan dos curvas: una para el conjunto de entrenamiento y otra para un conjunto de validación o prueba. Esta visualización permite detectar problemas de sobreajuste (overfitting) o infra-ajuste (underfitting):

- El sobreajuste ocurre cuando el modelo aprende demasiado los datos de entrenamiento, obteniendo muy buen rendimiento en ellos pero fallando al generalizar a datos nuevos.

- El infra-ajuste sucede cuando el modelo no logra capturar la complejidad del problema, mostrando bajo rendimiento tanto en entrenamiento como en validación.
  Las curvas de aprendizaje ayudan a determinar si agregar más datos, ajustar parámetros o cambiar la complejidad del modelo puede mejorar su desempeño.

La validación cruzada es una técnica que consiste en dividir el conjunto de datos en varios subconjuntos o “folds” para entrenar y evaluar el modelo múltiples veces. Cada subconjunto se utiliza como conjunto test, mientras los demás sirven para el entrenamiento, y el resultado devuelto es la media y desviación típica de todos los scores (métrica de clasificación elegida) conseguidos en los diferentes ajustes realizados. Esta metodología permite obtener una estimación robusta del rendimiento del modelo, que se compara con la obtenida en el ajuste inicial, y cuya desviación estándar se valora en términos de variabilidad encontrada al variar de muestra de entrenamiento. Para validar el modelo se esperan estimaciones del score próximas al obtenido en el modelo ajustado al inicio, y desviaciones estándar próximas a cero.

6.3 Software y hardware

Para el desarrollo del análisis se utilizó el lenguaje de programación Python, ampliamente reconocido en ciencia de datos por su versatilidad y la disponibilidad de librerías especializadas para análisis estadístico, visualización y modelización predictiva.

Las principales librerías y módulos utilizados fueron:

- pandas: para la manipulación y análisis de datos tabulares, incluyendo limpieza de datos, transformación de variables y cálculo de estadísticas descriptivas.
- numpy: para operaciones numéricas, manejo de arrays y cálculos matriciales.
- matplotlib y seaborn: para la creación de gráficos estadísticos y visualización de patrones en los datos, incluyendo histogramas, boxplots y gráficos de barras.
- scikit-learn: para la implementación de modelos de aprendizaje supervisado, como regresión logística, Random Forest y Gradient Boosting; así como para el preprocesamiento de datos, la partición de datos, la evaluación mediante validación cruzada y la obtención de métrica.
- kagglehub: para la descarga y manejo de conjuntos de datos disponibles públicamente en la plataforma Kaggle.

El uso de estas librerías permitió realizar todas las etapas del análisis de manera eficiente, desde la limpieza y transformación de datos hasta la construcción, evaluación y comparación de modelos predictivos, garantizando además la reproducibilidad del estudio.

7. Resultados

Se presentan en esta sección los resultados obtenidos en el análisis exploratorio, y a continuación los modelos de clasificación ajustados. Acabaremos con una comparativa entre ellos para concluir sobre el mejor modelo de clasificación para estos datos.

7.1 Análisis Exploratorio de Datos

Iniciamos el análisis exploratorio mostrando, en la Figura 1, el comportamiento de la variable objetivo referida a la aceptación por los clientes de alguna de las campañas publicitarias llevadas a cabo por la empresa.

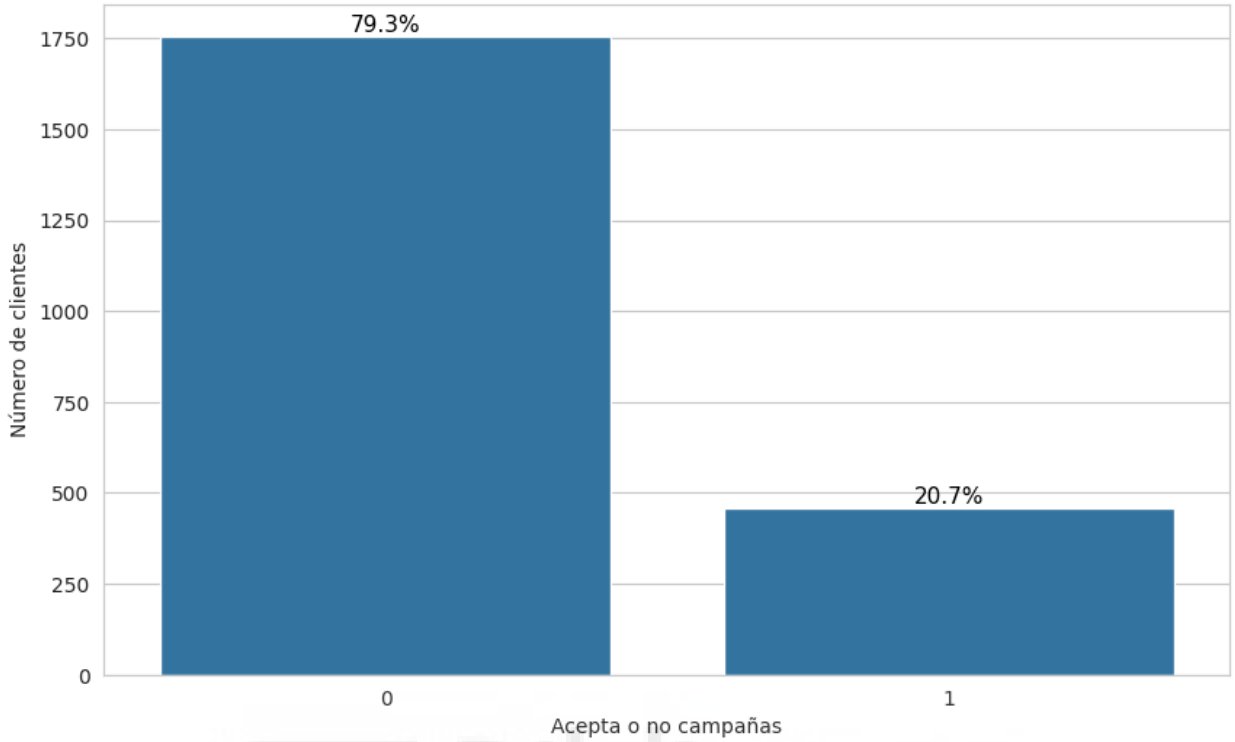

Figura 1: Distribución de la variable objetivo relativa a la aceptación de alguna campaña publicitaria por parte del cliente.

En la Figura 1 se observa que la mayoría de los clientes (el 79,3%) no aceptó ninguna campaña, mientras que tan sólo el 20,7% aceptó alguna. Esta distribución confirma lo observado en análisis previos y evidencia la necesidad de mejorar la personalización de las campañas, enfocándose mejor a los perfiles de los clientes según características como ingresos, edad o canal de compra.
Esta distribución nos resulta útil para reconocer el desequilibrio entre las dos categorías de clasificación, y por lo tanto a utilizar ésta como variable de estratificación en la creación de las muestras de entrenamiento y test.

A continuación podemos observar en la Figura 2 la relación entre la aceptación o no de las campañas por los clientes y el nivel de ingresos de los clientes.

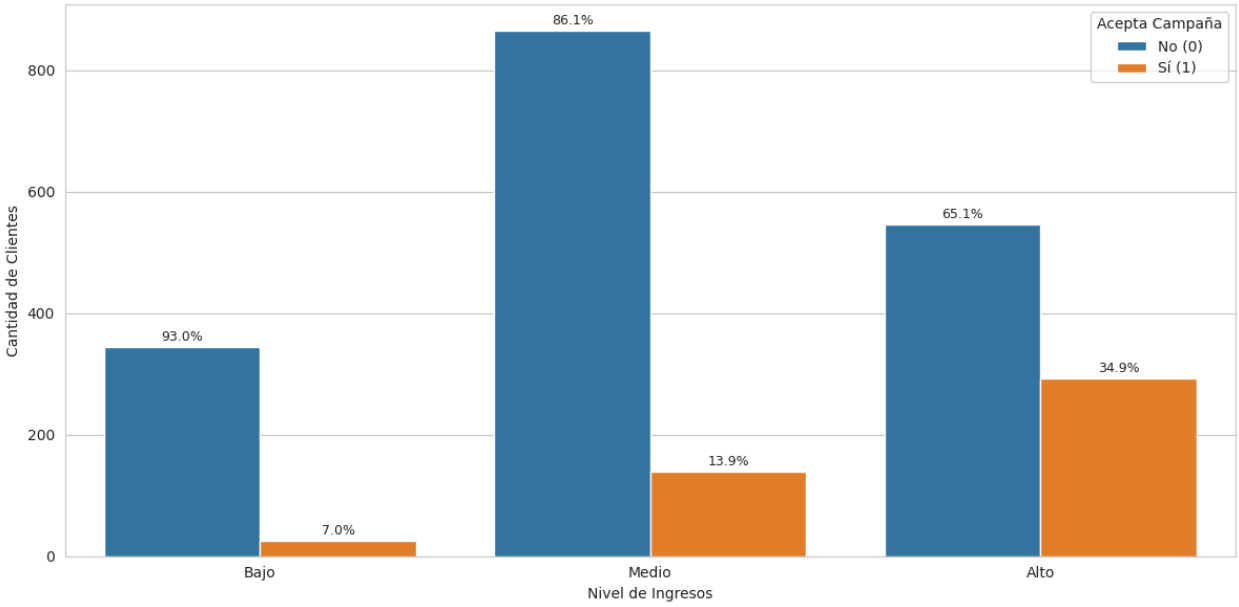

Figura 2: Número de clientes que aceptan o no alguna campaña, en función del nivel de ingresos.

En la Figura 2 se aprecia una distribución distinta de clientes que aceptan y no aceptan campañas, según el nivel de ingresos de los clientes. Los clientes de Ingresos Bajos presentan la menor aceptación (7,0%), mientras que los de Ingresos Altos alcanzan la más elevada (34,9%), lo que sugiere que las campañas son más efectivas entre los clientes con mayor poder adquisitivo, probablemente porque las ofertas resultan más atractivas o accesibles para ellos. El grupo de Ingresos Medios, aunque es el más numeroso (aproximadamente 970 clientes), mantiene una tasa de aceptación moderada (13,9%), lo que indica cierto potencial de mejora si se ajustan las estrategias y mensajes.

En conjunto,estos gráficos evidencian que las campañas están teniendo mayor acogida en el grupo de ingresos más altos, por lo que enfocar las campañas a ellos otorgará mayores garantías de éxito.

Seguidamente, la Figura 3 presenta un gráfico de barras que ilustra la distribución de clientes que aceptan campañas según su nivel educativo.

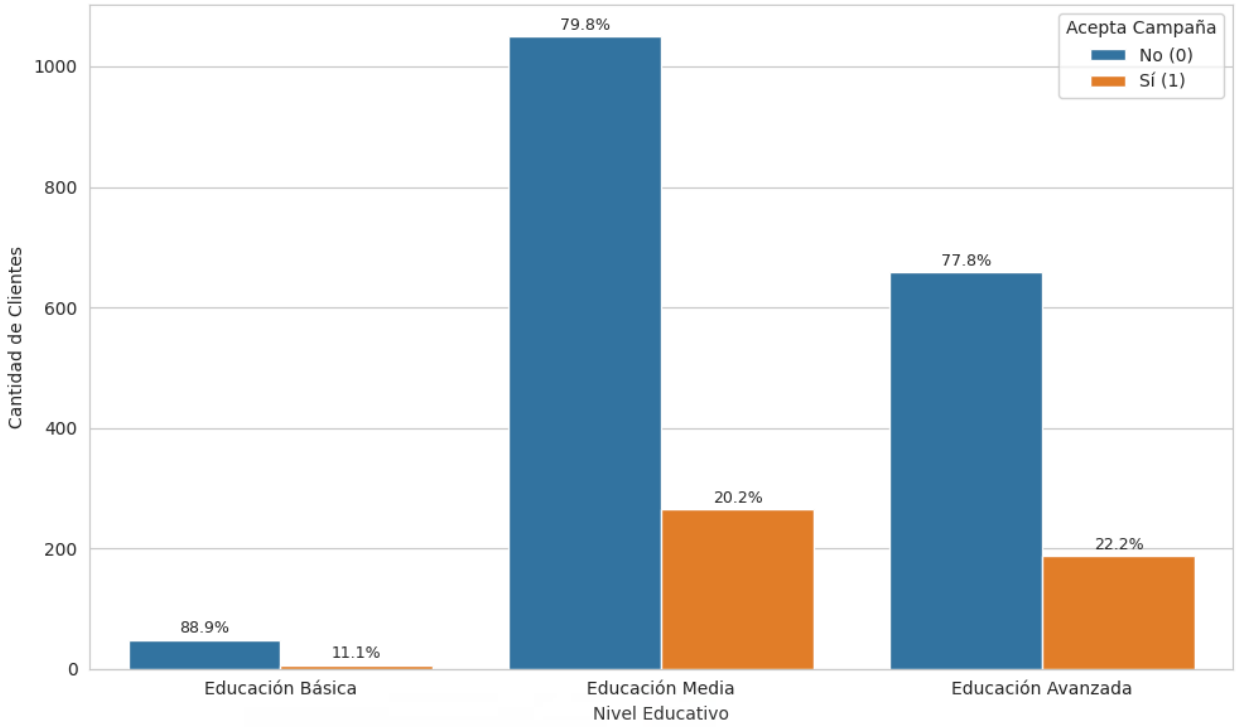

Figura 3: Número de clientes que acepta/no acepta alguna campaña en función del nivel educativo.

En la Figura 3 se observa que la mayoría de los clientes cuenta con un nivel de educación media, seguido por aquellos con educación avanzada, mientras que los clientes con educación básica representan una proporción muy reducida del total. En cuanto a la aceptación de campañas de marketing, los clientes con educación avanzada son los más sensibles a las campañas (un 22.2%), seguidos por los de educación media (con un 20.2%) y ya mucho más atrás los de educación básica (con un 11.1%).

Continuamos con la figura 4, que presenta la distribución de la variable objetivo en función del estado civil de los clientes.

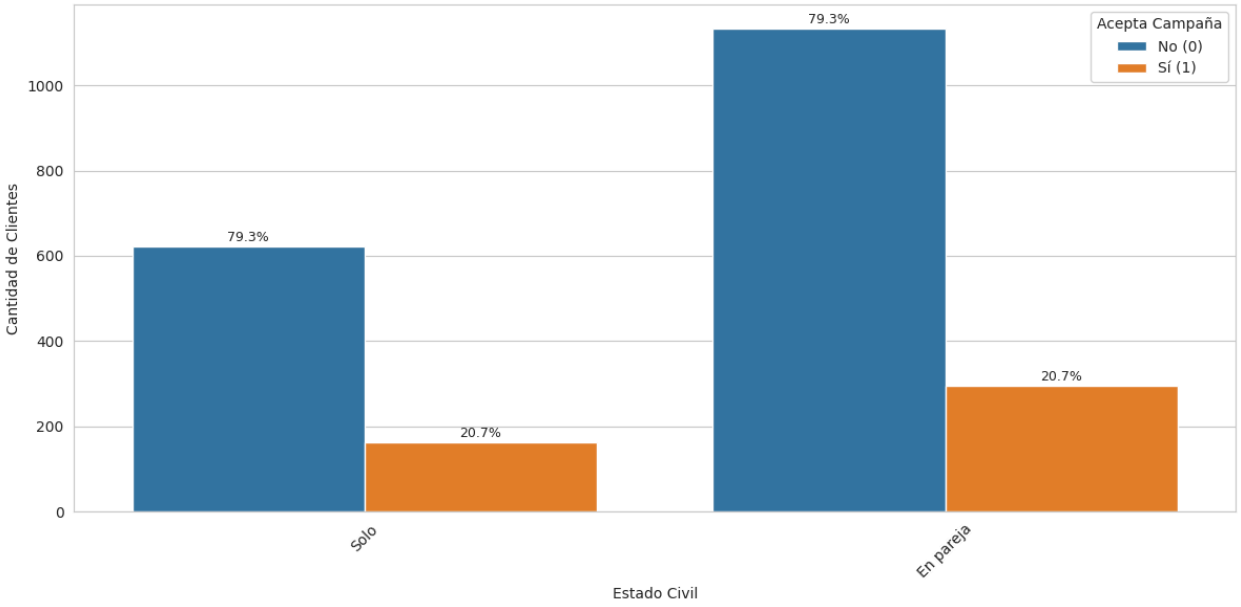

Figura 4: Número de clientes que acepta/no acepta alguna campaña en función del estado civil.

En la Figura 4 se observa que es mayoritario el grupo de los clientes en pareja, pero sin embargo la proporción de clientes sensibles a las campañas de marketing es similar para los que están solos que para los que están en pareja (20.7%). Esto quiere decir que el estado civil no parece tener relevancia para decidir las acciones de marketing.

Por otro lado, la Figura 5 muestra la relación entre la aceptación de campañas y la cantidad gastada por los clientes en cada tipo de producto.

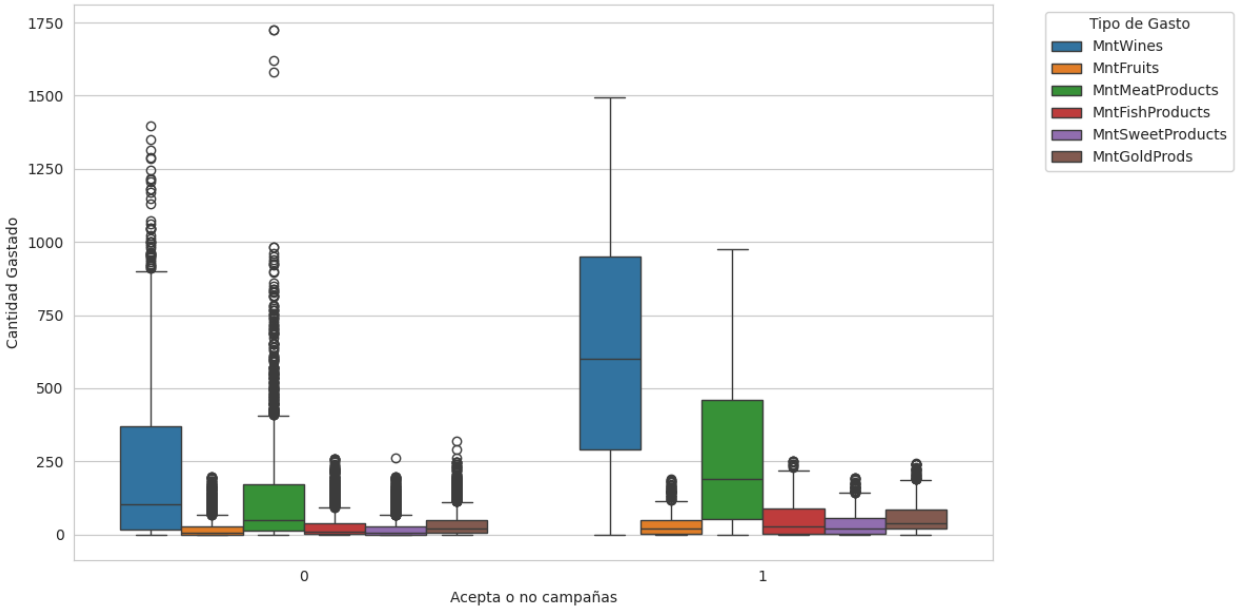

Figura 5: Dinero gastado por los clientes en cada una de las categorías de productos consideradas, en función de si acepta (1) o no (0) alguna campaña.

La Figura 5 nos muestra diferencias claras en los patrones de gasto entre los clientes que aceptan campañas de marketing y aquellos que no. Los clientes que aceptan las campañas (1) muestran un gasto considerablemente mayor y más constante en categorías como vinos (MntWines), productos cárnicos (MntMeatProducts) y frutas (MntFruits), con medianas notablemente superiores y menor presencia de valores atípicos. Esto sugiere un perfil de consumidor con mayor poder adquisitivo o gustos hacia productos de mayor valor.

Por otro lado, los clientes que no aceptan campañas (0) presentan menores niveles de gasto y una mayor dispersión en sus consumos, especialmente en vinos y carnes, evidenciando una mayor variabilidad en su comportamiento de compra. En categorías como pescado, dulces y oro, las diferencias entre ambos grupos son mínimas, indicando que las campañas no influyen de manera significativa o que el gasto en estos productos es bajo en general. En conjunto, el análisis sugiere que las campañas de marketing resultan más efectivas o atractivas para los clientes con hábitos de gasto altos y consistentes.

Continuamos con la figura 6, donde podemos ver la aceptación o no de campañas en función de la antigüedad del cliente.

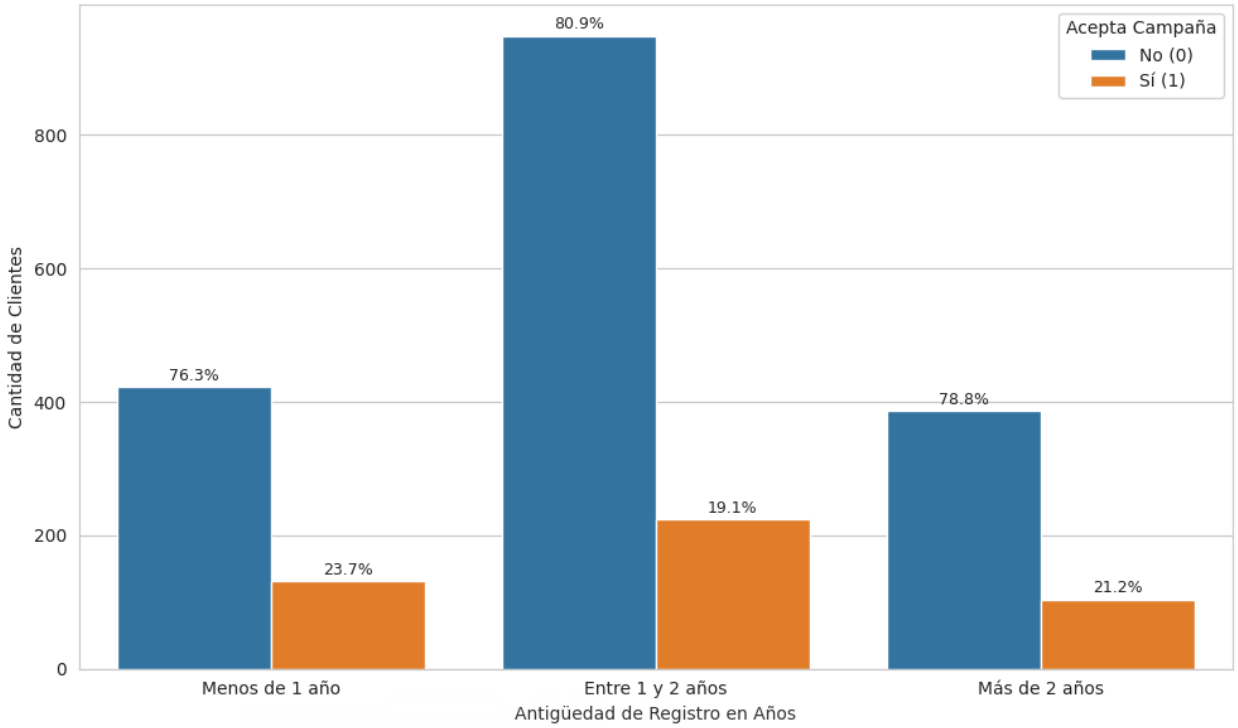

Figura 6: Cantidad de personas que aceptan/no aceptan las campañas de marketing en función de la antigüedad de su registro en la empresa (en años).

En la Figura 6 se aprecia que la mayor parte de las personas en la muestra son clientes con poca antigüedad en la empresa (menos de 2 años), especialmente entre 1 y 2 años. La figura presenta una mayor proporción de aceptación de campañas en los clientes que tienen menos de 1 año de antigüedad en la empresa (23,7%), seguido de los que tienen una antigüedad mayor a 2 años (21,2%). El grupo con una antigüedad entre 1 y 2 años años muestra el menor porcentaje de aceptación (19,1) respecto a los que no aceptan, lo que indica que la receptividad a las campañas no depende de la antigüedad del registro de los clientes en su empresa, sino a otros factores más determinantes.

Tras esta tenemos las figuras 7 y 8, que nos indican la aceptación o no de campañas de marketing en función de las visitas a la web que realizaron las personas el último mes.

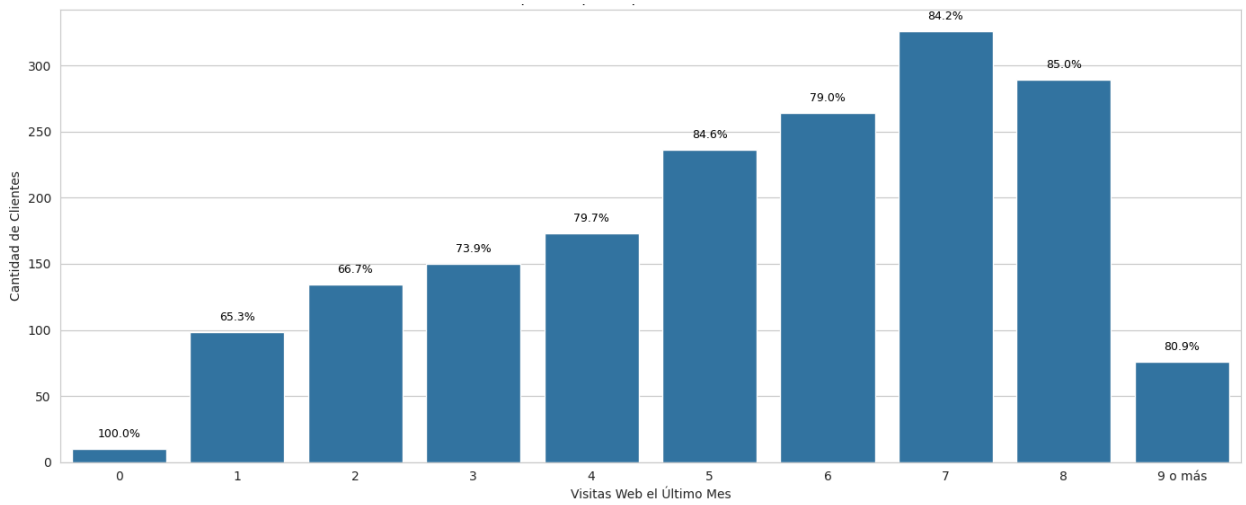

Figura 7: Cantidad de personas que no aceptan las campañas de marketing en función de las visitas que hicieron a la web el último mes.

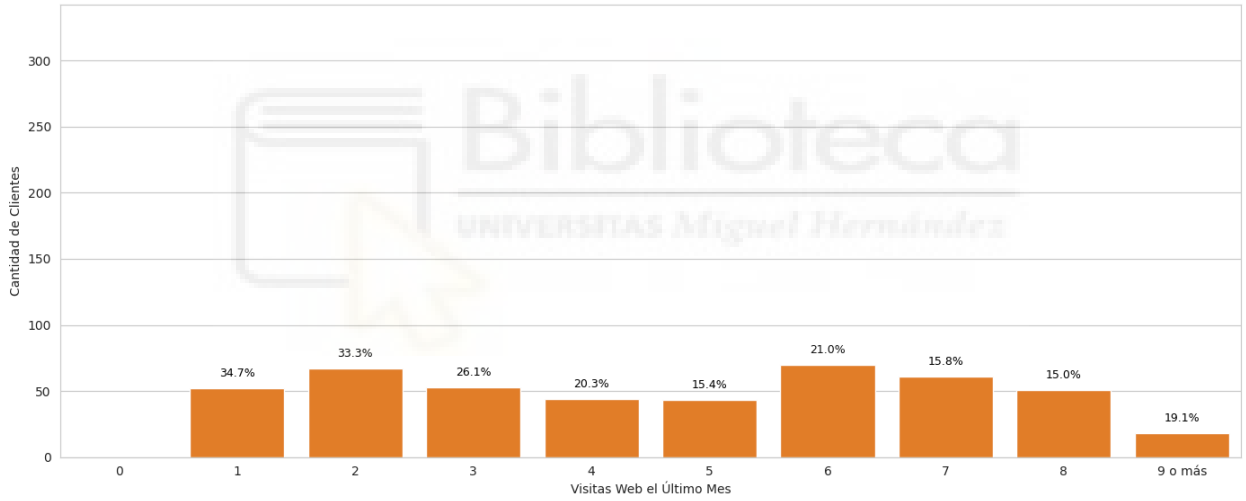

Figura 8: Cantidad de personas que aceptan las campañas de marketing en función de las visitas que hicieron a la web el último mes.

En las figuras 7 y 8 se observa una relación no lineal entre la frecuencia de visitas a la web y la aceptación de campañas. Los clientes con 0 visitas muestran un rechazo absoluto, mientras que aquellos con 1 visita presentan una ligera aceptación, aunque la mayoría sigue siendo reticente. En el rango de 2 a 6 visitas, la aceptación varía: los clientes con 2 o 3 visitas son relativamente más receptivos, mientras que la aceptación disminuye en quienes visitan la web 4 a 6 veces. Este segmento concentra un alto número de clientes que no aceptan campañas.

Entre los clientes con alta frecuencia de visitas (7 o más), el rechazo es mayor, superando el 80%. En conclusión, la aceptación de campañas depende del nivel de compromiso digital: los

clientes inactivos requieren estrategias de reactivación, los de frecuencia media muestran oportunidades mixtas, y los más activos pueden beneficiarse de campañas selectivas para evitar saturación.

Finalmente, investigamos el número de compras en diferentes plataformas, en función de si un cliente acepta o no alguna campaña publicitaria.

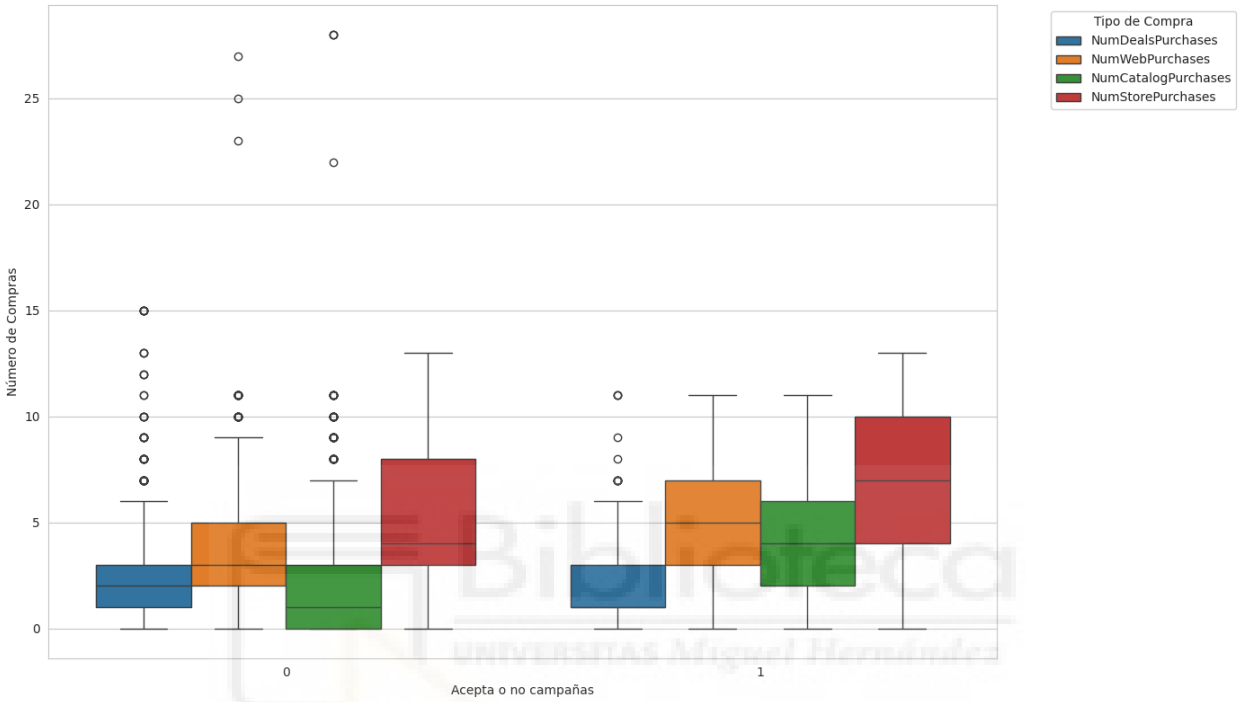

Figura 9: Boxplots que muestran la distribución de la variable aceptación o no de campañas por el cliente en función del número de compras realizadas de cada tipo.

Así, en la Figura 9 se aprecia que los clientes que aceptaron campañas presentan, en términos de valores intermedios como la mediana, un mayor número de compras totales, especialmente a través de los canales web y tienda física.

En conjunto, los análisis indican que la aceptación de campañas de marketing está fuertemente determinada por las características sociodemográficas y de comportamiento de los clientes. Aunque una proporción importante de clientes no responde a las campañas, se observa que aquellos con mayores ingresos y mayor nivel educativo presentan una mayor tasa de aceptación. Además, los clientes que aceptan campañas suelen gastar más, especialmente en categorías como vinos y carnes, y muestran un patrón de interacción moderada con el sitio web, especialmente aquellos que visitan la página entre una y tres veces al mes. También destacan por realizar compras tanto en tienda física como online con mayor frecuencia que el resto.

Estos hallazgos resaltan la relevancia de diseñar campañas de marketing personalizadas y segmentadas según el perfil y comportamiento de cada cliente. Esta estrategia permitirá maximizar la aceptación, aumentar el gasto promedio y mejorar la fidelización de los clientes.

7.2 Modelización y Clasificación.

En este apartado se presentan los resultados de la modelización de la aceptación de campañas por parte de los clientes mediante distintos métodos de clasificación. El objetivo es evaluar y comparar el desempeño de tres enfoques: regresión logística, Random Forest y Gradient Boosting, con base en métricas como precisión, sensibilidad, especificidad y área bajo la curva ROC.

7.2.1 Modelo logit

A continuación, en la Tabla 1 se presentan los principales resultados del modelo logístico, resumidos mediante las métricas de clasificación obtenidas sobre la muestra test.

Tabla 1: Métricas de clasificación del modelo logístico binario para la muestra test.

CLASIFICACIÓN	PRECISIÓN	SENSIBILIDAD	F1-SCORE	N
0 (No aceptan)	0.86	0.95	0.90	439
1 (Aceptan)	0.66	0.40	0.50	115
ACCURACY			0.83	554
MACRO AVG	0.76	0.67	0.70	554
WEIGHTED AVG	0.82	0.83	0.82	554

Los resultados indican un rendimiento global aceptable, con una exactitud del 83%, aunque el desempeño difiere entre clases. El modelo muestra mayor capacidad para identificar correctamente a los clientes que no aceptan la campaña (con un 95% de clasificación correcta según la sensibilidad y un f1-score del 90%), mientras que presenta mayor dificultad para clasificar correctamente a los que sí la aceptan (con una sensibilidad del 40% y un f1-score del 50%).

En conjunto, estos resultados sugieren que el modelo es más confiable para clasificar bien a los clientes que no aceptan campañas y nos da poca información sobre los que sí las aceptan.

La Figura 10 permite visualizar de manera más clara cómo se dan las clasificaciones incorrectas.

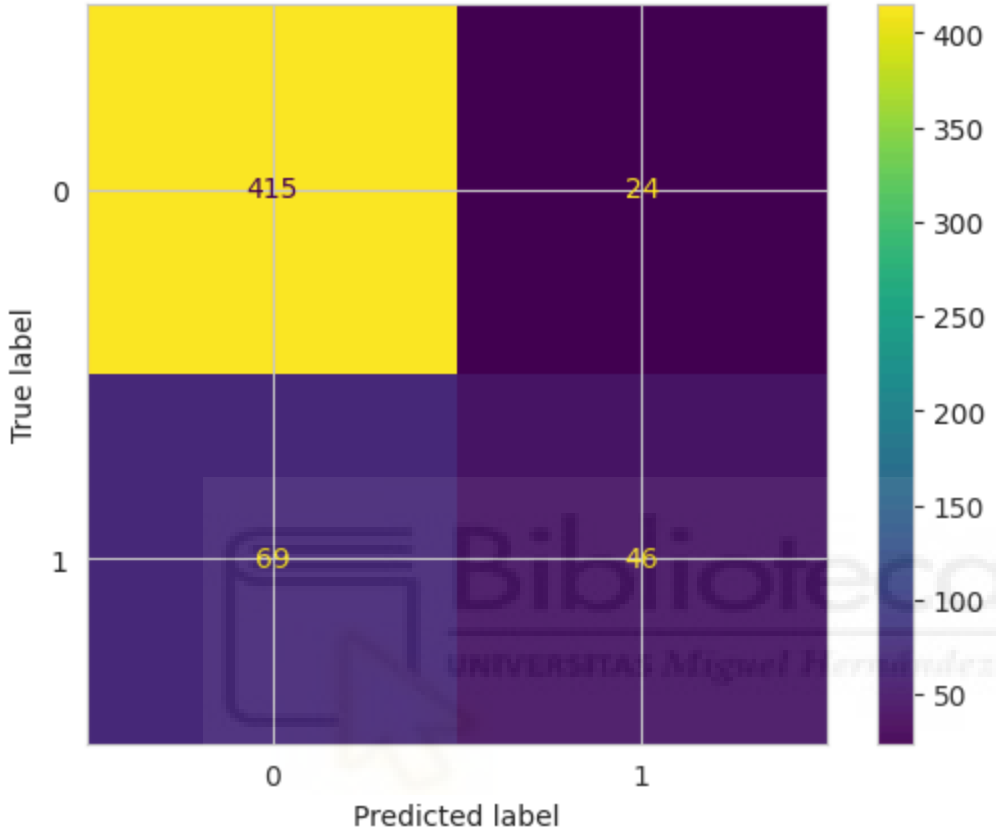

Figura 10: Matriz de confusión para el modelo logístico binario, que representa las predicciones que concuerdan con los datos y las que no.

La matriz de confusión muestra que el modelo identifica correctamente a 415 clientes que no aceptan la campaña y a 46 clientes que sí la aceptan. Sin embargo, predice incorrectamente que 24 clientes aceptarían la campaña cuando en realidad no lo hacen, y 69 clientes que sí aceptarían la campaña son clasificados como no participantes. Estos resultados confirman que el modelo es más preciso para predecir a los clientes que no aceptarán campañas que a los que sí lo harán, como nos indican las métricas presentadas en la Tabla 1.

A continuación, la gráfica 11 nos muestra la relación entre los verdaderos y los falsos positivos.

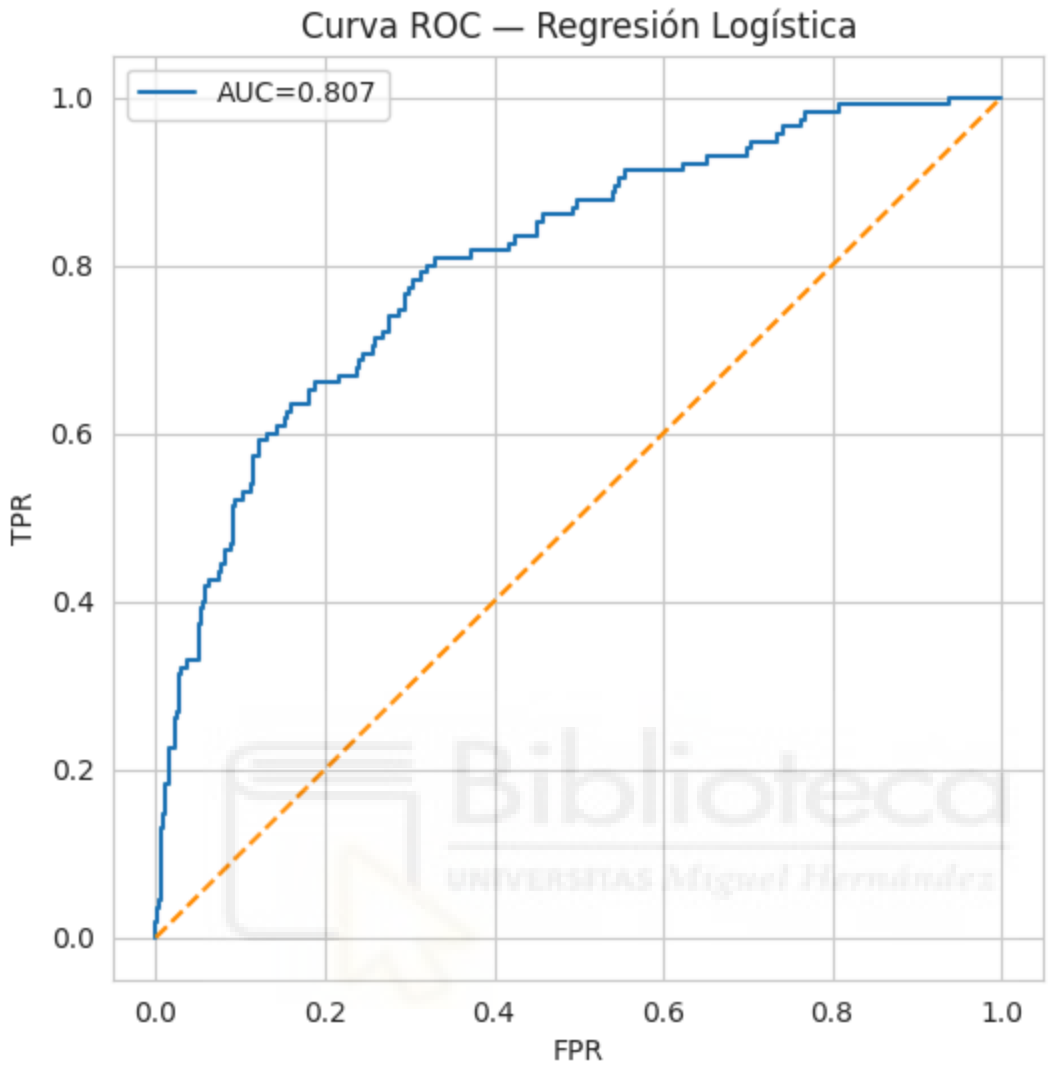

Figura 11 : Curva de ROC que representa la relación entre los verdaderos positivos (que aceptan campañas y son clasificados como tales) y los falsos positivos (que no aceptan campañas y son clasificados como que sí) del modelo logístico binario.

Un AUC de 0.807 indica que el modelo tiene una capacidad sólida para diferenciar entre clientes que aceptarán y los que no aceptarán las campañas. Este resultado complementa la exactitud global del 83%, si bien no captura las deficiencias de clasificación en el grupo de los clientes sensibles al marketing.

La Figura 12 muestra la curva de aprendizaje del modelo logístico.

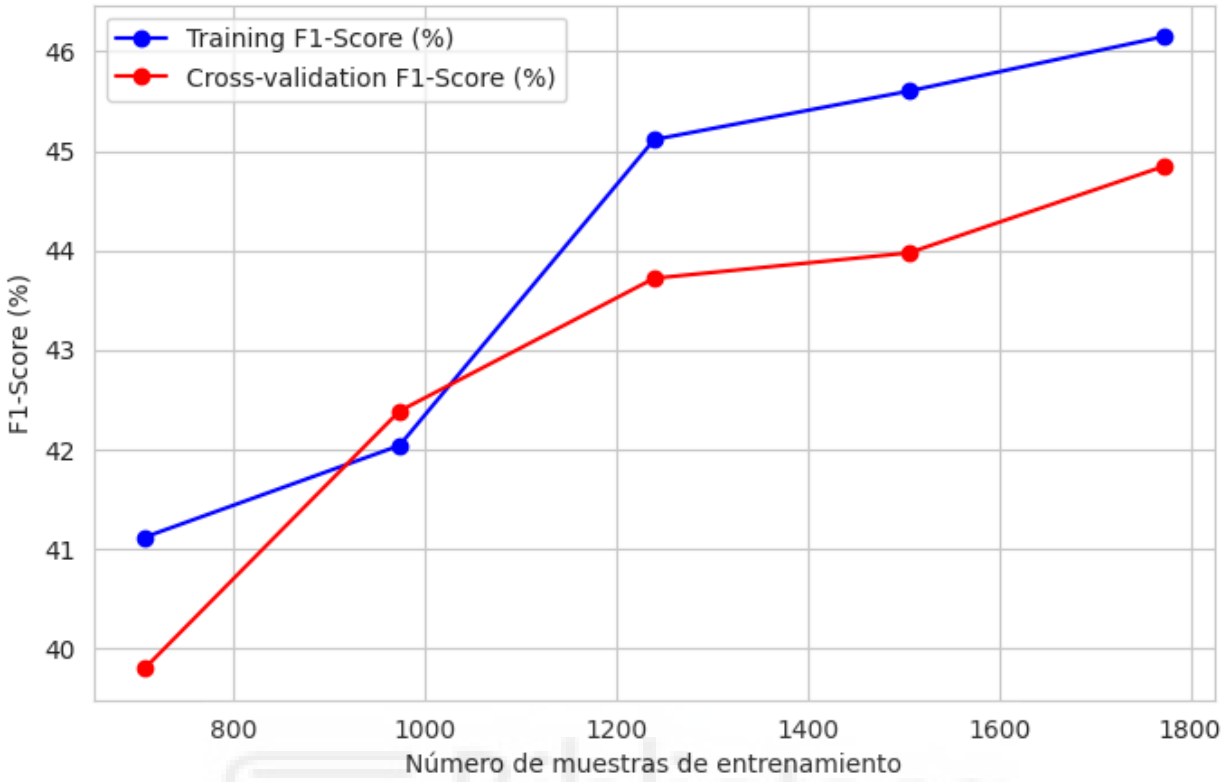

Figura 12: Curva de aprendizaje que representa cómo evoluciona el rendimiento del modelo (en términos del f1-score), a medida que aumenta el tamaño del conjunto de entrenamiento. El punto inicial entrena con el 40% de la muestra.

Atendiendo a la evolución del f1-score, observamos que el rendimiento del modelo aumenta con el tamaño de la muestra de entrenamiento, y a partir de muestras de entrenamiento con al menos el 70% de los datos (tercer punto en el gráfico, con algo más de 1200 datos) se estabiliza de modo razonable. Tenemos pues, que nuestro entrenamiento con el 75% de los datos es fiable.

La Figura 13 nos muestra visualmente cuáles son las variables que tienen mayor importancia en el modelo logístico ajustado, en función de la magnitud del coeficiente estimado en el modelo.

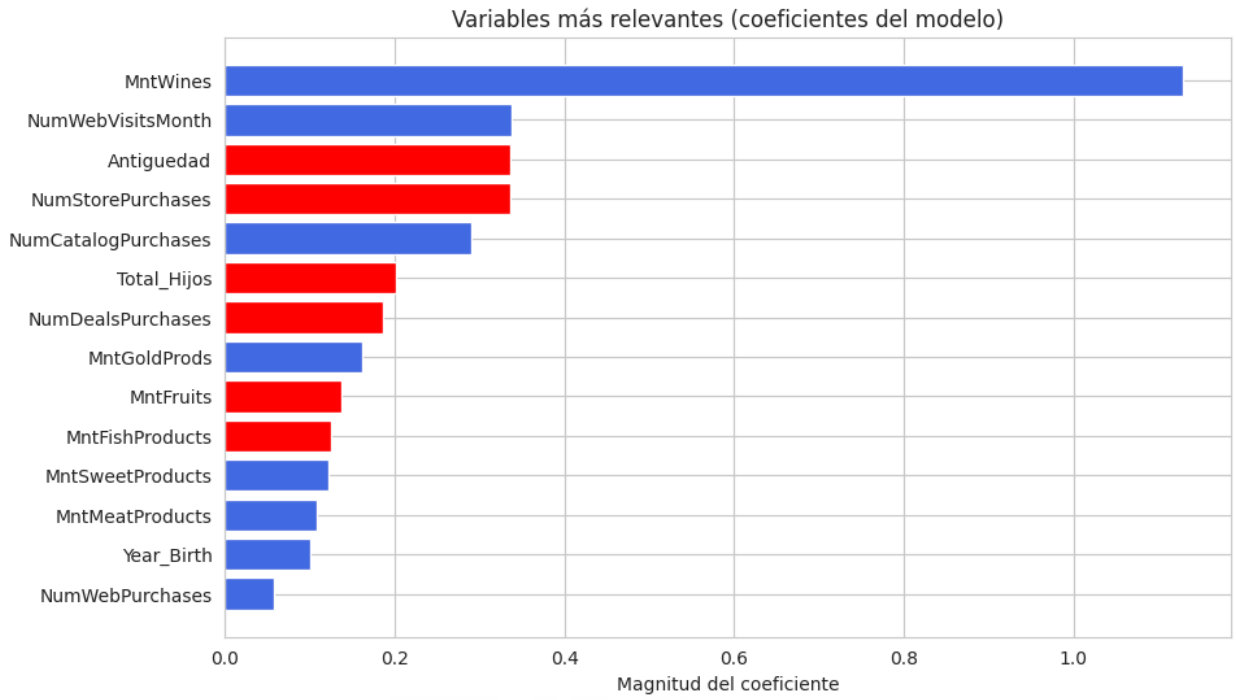

Figura 13 : Magnitud en valor absoluto del coeficiente estimado en el modelo logístico. En azul coeficientes positivos (variables que influyen positivamente en la aceptación) y en rojo coeficientes negativos (que influyen negativamente en la aceptación).

En el análisis de los coeficientes del modelo logístico, representados en la Figura 13, se observa que las variables con mayor influencia positiva en la aceptación de la oferta son principalmente el gasto en vinos (MntWines), el número de visitas mensuales a la web (NumWebVisitsMonth) y las compras a través de catálogo (NumCatalogPurchases). Estas variables, con coeficientes positivos, indican que a mayor gasto en vinos, mayor frecuencia de visitas a la web y mayor uso del canal de catálogo, aumenta la probabilidad de aceptación de las campañas de marketing.

Variables con mayor influencia negativa sobre la probabilidad de aceptar las campañas son la antigüedad del cliente (Antigüedad) y el número de compras en tienda física (NumStorePurchases): los clientes más antigüos y que hacen más compras en tienda física son los menos sensibles a las campañas de marketing. Otras variables con influencia media son el total de hijos (Total_Hijos, negativo), el número de compras en oferta (NumDealsPurchases, negativo) y el gasto en ciertos productos de alimentación como frutas o pescado (positivo).

En conjunto, estos resultados sugieren que los clientes más digitales y con un patrón de consumo orientado a productos premium como el vino muestran mayor propensión a aceptar la oferta, mientras que la antigüedad, las compras presenciales y las responsabilidades familiares se asocian a una menor probabilidad de aceptación. Este hallazgo refuerza la importancia de potenciar los canales de venta online y de catálogo para los segmentos más receptivos.

7.2.2 Random forest.

A diferencia del modelo de regresión logística, el Random Forest permite manejar interacciones no lineales entre las variables y reducir el riesgo de sobreajuste, mejorando la capacidad de predicción de la aceptación de las campañas.

Tabla 2: Métricas de evaluación del modelo random forest.

CLASIFICACIÓN	PRECISIÓN	RECALL	F1-SCORE	SUPPORT
0 (No aceptan)	0.92	0.72	0.81	439
1 (Aceptan)	0.41	0.75	0.53	115
ACCURACY			0.73	554
MACRO AVG	0.66	0.73	0.67	554
WEIGHTED AVG	0.81	0.73	0.75	554

Los resultados del modelo Random Forest muestran un rendimiento global aceptable, con una exactitud del 73%. Al igual que en el modelo anterior, el desempeño difiere entre las dos clases, aunque en este caso la tendencia se invierte en términos de Recall.

El modelo muestra mayor dificultad para predecir a los clientes que no aceptarán la campaña. Aun con una precisión del 92%, que indica predicciones negativas muy fiables, su Recall del 72% refleja cierta limitación para detectar correctamente a todos los clientes que no aceptan la oferta, obteniendo un F1-score de 0.81.

Por el contrario, se observa una mejor capacidad para identificar a los clientes que sí aceptarán la campaña. El modelo alcanza un Recall del 75%, lo que evidencia una mejor detección de la clase minoritaria, aunque con una precisión baja (41%), que implica un número elevado de falsos positivos. En conjunto, el modelo mejora la sensibilidad hacia los casos positivos, pero sacrifica precisión y reduce ligeramente la exactitud global.

Podemos comprobar el número de predicciones que concuerdan con los datos y las que no en la siguiente figura.

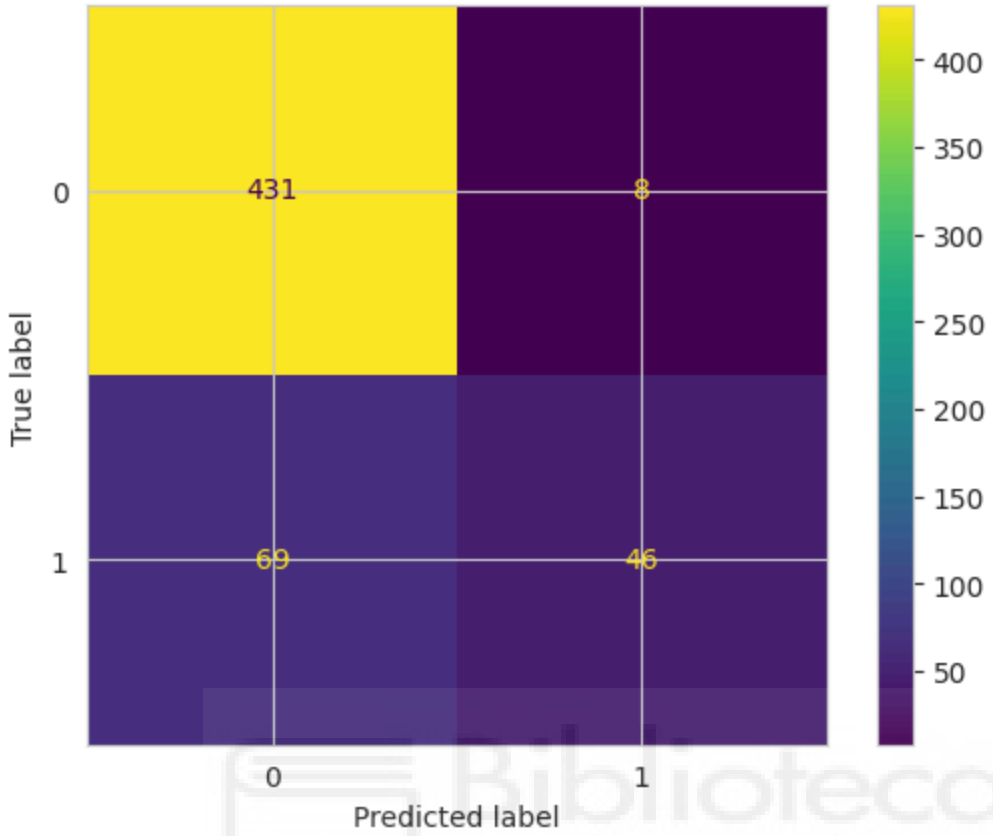

Figura 14: Matriz de confusión para el Random Forest, que representa las predicciones que concuerdan con los datos y las que no.

El modelo identifica correctamente a 431 clientes que no aceptarán la campaña y a 46 clientes que sí la aceptarán. No obstante, comete 69 Falsos Positivos, al predecir erróneamente que clientes no interesados aceptarían la campaña, y 8 Falsos Negativos, al clasificar como no receptivos a clientes que en realidad sí lo eran.

Estos resultados confirman la tendencia observada en las métricas: el bajo número de Falsos Negativos explica el alto Recall de la clase de aceptación, mientras que el elevado número de Falsos Positivos reduce significativamente su Precisión y afecta a la Exactitud global. En conjunto, el modelo prioriza la detección de clientes potencialmente receptivos, aunque a costa de generar numerosas falsas alarmas.

Seguidamente, en la figura 15 podemos observar la relación entre los verdaderos y los falsos positivos.

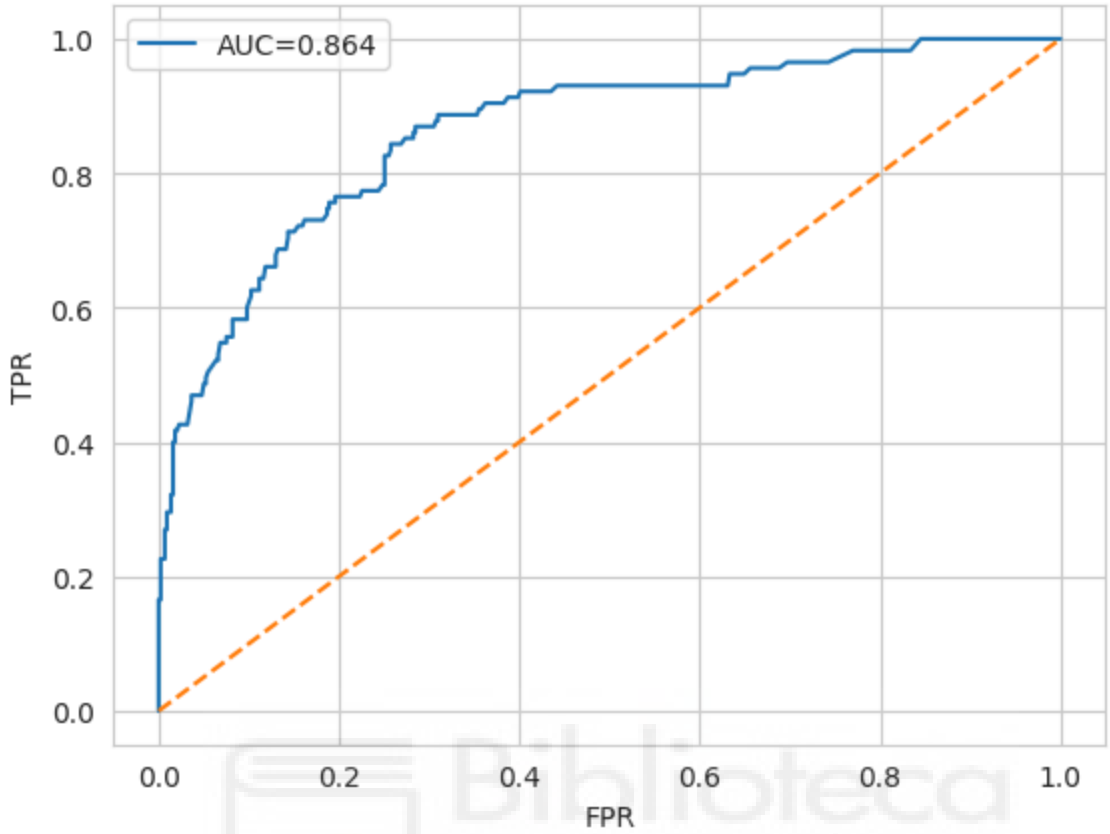

Figura 15 : Curva de ROC que representa la relación entre los verdaderos positivos (que aceptan campañas y son clasificados como tales) y los falsos positivos (que no aceptan campañas y son clasificados como que sí) del modelo Random Forest.

En la Figura 15 la línea azul, que representa su desempeño, se sitúa claramente por encima de la línea discontinua naranja, lo que confirma una capacidad de discriminación significativa entre los clientes que aceptarán la campaña y los que no.

El modelo alcanza un AUC de 0.864, lo que indica una buena capacidad predictiva y una probabilidad del 86,4% de clasificar correctamente a un cliente que acepte la campaña frente a uno que no lo haga.

Aunque su exactitud global es inferior a la de los otros modelos, el Random Forest demuestra un rendimiento estable y fiable al estimar las probabilidades de aceptación, manteniendo un equilibrio adecuado entre sensibilidad y especificidad.

Continuamos con la Figura 16, que muestra la curva de aprendizaje del modelo Random Forest.

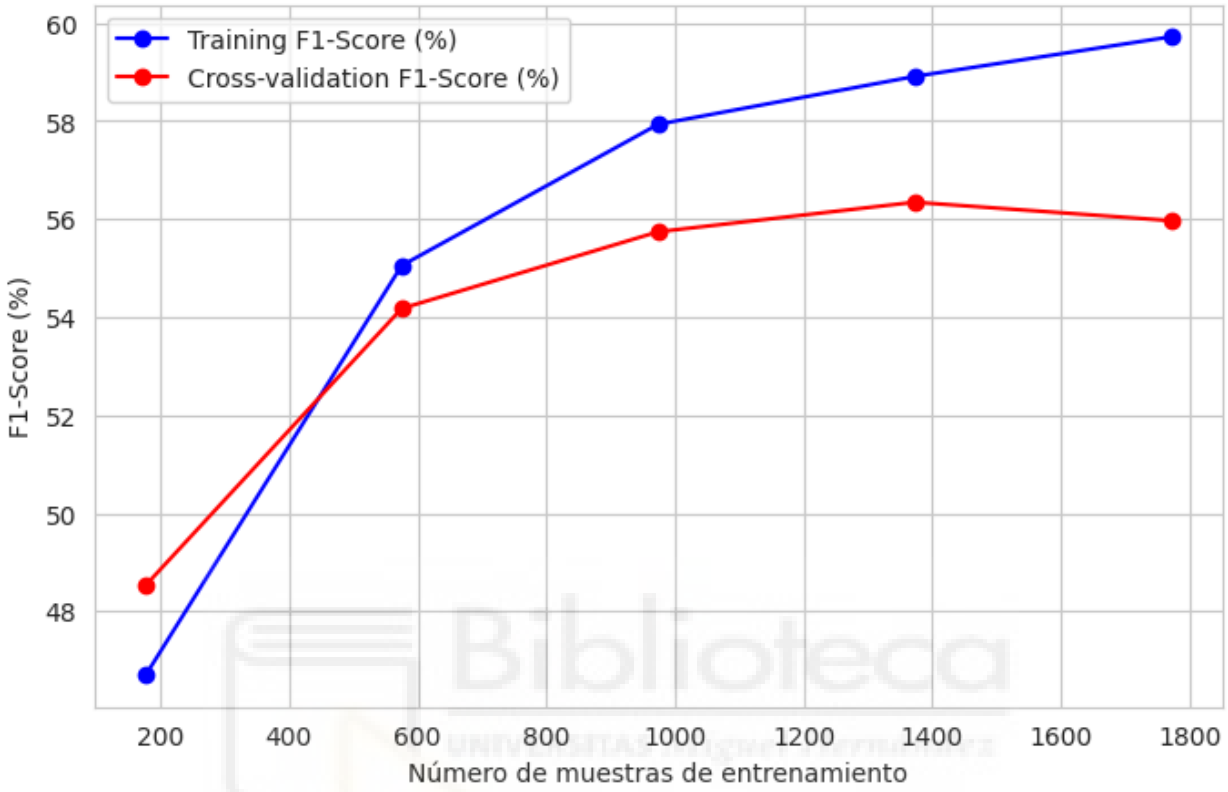

Figura 16: Curva de aprendizaje que representa cómo evoluciona el rendimiento del modelo (en términos del f1-score), a medida que aumenta el tamaño del conjunto de entrenamiento en el modelo Random Forest.

En la figura se observa que el F1-score de validación aumenta de forma pronunciada a medida que crece el tamaño de la muestra de entrenamiento, partiendo de valores bajos hasta estabilizarse alrededor en las 1000 muestras. La brecha entre ambas curvas es reducida, lo que indica que el modelo no presenta sobreajuste severo y generaliza de forma adecuada.

Aunque el rendimiento se mantiene en un nivel moderado, la tendencia sugiere que utilizar el conjunto completo de datos contribuye a alcanzar el máximo rendimiento posible del modelo, reflejando un equilibrio razonable entre sesgo y varianza.

La siguiente figura nos muestra visualmente cuáles son las variables que tienen mayor importancia en el modelo.

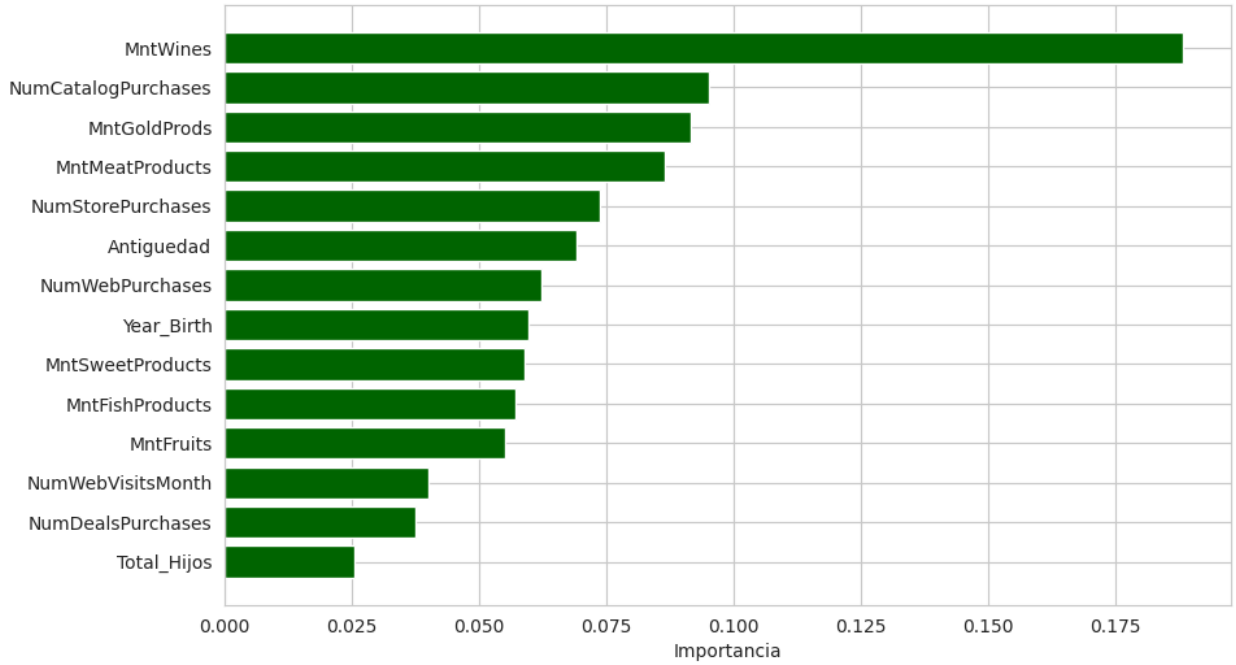

Figura 17 : Contribución relativa de las variables en el modelo Random Forest.

La Figura 17 refleja la contribución relativa de cada característica en la predicción de la aceptación de campañas. A diferencia de la regresión logística, esta métrica siempre toma valores positivos y representa la potencia predictiva de una variable, no la dirección de su efecto.

Las variables con mayor influencia son principalmente las relacionadas con el gasto y el comportamiento de compra. Destaca MntWines (Gasto en Vinos) como la más determinante, con una importancia que duplica a la siguiente variable. Le siguen NumCatalogPurchases (Compras por Catálogo), NumGoldProducts (Gasto en productos de oro) , MntMeatProducts (Gasto en Carnes) y NumStorePurchases (Compras en Tienda Física).

En cambio, Total_Hijos (total de hijos en casa) y NumDealsPurchases (Compras en Ofertas) muestran escasa relevancia. En conjunto, el modelo confirma que los patrones de consumo y los canales de compra son los factores más decisivos en la predicción, apoyando la idea de que el modelo se basa principalmente en dónde y cuánto gasta el cliente.

7.2.3 Gradient boosting

Para continuar con el análisis, se procede a aplicar el método de Gradient Boosting como uno de los modelos de predicción seleccionados. Este algoritmo permitirá evaluar su

rendimiento frente a los modelos anteriores y determinar su capacidad para predecir la aceptación de las campañas.

Tabla 3: Métricas de evaluación del modelo gradient boosting.

CLASIFICACIÓN	PRECISION	RECALL	F1-SCORE	SUPPORT
0	0.87	0.94	0.90	439
1	0.68	0.46	0.55	115
ACCURACY			0.84	554
MACRO AVG	0.77	0.70	0.73	554
WEIGHTED AVG	0.83	0.84	0.83	554

El modelo Gradient Boosting presenta un rendimiento global sólido, alcanzando una exactitud del 84%. Sin embargo, la capacidad predictiva se distribuye de manera desigual entre las clases, siguiendo un patrón similar al del modelo logístico binario y favoreciendo la predicción de la clase mayoritaria.

El modelo muestra gran eficacia al identificar a los clientes que no aceptarán la campaña, con una precisión del 87% y un recall del 94%, lo que indica una detección muy fiable y pocos errores de clasificación. En consecuencia, obtiene un f1-score elevado (0.90) para esta clase.

Por el contrario, presenta mayor dificultad en la predicción de los clientes que sí aceptarán la campaña, con una precisión del 68% y un recall del 46%, lo que limita su capacidad para identificar correctamente estos casos. En conjunto, el modelo mantiene un excelente desempeño en la clase de no aceptación, pero su sensibilidad hacia la clase minoritaria sigue siendo reducida, pese a la alta exactitud global alcanzada.

En la siguiente figura encontramos las predicciones del modelo que concuerdan con los datos y las que no.

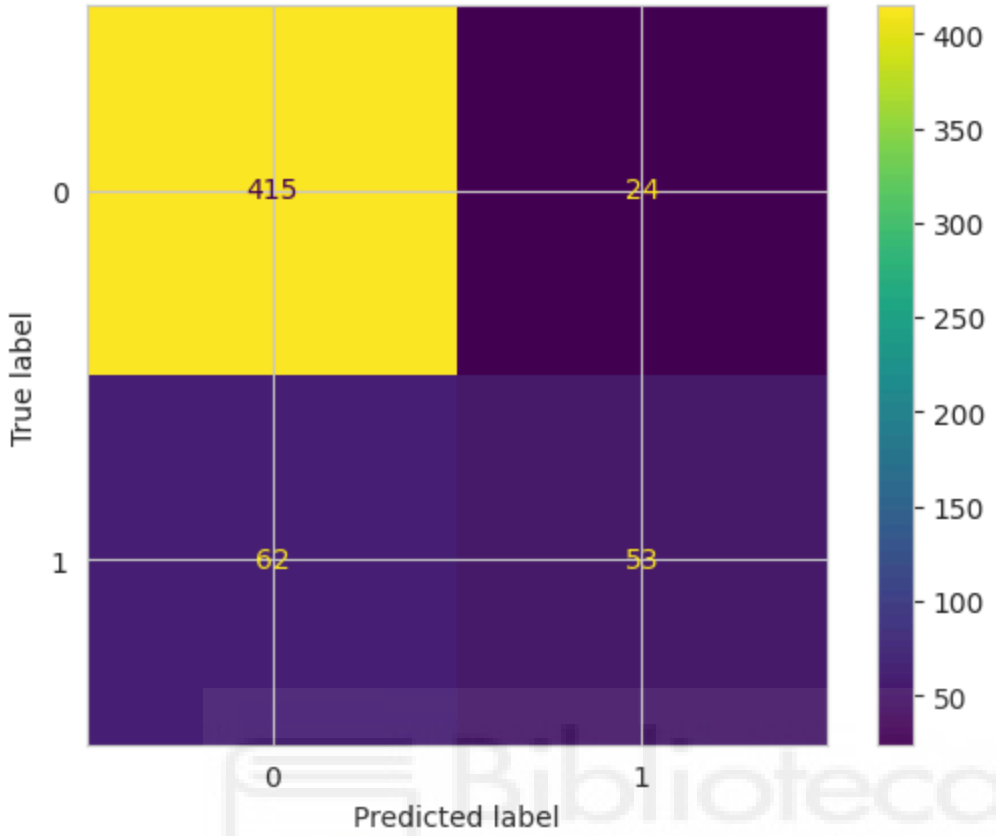

Figura 18: Matriz de confusión para el Gradient Boosting, que representa las predicciones que concuerdan con los datos y las que no.

El modelo identifica correctamente a 415 clientes que no aceptarán la campaña y a 53 que sí la aceptarán. En cuanto a los errores, produce únicamente 24 Falsos Positivos, lo que indica una excelente precisión en la clase de no aceptación, pero presenta 62 Falsos Negativos, lo que revela dificultades para reconocer a todos los clientes realmente interesados.

Estos resultados reflejan que el Gradient Boosting es el modelo más fiable para predecir la no aceptación, con un control sobresaliente de los errores en la clase mayoritaria. Sin embargo, el alto número de Falsos Negativos confirma su limitada capacidad de Recal para detectar a los clientes que sí aceptarían la campaña, lo que representa su principal área de mejora.

Seguidamente, la figura 19 presenta la relación entre los verdaderos y falsos positivos.

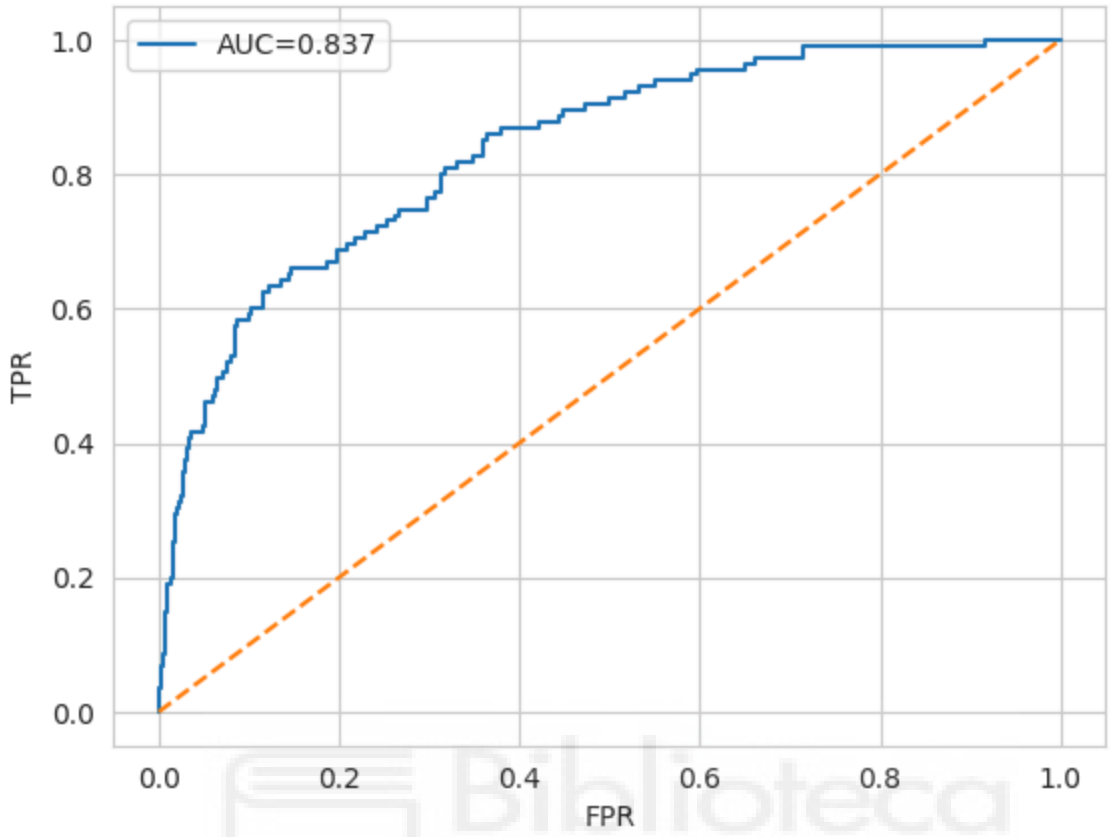

Figura 19 : Curva de ROC que representa la relación entre los verdaderos positivos (que aceptan campañas y son clasificados como tales) y los falsos positivos (que no aceptan campañas y son clasificados como que sí) del modelo Gradient Boosting.

La curva ROC del modelo Gradient Boosting muestra una clara capacidad discriminativa entre los clientes que aceptarán y los que no. La línea azul, correspondiente al modelo, se sitúa considerablemente por encima de la línea discontinua naranja que representa el azar, lo que demuestra una notable mejora predictiva.

El AUC de 0.837 es el más alto de los tres modelos, lo que indica que el Gradient Boosting es el más eficaz para ordenar y diferenciar las probabilidades de respuesta de los clientes. Este valor, superior al 80%, confirma su alto poder de discriminación.

Además, este resultado es coherente con su mayor exactitud global, reforzando la idea de que el Gradient Boosting ofrece las predicciones más precisas y fiables a la hora de identificar a los clientes con mayor probabilidad de aceptar la campaña.

Continuamos con la Figura 20, que muestra la curva de aprendizaje del modelo Gradient Boosting.

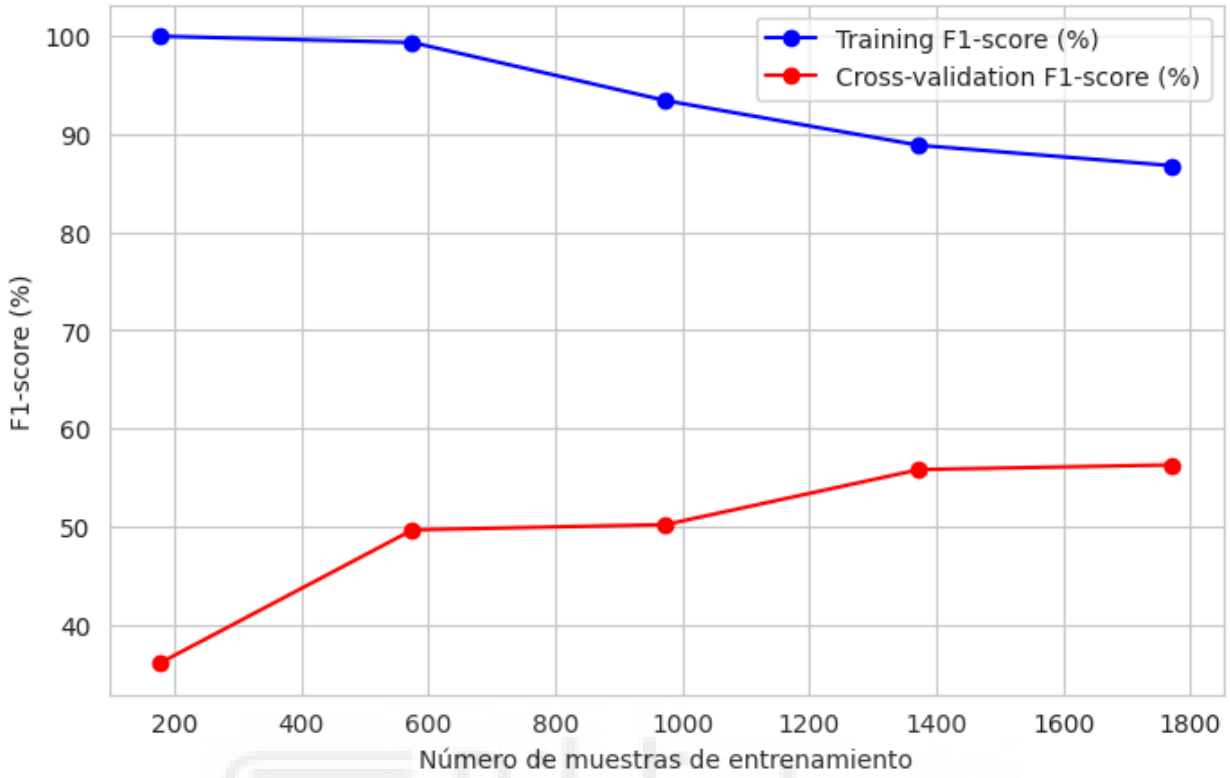

Figura 20: Curva de aprendizaje que representa cómo evoluciona el rendimiento del modelo (en términos del f1-score), a medida que aumenta el tamaño del conjunto de entrenamiento en el modelo Gradient Boosting.

La curva de entrenamiento comienza muy alta y desciende ligeramente conforme aumenta la muestra, mientras que la curva de validación cruzada parte de valores bajos y se incrementa progresivamente hasta estabilizarse en torno a las 1400 muestras.

La amplia distancia entre ambas curvas evidencia un claro sobreajuste: el modelo aprende en exceso los patrones del conjunto de entrenamiento, pero no generaliza adecuadamente a datos nuevos. Pese a ello, la tendencia ascendente de la curva de validación indica que el modelo mejora con un mayor volumen de datos.

El gráfico que encontramos a continuación nos muestra visualmente cuáles son las variables que tienen mayor importancia en el modelo.

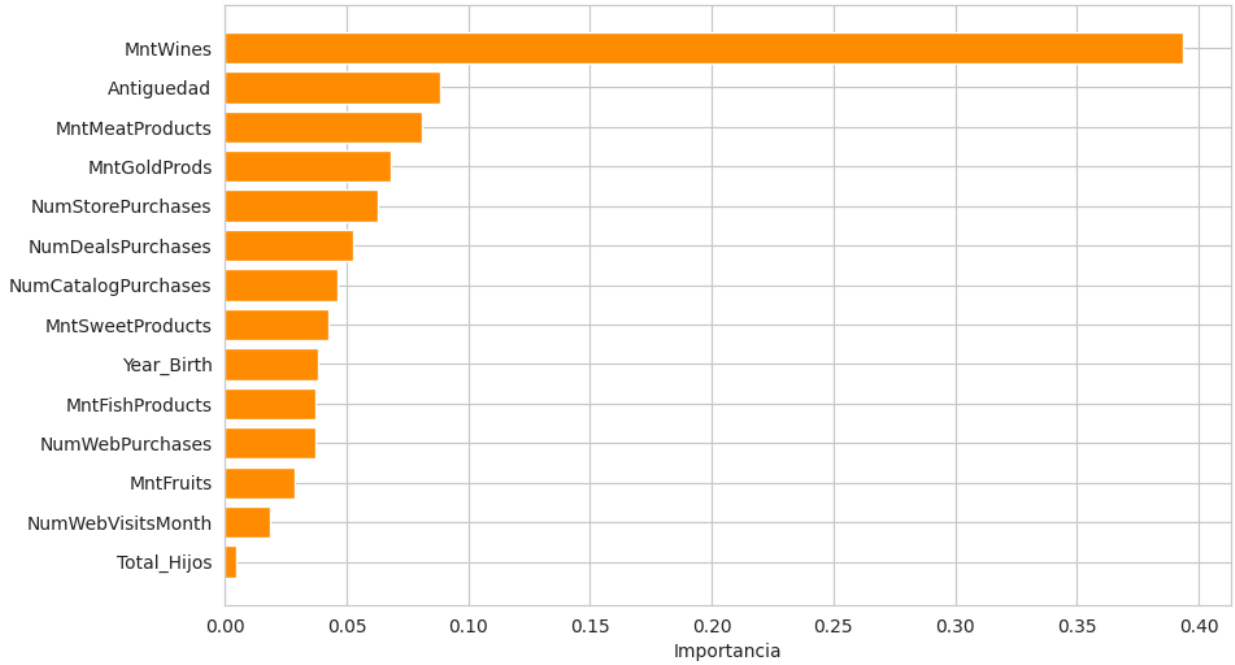

Figura 21 : Importancia de las variables en el modelo Gradient Boosting.

La Figura 21 está medida también por la reducción total de impureza a lo largo de los árboles que conforman el modelo. La jerarquía de importancia mantiene un patrón similar al observado en el Random Forest, aunque con ligeras variaciones en el orden y el peso relativo de las variables.

MntWines (Gasto en Vinos) vuelve a ser la variable dominante, con un peso cercano al 40% del total, lo que resalta su papel clave en la predicción de la aceptación. Le siguen Antiguedad, MntMeatProducts (Gasto en Carnes), MntGoldProds (Gasto en Productos de Oro) y NumStorePurchases (Compras en Tienda Física)

En contraste, Total_Hijos se mantiene como la variable con menor influencia. En conjunto, el análisis reafirma que las variables vinculadas al gasto en productos premium y los canales de compra son las que más contribuyen al poder predictivo de los modelos basados en árboles.

7.2.4 Comparaciones y conclusiones.

Se presentan las comparaciones entre los distintos modelos aplicados con el fin de evaluar su desempeño y determinar cuál ofrece las mejores capacidades predictivas.

Tabla 4: Comparación de métricas principales entre los métodos de clasificación logístico, Random Forest y Gradient Boosting.

MÉTRICAS	REGRESIÓN LOGÍSTICA	RANDOM FOREST	GRADIENT BOOSTING
F1-SCORE (1- ACEPTAN)	0.50	0.53	0.55
F1-SCORE (0- NO ACEPTAN)	0.90	0.81	0.90
F1-SCORE (AVG)	0.82	0.75	0.83
EXACTITUD (ACCURACY)	0.83	0.73	0.84

Al comparar los tres modelos de clasificación entrenados, Regresión Logística Binaria, Random Forest y Gradient Boosting, podemos valorar conjuntamente su capacidad predictiva y su utilidad práctica en la predicción de la aceptación de campañas de marketing. Cada enfoque ofrece ventajas diferenciadas en términos de exactitud, equilibrio entre clases y complejidad interpretativa, lo que contribuye a una comprensión más completa de cómo cada modelo se ajusta al comportamiento de los clientes ante las campañas promocionales. En este análisis, los indicadores clave de rendimiento (F1-score ponderado y exactitud general) evidencian diferencias notables entre los modelos.

En primer lugar, el modelo de Regresión Logística alcanzó una exactitud del 83 % y un F1-score promedio ponderado de 0.82, lo que demuestra un desempeño alto y robusto. Sin embargo, al observar los resultados por clase, se aprecia una desigualdad relevante: el F1-score para la clase positiva fue de apenas 0.50, mientras que para la clase negativa alcanzó 0.90. Esto sugiere que el modelo tiende a predecir mucho mejor a quienes no aceptan las campañas, mostrando dificultades para identificar correctamente a los clientes receptivos. No obstante, la Regresión Logística nos ayuda a identificar los factores más influyentes como el gasto en vinos o carnes, la frecuencia de compras online o el nivel de ingresos y sienta las bases conceptuales para modelos más avanzados.

Por otro lado, el modelo Random Forest obtuvo una exactitud del 73 % y un F1-score promedio ponderado de 0.75. Si bien su exactitud general es la más baja de los tres, destaca por tener un F1-score de 0.53 para la clase positiva, el cual es ligeramente mejor que el de la Regresión Logística, aunque menor que el del Gradient Boosting. Su rendimiento en la clase negativa fue de F1-score = 0.81. Esto indica que el modelo Random Forest tiene la peor capacidad de clasificación general (medida por exactitud y F1-score ponderado) de los tres modelos. El análisis de importancia de variables confirmó los hallazgos obtenidos con la regresión logística,

reforzando la influencia de los ingresos, la frecuencia de visitas web, el gasto en productos de alto coste (como vinos y carnes) y la actividad en el canal online. En términos prácticos, su rendimiento global es menos óptimo.

El modelo Gradient Boosting, por su parte, logró la mejor exactitud y un F1-score promedio ponderado de 0.83, mostrando el rendimiento más competitivo y equilibrado. Presentó el F1-score más alto en la clase positiva con 0.55 y un 0.90 en la clase negativa. Esto revela el mejor equilibrio entre la capacidad de detección de la clase minoritaria y la exactitud global. Su ventaja principal es su habilidad para optimizar gradualmente los errores residuales de predicciones anteriores, lo que se traduce en una mayor estabilidad y el mejor rendimiento global. Este modelo demuestra ser el más efectivo en términos predictivos.

En conjunto, los resultados comparativos evidencian que los tres modelos cumplen adecuadamente con el objetivo de clasificar a los clientes, aunque difieren en su grado de precisión e interpretabilidad. La Regresión Logística destaca por su transparencia y su valor explicativo. Gradient Boosting se consolida como el modelo más preciso y efectivo en términos de rendimiento global. Random Forest, aunque sigue siendo útil, mostró la exactitud más baja.

Considerando estos resultados, el Gradient Boosting destaca como el modelo más adecuado para la implementación práctica, ya que combina la mejor capacidad predictiva con una alta exactitud, constituyéndose en una herramienta confiable para la segmentación de clientes, la personalización de estrategias promocionales y la planificación de futuras campañas de marketing basadas en datos.

8. Conclusiones

Nuestro estudio ha permitido analizar de forma exhaustiva los factores que influyen en la aceptación de campañas de marketing, desarrollando modelos predictivos capaces de anticipar el comportamiento de los clientes. A continuación, se revisan los objetivos planteados y las conclusiones alcanzadas en relación con cada uno de ellos:

Objetivo 1. *Describir y comparar el comportamiento de los clientes en la aceptación de las cinco campañas de marketing implementadas por la empresa.*
Este objetivo se logró mediante un análisis exploratorio que permitió identificar patrones de aceptación vinculados a características sociodemográficas y hábitos de consumo. Los resultados muestran que la aceptación de las campañas es mayor entre los clientes con ingresos altos y mayor nivel educativo, así como entre aquellos con actividad web moderada (1 a 3 visitas al mes). Además, los clientes que aceptan campañas tienden a realizar más compras, tanto en tienda física como online, especialmente en categorías como vinos y carnes. Por el contrario, los clientes con ingresos bajos presentan menor aceptación y quienes tienen una frecuencia de visitas a la web muy elevada muestran tasas de rechazo superiores. Estos

hallazgos subrayan la importancia de segmentar las campañas según ingresos, nivel educativo y comportamiento digital para optimizar la respuesta de los clientes.

Objetivo 2. *Desarrollar modelos de clasificación válidos para diferenciar a los clientes en función de su comportamiento en la aceptación de campañas.*
Este objetivo se alcanzó mediante la implementación de tres modelos predictivos: Regresión Logística, Random Forest y Gradient Boosting. Los resultados obtenidos evidenciaron un buen nivel de rendimiento y capacidad discriminativa, con valores de F1-score promedio ponderado entre 0.75 y 0.83 y exactitudes que oscilaron entre el 73 % y el 84 %, dependiendo del modelo aplicado. Estos indicadores confirman la validez y eficacia de los métodos desarrollados, destacando al Gradient Boosting como el modelo con mejor desempeño global, al presentar la exactitud más alta y el F1-score más alto para la clase minoritaria.

Objetivo 3. *Identificar las variables relevantes en la clasificación y caracterizar los diferentes tipos de clientes según su respuesta.*
Los tres modelos coinciden en que el gasto en vinos, seguido de carnes y productos de oro, es el factor más determinante para predecir la aceptación de campañas, aunque existen diferencias en la jerarquía de otras variables según el método: regresión logística destaca visitas web, compras en catálogo y dulces; Random Forest resalta compras en tienda y antigüedad; y Gradient Boosting combina varios factores de compra y algunos sociodemográficos. En todos los modelos, variables como número total de hijos, nivel educativo, frutas y pescado resultan menos determinantes para la predicción. Estas diferencias reflejan que cada método captura distintos aspectos del comportamiento del cliente, desde relaciones lineales hasta interacciones complejas, haciendo que los hallazgos derivados de los modelos sean más exhaustivos y precisos que los obtenidos únicamente con el análisis exploratorio. En conjunto, se puede caracterizar a los clientes más receptivos como aquellos con mayor gasto en productos premium y hábitos activos de compra online y en tienda, proporcionando una base sólida para diseñar campañas segmentadas y efectivas.

Objetivo 4. *Ajustar y comparar varios modelos de clasificación para encontrar el más preciso en la diferenciación de clientes.*
La comparación evidenció que el modelo Gradient Boosting presentó el mejor rendimiento global, con 84 % de exactitud y un F1-score promedio ponderado de 0.83, superando a la regresión logística y al Random Forest. Por tanto, este modelo se considera el más adecuado para la predicción de la aceptación de campañas, equilibrando precisión, sensibilidad y estabilidad. Su implementación práctica permitiría optimizar recursos de marketing al enfocar las campañas en los clientes con mayor probabilidad de conversión.

El trabajo cumple plenamente el objetivo global de desarrollar un modelo predictivo que optimice la asignación de recursos en campañas de marketing, incrementando la tasa de respuesta y reduciendo los costes operativos. Los hallazgos confirman que la personalización

de las estrategias en función del perfil del cliente es esencial para mejorar la eficacia de las campañas.

Referencias

eMarketer. (2023). *Digital ad spend worldwide to pass $600 billion this year.* https://www.emarketer.com/content/digital-ad-spend-worldwide-pass-600-billion-this-year

Marketers by Adlatina. (2025). *Amazon invirtió 21 mil millones de dólares en publicidad y promoción en 2024: 679 dólares por segundo*. Marketers by Adlatina. https://www.marketersbyadlatina.com/articulo/13089_amazon-invirtio-21-mil-millones-de-dolares-en-publicidad-y-promocion-en-2024-679-dolares-por-segundo

Ekos. (2024). *Coca-Cola: Su nueva estrategia global de marketing con la que cerrará el 2024*. Ekos Negocios. https://ekosnegocios.com/articulo/coca-cola-su-nueva-estrategia-global-de-marketing-con-la-que-cerrara-el-2024

Red Bull. (2012). *Red Bull Stratos.* https://www.redbull.com/es-es/projects/red-bull-stratos

Marketing Ideas 101. (2024). *The genius of Red Bull's Stratos jump: Extreme marketing at 128,000 feet*. https://marketingideas101.com/marketing/the-genius-of-red-bulls-stratos-jump-extreme-marketing-at-128000-feet/

iArtificial. (2025). *¿Por qué las recomendaciones de Netflix y Spotify son tan precisas?* AI Blog. https://iartificial.blog/aprendizaje/por-que-las-recomendaciones-de-netflix-y-spotify-son-tan-precisas/

Javanmardi, S., Javanmardi, E., & Bucci, A. (2025). Quantifying drivers of virtual reality acceptance in tourism planning using a grey system theory-based approach. *Journal of Hospitality and Tourism Insights, 8*(4), 1308–1327. https://doi.org/10.1108/JHTI-04-2024-0367

Venkateswaran, P. S., & Ngo, N.-Q.-N. (2025). Quantitative Analysis of AI-Driven Predictive Analytics in Digital Marketing. En *Strategic Blueprints for AI-Driven Marketing in the Digital Era* (pp. 221–252). IGI Global. https://doi.org/10.4018/979-8-3373-3897-2.ch007

Usman, S. M., Khalid, S., Tanveer, A., Imran, A. S., & Zubair, M. (2024). *Multimodal consumer choice prediction using EEG signals and eye tracking*. *Frontiers in Computational Neuroscience, 18*, Article 1516440. Frontiers | Multimodal consumer choice prediction using EEG signals and eye tracking

Othman, A. H. A., Alkhafaji, M. A., Hussam, R., Haroon, N. H., & Majeed, M. G. (2024). Predicting Consumer Behaviour and Results Using Social Media and Deep Learning. En *2023*

Annual International Conference on Emerging Research Areas: International Conference on Intelligent Systems (AICERA/ICIS). IEEE. https://ieeexplore.ieee.org/document/10420394/

Ullah, A., Baloch, G., Ali, A., Buriro, A. B., Ahmed, J., Ahmed, B., & Akhtar, S. (2022). *Neuromarketing solutions based on EEG signal analysis using machine learning*. *International Journal of Advanced Computer Science and Applications, 13*(1), 298-304. Neuromarketing Solutions based on EEG Signal Analysis using Machine Learning

DataCamp. (2023). *A guide to the gradient boosting algorithm*. DataCamp. https://www.datacamp.com/tutorial/guide-to-the-gradient-boosting-algorithm

DataCamp. (2023). *Random forest algorithm: A complete guide*. DataCamp. https://www.datacamp.com/tutorial/random-forests-classifier-python

GeeksforGeeks. (2023, 18 de diciembre). *Gradient boosting in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/machine-learning/ml-gradient-boosting/

GeeksforGeeks. (2023, 22 de junio). *Random forest algorithm in machine learning*. GeeksforGeeks. https://www.geeksforgeeks.org/random-forest-algorithm-in-machine-learning/

IBM. (s. f.). *What is gradient boosting?* IBM Think. https://www.ibm.com/think/topics/gradient-boosting

IBM. (s. f.). *What is random forest?* IBM Think. https://www.ibm.com/think/topics/random-forest

Networkianos, J. C. (s. f.). *Qué es la regresión logística binaria y cómo analizarla en 6 pasos*. Networkianos. Recuperado de https://networkianos.com

Wikipedia. (2024, 29 de septiembre). *Random forest*. En *Wikipedia*. Recuperado de https://en.wikipedia.org/wiki/Random_forest

Wikipedia. (2024, 21 de noviembre). *Gradient boosting*. En *Wikipedia*. Recuperado de https://es.wikipedia.org/wiki/Gradient_boosting

Apuntes de clase. (s. f.). *Regresión logística binaria y aprendizaje supervisado en R*. [Material docente no publicado].

Google Developers. (2025). *Classification*. Machine Learning Crash Course. Google. https://developers.google.com/machine-learning/crash-course/classification?hl=es-419 d

Pandas Development Team. (2025). *pandas – Python data analysis library*. https://pandas.pydata.org/

NumPy Development Team. (2025). *NumPy – The fundamental package for scientific computing with Python*. https://numpy.org/

Matplotlib Development Team. (2025). *Matplotlib – Python plotting library*. https://matplotlib.org/

Waskom, M. L. (2023). *seaborn: statistical data visualization*. https://seaborn.pydata.org/

Scikit-learn Development Team. (2025). *scikit-learn: Machine learning in Python*. https://scikit-learn.org/

Kaggle. (2025). *kagglehub*. https://pypi.org/project/kagglehub/

OpenAI. (2025). *ChatGPT* (GPT-5.1). https://chat.openai.com/

Google. (2025). *Gemini* (Flash 2.5).https://copilot.microsoft.com/

Microsoft. (2025). *Microsoft Copilot*. https://copilot.microsoft.com/

Google Developers. (s. f.). *Classification*. Machine Learning Crash Course. https://developers.google.com/machine-learning/crash-course/classification?hl=es-419

Borrás, F., Botella, F., Hernández, I., Martínez Mayoral, M. A., Moltó, J., & Morales, J. (2023, diciembre 28). *Visualización de datos multivariantes [Cuaderno de Google Colab]*. Google Colab. https://colab.research.google.com/drive/1sqwOR-4LhjAcOQjb5EnY6O-iptmiKekC?usp=sharing

Equipo de IA4LEGOS. (2024, 30 de junio). *Introducción al aprendizaje automático [Cuaderno de Google Colab]*. Google Colab. https://colab.research.google.com/drive/10hY4Qrn65PkWbKwVMJCXCyELWazR76SA?authuser=1

Equipo de IA4LEGOS. (2024, 30 de junio). *Librería Scikit-Learn para el aprendizaje automático* [Cuaderno de Google Colab]. Google Drive. https://drive.google.com/file/d/1Cf8QoYZ0HkqBJX7B9LCcG2x-3B_6toqJ/view?usp=sharing

Equipo de IA4LEGOS. (2024, 2 de julio). *Modelos de regresión logística para respuesta multinomial* [Cuaderno de Google Colab]. Google Drive. https://drive.google.com/file/d/1HeGYVoFeUwv-SXetMLZoiW4gf-ZSOitU/view?usp=sharing

ANEXOS

El código utilizado para el desarrollo de este estudio está íntegramente contenido en el Anexo I: Código para el análisis exploratorio y procesado de los datos.