

**UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS ADMINISTRATIVAS**



**PREDICCIÓN DEL ÉXITO DE CAMPAÑAS DIGITALES UTILIZANDO
MACHINE LEARNING: CASO DE ESTUDIO EN EXE VIVIENDAS**

**TESIS QUE
PARA OBTENER EL GRADO DE:
LICENCIADO EN INTELIGENCIA DE NEGOCIOS**

PRESENTA

MARIANA HERNÁNDEZ GONZÁLEZ

1. Introducción.....	3
2. Marco Empresarial.....	7
3. Marco Referencial.....	9
3.1 Fundamentos generales del marketing digital.....	10
3.2 Plataformas de anuncios digitales: Facebook y su papel estratégico.....	12
3.3 Inteligencia de negocios y análisis de datos.....	14
3.4 Machine Learning como herramienta predictiva.....	16
3.5 Comparación y justificación del modelo elegido.....	18
3.6 Estudios previos aplicados al marketing predictivo.....	20
3.7 Vacíos en la literatura y contribución del estudio.....	22
4. Metodología.....	24
4.1. Descripción del problema de estudio.....	24
4.2 Descripción del instrumento.....	24
4.3 Definición conceptual de las variables.....	26
4.4. Técnicas utilizadas.....	27
5. Análisis y Resultados de Modelo Machine Learning.....	30
1. 5.1 Carga de Datos.....	30
2. 5.2 Definición de Éxito por Residencial.....	32
3. 5.3 Pre Entrenamiento y Entrenamiento de Modelo Regresión Logística..	34
4. 5.4 Evaluación de Modelo.....	37
5. 5.5 Creación de Diccionario de Categorías.....	39
6. 5.6 Creando Campañas Nuevas.....	41
7. 5.7 Justificación de Resultados.....	43
8. 5.8 Visualización tipo Anillo de Resultados de Modelo.....	46
9. 5.9 Visualización Mapa de Calor, Correlación entre Variables.....	48
10. 5.10 Recomendaciones del Modelo en Prueba.....	51
6. Conclusiones y Recomendaciones.....	53
Anexos.....	56
Referencias Bibliográficas.....	57

1. Introducción

En la era digital, las empresas enfrentan una presión creciente por adaptar sus estrategias a entornos altamente competitivos, donde las decisiones deben estar sustentadas en datos concretos más que en la intuición. El marketing digital ha dejado de ser una herramienta complementaria para convertirse en un eje central de las estrategias comerciales, particularmente en sectores como el inmobiliario, donde cada campaña representa una inversión significativa y de alto impacto.

Sin embargo, a pesar de contar con plataformas publicitarias avanzadas como Meta Ads (anteriormente Facebook Ads), muchas empresas aún ejecutan campañas sin herramientas que les permitan anticipar su rendimiento antes del lanzamiento. Esto genera un margen de error considerable, derivando en decisiones que pueden comprometer el presupuesto y la eficiencia comercial.

La presente investigación surge de una necesidad real dentro de EXE Viviendas, una desarrolladora inmobiliaria ubicada en Mexicali, Baja California. A través del uso exclusivo de datos históricos de campañas ejecutadas en Meta Ads, se propone el desarrollo de un modelo de aprendizaje automático —específicamente una regresión logística binaria— capaz de predecir si una campaña será exitosa (1) o no exitosa (0) utilizando únicamente la información disponible antes de su publicación.

Este enfoque busca transformar la toma de decisiones publicitarias dentro de la empresa, pasando de una lógica reactiva a una predictiva, basada en evidencia y respaldada por tecnología. Además de representar una solución práctica, el proyecto tiene un valor académico, al aplicar conceptos de análisis de datos, inteligencia de negocios y modelado supervisado en un caso real dentro del sector inmobiliario.

Con ello, se busca no solo optimizar el uso de los recursos comerciales de EXE Viviendas, sino también demostrar cómo el uso estratégico de los datos puede convertirse en una herramienta efectiva para mejorar el rendimiento de campañas digitales en entornos altamente competitivos

Planteamiento del Problema.

EXE Viviendas, como muchas empresas del sector inmobiliario, realiza campañas de publicidad digital de forma constante a través de la plataforma Meta Ads (Facebook). No obstante, a pesar de contar con una base de datos histórica sobre el desempeño de estas campañas, actualmente no dispone de una herramienta que permita anticipar si una nueva campaña será exitosa o no antes de su lanzamiento.

Esta carencia limita significativamente la capacidad del área de marketing para tomar decisiones fundamentadas sobre la planeación, segmentación y asignación de presupuesto. Las campañas se ejecutan con un alto grado de incertidumbre, lo que puede derivar en ineficiencia operativa, desperdicio de recursos publicitarios y bajo retorno de inversión (ROI).

Además, la ausencia de un modelo predictivo impide identificar qué combinaciones de variables —como el tipo de mensaje, presupuesto, segmentación de edad o formato de anuncio— han sido más efectivas en campañas anteriores. Esto representa una oportunidad desaprovechada para mejorar la toma de decisiones y replicar estrategias exitosas de forma sistemática.

Ante este contexto, se plantea la siguiente pregunta de investigación:

¿Cómo puede EXE Viviendas predecir el éxito o no de una campaña publicitaria en Facebook utilizando datos históricos y un modelo de regresión logística binaria?

Resolver esta interrogante permitiría a la empresa transitar hacia un modelo de decisiones más estratégico, basado en evidencia y apoyado por herramientas de aprendizaje automático accesibles y comprensibles para su equipo.

Justificación.

El presente estudio se justifica por su potencial de impacto tanto en el ámbito profesional como en el académico. Desde una perspectiva empresarial, brindar a EXE Viviendas una herramienta capaz de predecir si una campaña en Meta Ads será exitosa o no permitirá optimizar decisiones clave en marketing digital. Al anticipar el desempeño con base en datos históricos y un modelo de regresión logística binaria, la empresa podrá asignar mejor su presupuesto, reducir la ejecución de campañas ineficaces y reconocer con mayor claridad los factores que influyen en el éxito de sus estrategias.

En el plano académico, esta investigación constituye una aplicación concreta de técnicas de clasificación supervisada dentro del sector inmobiliario, un ámbito donde aún existe una adopción limitada de modelos predictivos. Mediante el uso exclusivo de datos provenientes de campañas publicitarias en Facebook, se busca aportar evidencia empírica sobre la viabilidad y utilidad de estos enfoques en contextos reales de negocio.

Además, el desarrollo del proyecto fortalece competencias clave en inteligencia de negocios. Entre ellas se incluyen: la interpretación de datos con una visión estratégica orientada a la toma de decisiones; la construcción de modelos estadísticos que permiten identificar patrones relevantes entre variables categóricas y numéricas; y la evaluación de métricas de desempeño mediante indicadores como precisión, recall o F1-score. Estas habilidades consolidan un perfil profesional preparado para afrontar retos analíticos en sectores altamente competitivos.

Objetivo general

Desarrollar un modelo de Machine Learning basado en clasificación binaria que prediga si una campaña publicitaria inmobiliaria será exitosa o no, utilizando datos históricos de campañas digitales ejecutadas por EXE Viviendas en la plataforma de Facebook Ads.

Objetivos específicos:

- Recopilar y estructurar una base de datos con campañas digitales históricas de EXE Viviendas.
- Definir el criterio de “éxito” de una campaña en función del cumplimiento de sus KPIs.
- Aplicar técnicas de análisis y transformación de datos para preparar las variables del modelo.
- Entrenar un modelo de clasificación binaria (regresión logística) para predecir el éxito de campañas.
- Evaluar el rendimiento del modelo y generar recomendaciones para su uso práctico dentro de la empresa.

2. Marco Empresarial

EXE Viviendas es una empresa desarrolladora de conjuntos habitacionales con sede en Mexicali, Baja California. Forma parte del grupo Hermosillo y Asociados, firma reconocida en la región por su participación en proyectos inmobiliarios de mediana y gran escala. La misión principal de EXE es ofrecer soluciones habitacionales accesibles, modernas y bien ubicadas para familias de clase media, con un enfoque en diseño urbano, calidad de construcción y eficiencia en procesos comerciales.

Dentro de su portafolio de desarrollos destacan residenciales como Condesa, El Marqués, Antea, Adara y Reforma, los cuales han contribuido significativamente a la expansión de zonas habitacionales en la ciudad. Cada proyecto es acompañado por campañas de marketing digital con el objetivo de promocionar los inmuebles, generar interés en el mercado y captar prospectos calificados que puedan convertirse en clientes finales.

La empresa ha apostado por una estrategia comercial basada principalmente en publicidad digital a través de Facebook Ads, una plataforma que le permite segmentar su audiencia, medir métricas clave en tiempo real y realizar ajustes sobre la marcha. Estas campañas han sido una herramienta clave para conectar con el público objetivo, especialmente en etapas de preventa y lanzamiento de nuevos residenciales. Sin embargo, la empresa no cuenta con un mecanismo que permita predecir con anticipación el rendimiento de estas campañas, lo cual limita su capacidad para optimizar recursos y mejorar los resultados desde una etapa temprana.

Actualmente, las decisiones sobre presupuesto, diseño y segmentación se toman en gran medida con base en experiencia pasada o criterios subjetivos, sin un respaldo analítico que indique qué factores aumentan la probabilidad de éxito. Esta situación representa un área de oportunidad para fortalecer la gestión comercial mediante el uso de herramientas analíticas y modelos predictivos basados en datos reales.

En este contexto, el presente proyecto plantea el desarrollo de un modelo de Machine Learning orientado a predecir si una campaña será exitosa (1) o no (0), utilizando únicamente la información disponible antes de su ejecución. Este modelo

considera variables como el tipo de mensaje, la segmentación demográfica, el presupuesto asignado, la duración de la campaña y otras características relevantes, con el objetivo de brindar un apoyo estratégico al área de marketing en la toma de decisiones.

Al integrar esta herramienta en sus procesos comerciales, EXE Viviendas podrá identificar patrones de éxito, reducir riesgos y asignar su inversión publicitaria de forma más eficiente, lo que representa un paso importante hacia la consolidación de un enfoque basado en datos dentro del sector inmobiliario.

3. Marco Referencial

Tras una revisión exhaustiva de literatura científica, estudios de caso y documentación especializada, esta sección se estructura sobre seis pilares fundamentales:

- 3.1. Fundamentos del marketing digital,
- 3.2. Facebook Ads como plataforma publicitaria estratégica,
- 3.3. Inteligencia de negocios y análisis de datos,
- 3.4. Machine Learning como herramienta predictiva,
- 3.5. Comparación y justificación del modelo elegido
- 3.6. Estudios previos aplicados al marketing, y
- 3.7. Vacíos en la literatura.

Esta organización permite avanzar de lo general a lo específico, trazando una línea argumentativa clara que vincula el contexto del marketing digital con las oportunidades que ofrece el análisis predictivo. Se busca así sustentar metodológicamente el uso de un modelo de clasificación binaria, específicamente una regresión logística, como apoyo en la toma de decisiones comerciales en EXE Viviendas.

La revisión documental revela que, si bien existe literatura sobre la aplicación de modelos predictivos en campañas publicitarias, no se han encontrado estudios centrados exclusivamente en el sector inmobiliario mexicano utilizando datos de campañas ejecutadas en Facebook Ads, lo cual refuerza la pertinencia y originalidad del presente estudio.

3.1 Fundamentos generales del marketing digital.

El marketing digital representa una evolución fundamental en la manera en que las organizaciones interactúan con sus audiencias, marcando una ruptura respecto al paradigma tradicional centrado en medios masivos y mensajes generalizados. A lo largo de las últimas décadas, el marketing ha transitado hacia un enfoque multidimensional, donde la hipersegmentación, la interacción constante y la adaptabilidad en tiempo real definen la relación entre marcas y consumidores (Chaffey & Ellis-Chadwick, 2019).

Este cambio ha sido impulsado, principalmente, por el crecimiento acelerado del acceso a internet, el uso cotidiano de dispositivos móviles y la consolidación de las redes sociales como canales primarios de comunicación. Según Statista (2023), más del 60% de la población mundial utiliza internet, lo que ha generado un ecosistema digital altamente activo y competitivo. En este entorno, las empresas deben entender que ya no basta con estar presentes en línea, sino que deben construir experiencias personalizadas, significativas y basadas en datos para captar y retener la atención del usuario (Kotler, Kartajaya & Setiawan, 2017).

Una de las principales ventajas del marketing digital frente a los enfoques tradicionales radica en su capacidad de medición. Herramientas como Google Analytics, Meta Ads Manager o HubSpot permiten rastrear con precisión el comportamiento del usuario: desde la primera interacción con un anuncio hasta la conversión final. Esta trazabilidad promueve una cultura de decisiones basadas en evidencia, donde la intuición cede terreno al análisis de métricas clave como el CTR (Click Through Rate), el ROAS (Return on Ad Spend), o el engagement rate (Ryan, 2016).

Además, el entorno digital ha favorecido el surgimiento de una publicidad bidireccional. El consumidor ya no es un receptor pasivo, sino un actor que participa activamente, expresa opiniones, comparte experiencias y espera respuestas inmediatas. Esta dinámica ha obligado a las marcas a redefinir su propuesta de valor y adoptar estrategias centradas en el usuario. Conceptos como el marketing conversacional, la personalización predictiva o el contenido interactivo reflejan esta transformación, donde la experiencia del cliente se convierte en el eje estratégico de

las campañas (Tuten & Solomon, 2020).

Por otro lado, tecnologías como el big data, el aprendizaje automático (machine learning) y la inteligencia artificial han permitido avanzar hacia escenarios donde la predicción, automatización y optimización son posibles incluso antes de lanzar una campaña. Por ejemplo, mediante modelos de regresión logística, árboles de decisión o redes neuronales, hoy es factible estimar con alto grado de precisión si una campaña será exitosa o no, utilizando únicamente datos previos a su publicación (Kumar et al., 2021). Esto representa un cambio paradigmático: las decisiones ya no se basan en el resultado de campañas pasadas, sino en proyecciones sustentadas por modelos predictivos construidos a partir de históricos enriquecidos.

El presente estudio se enmarca dentro de esta lógica contemporánea del marketing digital. A través del desarrollo de un modelo predictivo aplicado a campañas de EXE Inmobiliaria, se busca demostrar que el uso estratégico de los datos —incluso antes del lanzamiento de una campaña— puede ser una herramienta clave para anticipar resultados y orientar la toma de decisiones. Esta perspectiva no solo mejora la eficiencia operativa, sino que también reduce el riesgo asociado a la inversión publicitaria y permite enfocar recursos en estrategias con mayor probabilidad de éxito.

Por lo tanto, el marketing digital, más allá de ser una adaptación tecnológica, representa una transformación sistémica que redefine los procesos, capacidades y mentalidades dentro de las organizaciones. Aquellas empresas que logren integrar de forma eficaz la analítica de datos, la personalización y la automatización tendrán una ventaja competitiva significativa en un entorno donde la atención es cada vez más efímera y la exigencia del usuario más alta que nunca.

3.2 Plataformas de anuncios digitales: Facebook y su papel estratégico

En el ecosistema actual del marketing digital, Facebook (hoy parte del conglomerado Meta Platforms, Inc.) se ha consolidado como una de las plataformas publicitarias más influyentes y robustas del mundo. Con más de 3 mil millones de usuarios activos mensuales en sus diferentes aplicaciones (Facebook, Instagram, Messenger y WhatsApp), la red de Meta no solo ofrece un alcance masivo, sino también herramientas altamente sofisticadas para la segmentación, personalización y medición de campañas (Meta Business, 2024).

Uno de los factores que ha posicionado a Facebook como una plataforma estratégica para los anunciantes es su capacidad de recopilación y análisis de datos del comportamiento digital de sus usuarios. A través de interacciones como reacciones, comentarios, clics, vistas de video y formularios completados, la plataforma construye perfiles detallados que permiten a las marcas dirigirse a audiencias específicas con base en intereses, hábitos de navegación, eventos de vida, ubicación geográfica y conexiones sociales (Tuten & Solomon, 2017). Esta riqueza de datos posibilita una microsegmentación avanzada, que supera las limitaciones de las audiencias generales del marketing tradicional y permite enviar mensajes mucho más relevantes y personalizados.

La plataforma de anuncios de Facebook funciona bajo un sistema de subasta automatizada, donde cada anuncio compite por impresiones en función de su presupuesto, relevancia y nivel de engagement esperado. Esta lógica no solo democratiza el acceso a la publicidad digital —permitiendo a pequeñas empresas competir con grandes corporaciones—, sino que también fomenta la optimización constante de los anuncios en tiempo real (Chaffey & Ellis-Chadwick, 2019). El algoritmo prioriza mostrar contenidos con mayor probabilidad de interacción para mejorar la experiencia del usuario y aumentar la efectividad publicitaria.

Entre las ventajas más destacadas de Facebook Ads se encuentran la diversidad de formatos (imágenes, carruseles, video, formularios instantáneos, etc.), la posibilidad de realizar pruebas A/B, la creación de audiencias personalizadas basadas en datos de comportamiento previo (como visitas al sitio web, interacción con el contenido o

apertura de correos electrónicos), así como la opción de audiencias similares (lookalike) que permiten escalar campañas a públicos con características análogas a los clientes existentes (Ryan, 2016).

En el contexto de la presente tesina, centrada en el sector inmobiliario, Facebook representa una herramienta particularmente útil por su capacidad para captar leads mediante formularios incrustados y por su compatibilidad con CRMs y plataformas externas. Su alcance geográfico —que puede limitarse a zonas específicas como Poniente, Centro o Oriente de Mexicali— y su posibilidad de segmentar por edad, intereses o intención de compra, lo convierten en un canal ideal para evaluar la efectividad de campañas digitales inmobiliarias.

Por esta razón, el presente modelo predictivo se basa exclusivamente en datos extraídos de campañas publicitarias gestionadas en la plataforma de Meta. Esta decisión metodológica permite mantener la homogeneidad del origen de los datos, asegurar la consistencia en las métricas analizadas (como CTR, CPC, alcance o conversiones) y reflejar de manera más precisa los patrones de comportamiento de los prospectos que interactúan con los anuncios de EXE Viviendas. En consecuencia, el modelo no solo pretende evaluar retroactivamente el éxito de las campañas, sino también anticipar su probabilidad de éxito futuro antes de su lanzamiento, contribuyendo a una toma de decisiones más informada y estratégica dentro del área comercial y de marketing de la empresa.

3.3 Inteligencia de negocios y análisis de datos

En un entorno digital cada vez más complejo, caracterizado por un crecimiento exponencial en la generación de datos, la capacidad para convertir esa información en conocimiento útil y accionable se ha vuelto un diferenciador clave para las organizaciones competitivas. La inteligencia de negocios (Business Intelligence, BI) surge como una disciplina estratégica que integra herramientas, metodologías y procesos para recolectar, transformar y analizar grandes volúmenes de datos con el fin de apoyar la toma de decisiones fundamentadas (Chen, Chiang & Storey, 2012). Esta capacidad analítica no solo mejora la eficiencia operativa, sino que fortalece la planificación táctica y la visión estratégica de las empresas en sectores altamente competitivos, como el inmobiliario.

A diferencia de enfoques tradicionales donde el análisis de datos se limitaba a generar reportes retrospectivos, la BI moderna incorpora modelos avanzados de análisis que permiten descubrir patrones ocultos, predecir comportamientos y simular escenarios. Como lo señalan Wixom y Watson (2010), la inteligencia de negocios no debe entenderse únicamente como una herramienta técnica, sino como una filosofía de gestión basada en la evidencia. Es decir, las organizaciones que integran la BI de manera efectiva son aquellas capaces de responder en tiempo real a cambios del mercado, optimizar sus campañas, anticiparse a riesgos y capitalizar oportunidades.

Dentro de la BI, el análisis de datos se estructura comúnmente en tres niveles: análisis descriptivo, análisis predictivo y análisis prescriptivo. El análisis descriptivo se enfoca en interpretar qué ha sucedido, proporcionando resúmenes estadísticos y visualizaciones de datos pasados. El análisis predictivo, por su parte, utiliza modelos matemáticos y estadísticos para estimar lo que podría ocurrir en el futuro. Finalmente, el análisis prescriptivo busca determinar qué debería hacerse, recomendando decisiones óptimas a partir de simulaciones o algoritmos de optimización (Turban, Sharda & Delen, 2020). Cada uno de estos niveles aporta valor en distintas etapas del proceso de toma de decisiones.

Este proyecto se sitúa en el nivel predictivo, al aplicar técnicas de Machine Learning particularmente la regresión logística para estimar la probabilidad de éxito de campañas digitales en EXE Inmobiliaria. El enfoque adoptado permite modelar el comportamiento de campañas anteriores a partir de características conocidas antes

de su lanzamiento, entrenando así al modelo con patrones históricos que posteriormente permiten anticipar resultados futuros. Este tipo de análisis es especialmente útil en contextos donde la planificación de campañas requiere precisión, optimización de presupuesto y una alta probabilidad de conversión.

Además, la implementación de la inteligencia de negocios no solo fortalece la toma de decisiones en el corto plazo, sino que construye una base de conocimiento organizacional que se enriquece con cada nueva campaña. Cada dato recolectado, correctamente almacenado y analizado, contribuye a afinar las estrategias futuras. En el caso de EXE Inmobiliaria, la incorporación de BI y análisis predictivo les permite pasar de un enfoque reactivo a uno proactivo: identificar qué combinaciones de segmentación, formato y mensaje tienden a campañas exitosas, antes de invertir en su ejecución.

Por lo tanto, la inteligencia de negocios no solo se posiciona como un soporte tecnológico, sino como un eje transversal que redefine la manera en que se conciben, ejecutan y evalúan las campañas de marketing digital. Su integración con técnicas analíticas, como la regresión logística, transforma los datos en una ventaja competitiva tangible para empresas como EXE Inmobiliaria, permitiéndoles alinear mejor sus decisiones con los resultados que desean alcanzar.

3.4 Machine Learning como herramienta predictiva

El avance de las tecnologías analíticas ha permitido a las empresas no solo observar el pasado, sino también anticiparse al futuro mediante modelos predictivos. En este escenario, el *Machine Learning* se ha convertido en una herramienta estratégica para identificar patrones complejos en los datos y generar predicciones basadas en evidencia. A diferencia de los métodos estadísticos tradicionales, el aprendizaje automático puede adaptarse dinámicamente a los cambios en los datos, mejorando la capacidad de predicción conforme se dispone de más información (James et al., 2021).

Desde un enfoque general, el *Machine Learning* se clasifica en tres tipos principales: **aprendizaje supervisado**, **no supervisado** y **por refuerzo**. En el aprendizaje supervisado, el modelo se entrena utilizando un conjunto de datos etiquetado, es decir, donde se conoce el resultado o “respuesta correcta” para cada observación. Este tipo de aprendizaje es útil cuando se desea predecir una variable objetivo a partir de atributos conocidos, como ocurre en la presente investigación al estimar el éxito o no de una campaña. Por otro lado, el aprendizaje no supervisado se emplea cuando no se cuenta con una variable objetivo, y el propósito es identificar estructuras o patrones ocultos en los datos, como grupos o segmentos de clientes. Finalmente, el aprendizaje por refuerzo se basa en la interacción del modelo con un entorno, aprendiendo a tomar decisiones óptimas a partir de recompensas o penalizaciones, siendo más común en áreas como robótica o videojuegos (Russell & Norvig, 2021; Géron, 2019).

Dado que el objetivo de este estudio es predecir un resultado específico (éxito o no de la campaña) a partir de características previamente conocidas —como el tipo de mensaje, segmentación o formato del anuncio—, se optó por el enfoque de aprendizaje supervisado. Este permite que el modelo aprenda de datos históricos etiquetados, identificando patrones que se relacionan con el éxito en campañas anteriores, para así realizar predicciones en nuevas campañas aún no lanzadas.

En este marco del aprendizaje supervisado —una de las ramas más utilizadas del *Machine Learning*— la **regresión logística** ha sido ampliamente aplicada en problemas de clasificación binaria, como lo es determinar si una campaña será

exitosa o no. Esta técnica permite estimar la probabilidad de ocurrencia de un evento categórico (éxito = 1, sin éxito = 0) en función de un conjunto de variables independientes, proporcionando una salida probabilística fácilmente interpretable (Kuhn & Johnson, 2013).

La elección de regresión logística para este estudio responde a su capacidad de ofrecer modelos simples, transparentes y estadísticamente robustos. A diferencia de otros algoritmos más complejos como *Random Forest* o redes neuronales, la regresión logística facilita la interpretación de los coeficientes, lo que resulta clave en entornos donde se requiere justificar las decisiones ante partes interesadas no técnicas (James et al., 2021). En el contexto de EXE Inmobiliaria, esta interpretabilidad es un factor crucial para adoptar el modelo dentro de los procesos operativos y estratégicos de la empresa.

El modelo desarrollado utiliza variables estructuradas generadas a partir de campañas históricas en Facebook, tales como tipo de mensaje, presupuesto, segmentación demográfica y comportamiento del público. Estas variables fueron seleccionadas en función de su correlación con la variable objetivo y su disponibilidad constante en futuras campañas. Con estos datos, se entrenó un modelo de regresión logística que permite estimar, con base en datos previos, la probabilidad de que una campaña tenga éxito según la definición establecida por EXE.

Este enfoque se alinea con lo planteado por Provost y Fawcett (2013), quienes destacan que los modelos predictivos en marketing digital no solo deben ser precisos, sino también comprensibles y operables. Así, la regresión logística se posiciona como una alternativa eficaz y adecuada para empresas que buscan incorporar modelos analíticos a su operación sin requerir infraestructura técnica compleja ni conocimientos avanzados en ciencia de datos.

Tabla 1.1

Modelo	Ventajas	Limitaciones	Referencia
Regresión logística	Rápido, interpretable, útil para clasificación binaria sencilla.	Bajo desempeño en relaciones no lineales complejas.	Provost & Fawcett, 2013
Árbol de decisión	Intuitivo, visualizable, útil para generar reglas comprensibles.	Alta propensión al sobreajuste.	Witten et al., 2016
Random Forest	Robusto, preciso, resistente al sobreajuste, eficaz con datos mixtos.	Menor interpretabilidad que un solo árbol.	Breiman, 2001
K-Nearest Neighbors	Simple de implementar, no requiere entrenamiento previo.	Lento en grandes datasets, sensible al ruido.	Shalev-Shwartz & Ben-David, 2014
SVM	Alta precisión en espacios complejos.	Requiere normalización, difícil de interpretar y ajustar.	Cortes & Vapnik, 1995
Red neuronal	Procesa patrones complejos y texto.	Mayor entrenamiento, menor transparencia.	Goodfellow et al., 2016

3.5 Comparación y justificación del modelo elegido

Durante la fase de experimentación, se compararon distintos algoritmos de clasificación supervisada (Tabla 1.1) para determinar cuál ofrecía el mejor equilibrio entre precisión predictiva, interpretabilidad y aplicabilidad práctica en el entorno operativo de EXE Inmobiliaria. Si bien modelos como los árboles de decisión o

Random Forest demostraron una capacidad predictiva ligeramente superior en métricas como la exactitud o el F1-score, su complejidad interpretativa los hace menos idóneos en contextos donde es necesario explicar con claridad los factores que inciden en el éxito de una campaña. Por esta razón, se optó por emplear regresión logística como modelo final, dado que permite estimar probabilidades de forma transparente y comprensible, facilitando la interpretación de los coeficientes y su impacto sobre la variable objetivo. Esta decisión se alinea con recomendaciones metodológicas en entornos de marketing digital, donde la facilidad de implementación y la claridad en la explicación del modelo suelen primar sobre una mínima ganancia en precisión (Kuhn & Johnson, 2013; James et al., 2021; Provost & Fawcett, 2013).

3.6 Estudios previos aplicados al marketing predictivo

El uso de técnicas de Machine Learning (ML) ha crecido significativamente en los últimos años como una herramienta esencial para automatizar procesos, optimizar decisiones y generar predicciones a partir de grandes volúmenes de datos. Como señala Marr (2016), industrias como el comercio electrónico, los servicios financieros, el retail y la logística han sido pioneras en la implementación de modelos predictivos para tareas como la recomendación de productos, la detección de fraude, la segmentación de clientes o la predicción del abandono. Estos casos han demostrado el valor tangible del ML como una tecnología aplicable y rentable en escenarios de negocio reales.

Sharda, Delen y Turban (2018) confirman que en el ámbito del marketing digital, los algoritmos de aprendizaje automático permiten anticipar comportamientos del consumidor, asignar presupuestos de forma más eficiente y personalizar mensajes en tiempo real. Según los autores, esta capacidad predictiva se ha traducido en mayores tasas de conversión y una optimización significativa del retorno sobre la inversión publicitaria (ROI). En cuanto a los algoritmos utilizados, diversos estudios coinciden en que los más empleados en contextos industriales y comerciales son la regresión logística, los árboles de decisión, el random forest y las redes neuronales.

Según Provost y Fawcett (2013), estos modelos ofrecen un buen equilibrio entre precisión, velocidad de ejecución y facilidad de interpretación, lo cual es fundamental para que los analistas y equipos de marketing puedan tomar decisiones basadas en sus salidas.

Casos documentados como los de Amazon, Netflix y Alibaba muestran cómo estas organizaciones utilizan modelos de clasificación para anticipar intención de compra, algoritmos de regresión para predecir montos de compra futura, y técnicas de agrupamiento (clustering) para segmentar usuarios de manera más específica (Marr, 2016; Witten, Frank & Hall, 2016).

Estas aplicaciones no solo optimizan el proceso de targeting, sino que también permiten implementar estrategias automatizadas de retención o reactivación, con base en predicciones obtenidas en tiempo real. Tiago y Veríssimo (2014) también destacan que el valor del marketing predictivo se incrementa cuando los modelos son alimentados con datos provenientes de distintas fuentes, como el CRM,

plataformas publicitarias y redes sociales, generando una visión 360° del cliente.

Esta integración de datos mejora la calidad de las predicciones y habilita estrategias mucho más personalizadas, adaptadas al contexto de cada usuario. En resumen, la literatura muestra una tendencia clara hacia la adopción de modelos de ML en el marketing moderno, tanto en grandes corporaciones como en empresas más pequeñas que desean tomar decisiones informadas a partir de sus datos. Estos antecedentes confirman que el desarrollo de un modelo predictivo, como el que se propone en esta investigación para EXE Inmobiliaria, no solo es viable, sino alineado con las mejores prácticas actuales en análisis de campañas y comportamiento del consumidor.

3.7 Vacíos en la literatura y contribución del estudio

A pesar del crecimiento sostenido de las tecnologías de análisis predictivo, la adopción del Machine Learning en América Latina sigue siendo desigual. Como reporta la Organización para la Cooperación y el Desarrollo Económicos (OCDE, 2022), la mayoría de las implementaciones en la región se concentran en grandes corporativos o startups tecnológicas, mientras que las pequeñas y medianas empresas aún enfrentan barreras significativas como la escasez de talento especializado, desconocimiento técnico o una baja cultura de toma de decisiones basada en datos. Esta brecha tecnológica es más evidente en sectores como el inmobiliario, donde el uso de modelos predictivos aún no se ha masificado. Según López-Morales, Romero-Torres y Ríos-Carmenado (2020), los estudios aplicados a machine learning en países en desarrollo tienden a estar dispersos, con poca sistematización, y enfocados principalmente en sectores como el financiero, manufactura, educación o salud. México no es la excepción. En la revisión documental realizada para este trabajo, se identificaron aplicaciones de ML en escenarios urbanos, como predicción de precios, segmentación de compradores o automatización de avalúos, pero no en el análisis estratégico de campañas publicitarias desde un enfoque de marketing digital. Como señala Fernández Loayza (2021), algunas tesis y artículos divulgativos han explorado temas como la velocidad de venta de propiedades o la evaluación de terrenos habitacionales mediante modelos de regresión, pero estos estudios tienden a tratar al inmueble como producto, y no abordan el rendimiento de las campañas como variable analítica. Es decir, se enfocan en el qué se vende, pero no en cómo se promueve, ni qué tan efectiva fue la estrategia digital que lo posicionó. De acuerdo con la OCDE (2022), incluso dentro del sector inmobiliario, los enfoques documentados se centran en:

- Predicción de precios mediante modelos de regresión (lineal, múltiple o con árboles de decisión),
- Modelos automatizados de valuación (AVMs),
- Segmentación de clientes mediante clustering,
- Estimación del tiempo promedio de venta de propiedades.

Sin embargo, no se han identificado estudios que integren información interna proveniente de plataformas como Facebook Ads para predecir el éxito de campañas

digitales en empresas desarrolladoras. Esta brecha se vuelve aún más evidente al reducir el enfoque geográfico. En estados como Baja California, y particularmente en la ciudad de Mexicali, no existen investigaciones académicas publicadas que documenten el uso de algoritmos de clasificación supervisada binaria, como Regresión Logística, aplicados al análisis del desempeño de campañas digitales en desarrolladoras inmobiliarias. Como sostienen Tiago y Veríssimo (2014), el éxito del marketing digital moderno radica precisamente en la capacidad de integrar múltiples fuentes de datos —como el CRM y las plataformas de anuncios— en modelos de decisión que anticipen y optimicen resultados. Aunque es posible que algunas empresas locales ya estén explorando este tipo de tecnologías, la ausencia de documentación científica formal confirma el vacío metodológico y práctico en este contexto. En ese sentido, el presente estudio representa una contribución original e innovadora, al documentar de manera detallada el desarrollo y entrenamiento de un modelo predictivo binario aplicado a un caso inmobiliario real en Mexicali. La base utilizada, compuesta por datos históricos reales y simulados estructurados, se integra bajo una metodología replicable y alineada con las mejores prácticas del análisis predictivo, abriendo la puerta a futuras aplicaciones en el sector inmobiliario regional.

4. Metodología

4.1. Descripción del problema de estudio

La presente investigación adopta un enfoque cuantitativo, ya que se basa en el análisis de datos numéricos como clics, leads, presupuesto, duración, entre otros. Se clasifica como un estudio explicativo y predictivo, al buscar identificar relaciones entre variables que influyen en el éxito de una campaña, y al implementar un modelo supervisado de clasificación binaria regresión logística para anticipar dicho éxito. Es un estudio aplicado, pues busca resolver una situación específica en un contexto real: mejorar la toma de decisiones sobre campañas publicitarias de EXE Viviendas. Asimismo, es un estudio transversal y prospectivo, ya que analiza datos históricos de un periodo definido y tiene como propósito proyectar resultados futuros de campañas aún no lanzadas.

4.2 Descripción del instrumento

Población, muestra y características del dataset;

El dataset contiene información histórica sobre **2,011 campañas publicitarias** realizadas por la empresa **EXE Viviendas** en la plataforma **Meta Ads (Facebook e Instagram)**. Cada registro en el conjunto de datos representa una campaña individual y está compuesto por una combinación de **variables categóricas** que describen las decisiones estratégicas tomadas antes de su ejecución, así como **métricas numéricas de desempeño** obtenidas tras su lanzamiento.

Entre las variables categóricas se incluyen:

- **Residencial:** proyecto inmobiliario al que está dirigida la campaña (por ejemplo, ANTEA, ADARA, etc.).
- **Segmentación de edad y género:** público objetivo al que se orienta la campaña.
- **Formato del anuncio:** tipo de recurso utilizado (imagen, video o carrusel).
- **Tipo de mensaje:** estilo de comunicación (emocional, familiar, económico,

etc.).

- **Call to Action:** acción sugerida al usuario (registrarse, enviar mensaje, más información).
- **Dispositivo objetivo:** plataforma de visualización prioritaria (celular, computadora o ambos).
- **Temporada:** momento del año en que se ejecutó (primavera, verano, otoño, invierno).

Las variables numéricas del dataset reflejan el rendimiento de cada campaña, incluyendo:

- **Leads generados** (clientes potenciales).
- **Conversaciones iniciadas** (mensajes recibidos).
- **Alcance total.**
- **CTR** (Click Through Rate).
- **Número de impresiones y clics.**

En total, el dataset cuenta con **2,011 filas** (una por cada campaña) y **17 columnas**. Como parte del procesamiento, se añadió una columna adicional llamada '**éxito**', que etiqueta cada campaña como **exitosa (1)** o **no exitosa (0)** según si cumple al menos **2 de los 4 KPIs definidos por el área de marketing de EXE Viviendas**.

De todas las campañas, **1,314 fueron exitosas (65.35%)** y **697 no exitosas (34.65%)**, lo que permitió entrenar un modelo de clasificación supervisado para **predecir el éxito potencial de futuras campañas** en función de las características previas a su ejecución.

Una característica destacada del dataset es que **la definición de éxito no es uniforme**, sino que **se adapta a las metas de cada residencial**. Esto significa que los **umbrales de éxito para leads, mensajes, alcance y CTR** varían dependiendo del proyecto inmobiliario. Por ejemplo, lo que se considera una campaña exitosa para el residencial **ANTEA** puede diferir de lo esperado para **PORTAGO** o **IRIA**. Esta diferenciación contextual permite un análisis más realista y estratégico, alineado con las metas comerciales de cada zona de ventas.

Kpis Definidos por equipo de Marketing de Exe:

- Para ADARA (297 campañas): Leads ≥ 35 , Conversaciones ≥ 25 , Alcance $\geq 50,000$, CTR $\geq 1.2\%$
- Para ANTEA (387 campañas): Leads ≥ 40 , Conversaciones ≥ 30 , Alcance $\geq 60,000$, CTR $\geq 1.5\%$
- Para REFORMA (621 campañas), CONDESA (675 campañas), PORTAGO (10 campañas) e IRIA (21 campañas): se establecieron combinaciones específicas de KPIs con valores ajustados a su desempeño histórico

Para mantener coherencia con la lógica comercial de la empresa, se consideró que una campaña era exitosa si cumplía al menos dos de los KPIs definidos para su residencial. Esta decisión permitió capturar el matiz operativo real de cada desarrollo y evitar una interpretación reduccionista del éxito publicitario. A partir de esta evaluación se construyó la variable binaria **exito_campaña**, que fue utilizada como variable dependiente en el modelo de clasificación.

4.3 Definición conceptual de las variables

Cada campaña está descrita a través de una combinación de variables numéricas y categóricas, que permiten capturar sus resultados. A continuación, se muestra una clasificación general de los tipos de variables presentes:

Variables numéricas:

- presupuesto_final
- Conversaciones con mensajes iniciadas
- Clientes potenciales

- Alcance
- Impresiones
- Interacciones_Totales
- Resultados
- Costo por resultado
- Frecuencia
- Clics en el enlace
- Clics únicos en el enlace
- Clics salientes
- Clics salientes únicos
- Clics (todos)

Variables categóricas:

- residencial
- segmentación_edad
- segmentación_genero
- formato_anuncio
- tipo_mensaje
- call_to_action
- dispositivo_objetivo
- temporada
- tipo de resultado

Variable objetivo:

- 1= Exito, 0=No Exito

4.4. Técnicas utilizadas

Para llevar a cabo la presente investigación, se recurrió a una combinación de herramientas tecnológicas y metodologías de análisis que permitieron gestionar de forma eficaz las distintas etapas del proceso: desde la recolección y limpieza de los datos hasta la construcción, evaluación y visualización del modelo predictivo. A continuación, se detallan las técnicas y recursos empleados:

Microsoft Excel

Esta herramienta fue utilizada en las etapas iniciales del tratamiento de los datos. Su interfaz amigable y funciones de filtrado, ordenamiento y validación manual facilitaron la detección de errores, valores atípicos y datos faltantes en los archivos provenientes de Meta Ads. Asimismo, permitió estructurar y homogeneizar las bases de datos correspondientes a distintas campañas, asegurando que todas compartieran una misma nomenclatura y orden de columnas, lo cual resultó esencial para la posterior integración en Python.

Meta Ads (Facebook Ads Manager)

La totalidad de las métricas empleadas para el análisis provienen de esta plataforma, que constituyó la única fuente de datos del estudio. Meta Ads proporciona información detallada y estructurada sobre el rendimiento de las campañas publicitarias, permitiendo la extracción de indicadores clave como: impresiones, clics, alcance, leads, CTR (Click Through Rate), CPC (Costo por Clic), frecuencia, conversiones, entre otros. Esta selección de métricas permitió construir una visión integral del comportamiento de las campañas y sentó las bases para definir la variable objetivo de “éxito”.

Python

El lenguaje de programación Python fue seleccionado como herramienta principal para la manipulación de datos, desarrollo del modelo de machine learning y visualización de resultados, dada su versatilidad, comunidad activa y disponibilidad de bibliotecas especializadas en ciencia de datos. En concreto, se utilizaron las siguientes, entre otras que se explicarán más adelante a detalle:

- **pandas y numpy:** estas bibliotecas fueron fundamentales para el manejo estructurado de los datos. Permitieron la lectura de archivos, transformación de columnas, codificación de variables categóricas y creación de nuevas variables derivadas, como el mes y año de inicio de campaña. pandas también facilitó operaciones como el filtrado, agrupamiento y combinación de conjuntos de datos.
- **scikit-learn:** biblioteca clave en la construcción del modelo de regresión logística. Se utilizó para dividir los datos en subconjuntos de entrenamiento y

prueba, normalizar variables numéricas, aplicar codificadores a variables categóricas, entrenar el modelo, y evaluar su desempeño mediante métricas como precisión (accuracy), matriz de confusión y reporte de clasificación.

- **matplotlib y seaborn:** estas herramientas de visualización fueron utilizadas para representar gráficamente las relaciones entre variables, así como para analizar la importancia relativa de cada variable predictora dentro del modelo. Los gráficos generados facilitaron la interpretación de patrones y contribuyeron a la elaboración de conclusiones más comprensibles tanto para un público técnico como no técnico.

Jupyter Notebook

Como entorno de desarrollo, Jupyter Notebook permitió integrar de manera ordenada todo el flujo de trabajo. Esta plataforma facilitó la documentación en tiempo real del código, la visualización inmediata de resultados y la trazabilidad de cada etapa del análisis. Su formato interactivo resultó especialmente útil para combinar celdas de código, texto explicativo, visualizaciones y resultados en un solo documento, promoviendo así la transparencia y reproducibilidad del proyecto.

En conjunto, estas herramientas permitieron llevar a cabo un proceso de análisis riguroso, estructurado y replicable, alineado con las mejores prácticas en ciencia de datos aplicada al marketing digital.

5. Análisis y Resultados de Modelo Machine Learning

Antes de detallar cada fase del desarrollo, es importante contextualizar que, de manera inicial, el modelo fue entrenado con campañas históricas completas. Esto implica considerar tanto las variables disponibles antes del lanzamiento (como `segmentación_edad`, tipo de mensaje, `segmentación_genero`, temporada) como los resultados obtenidos después de su ejecución, tales como clics, leads, mensajes recibidos. Una vez que el modelo ha sido expuesto a este historial completo y ha aprendido qué combinaciones de características iniciales tienden a asociarse con campañas exitosas frente a las que no lo fueron, puede considerarse entrenado. A partir de ese punto, está en condiciones de recibir únicamente variables preliminares (antes del lanzamiento) y predecir con base en ellas la probabilidad de éxito de una nueva campaña.

El desarrollo del modelo se dividió en las siguientes fases:

1. 5.1 Carga de Datos

En esta primera sección se importan todas las librerías necesarias que permiten construir, entrenar y evaluar un modelo de aprendizaje automático. La librería **pandas** se utiliza para el manejo y análisis de datos tabulares, permitiendo cargar el archivo fuente y realizar operaciones de limpieza y transformación. Posteriormente, se importan funciones del módulo **sklearn**, específicamente herramientas como **train_test_split**, que se emplea para dividir los datos en conjuntos de entrenamiento y prueba; **OneHotEncoder** y **ColumnTransformer**, que permiten transformar variables categóricas en variables numéricas codificadas; **Pipeline**, que facilita la organización de los pasos del preprocesamiento y el modelo en una única estructura coherente; y **LogisticRegression**, que se utiliza como clasificador principal para predecir si una campaña será exitosa o no. Además, se incluye **resample** para balancear las clases en caso de desbalanceo entre campañas exitosas y no exitosas, y se incorporan métricas como **classification_report** y **confusion_matrix** para evaluar el rendimiento del modelo. Finalmente, se integran **matplotlib.pyplot** y **seaborn**, dos bibliotecas para generar

visualizaciones gráficas que ayudan a interpretar los resultados del modelo y explorar el comportamiento de las variables.

```
# -----  
# 1. Carga de datos  
# -----  
import pandas as pd  
from sklearn.model_selection import train_test_split  
from sklearn.preprocessing import OneHotEncoder  
from sklearn.compose import ColumnTransformer  
from sklearn.pipeline import Pipeline  
from sklearn.linear_model import LogisticRegression  
from sklearn.utils import resample  
from sklearn.metrics import classification_report, confusion_matrix  
import matplotlib.pyplot as plt  
import seaborn as sns
```

Una vez importadas las librerías, se procede a cargar el archivo de datos “TESINA FINAL.xlsx” a través de la función `read_excel()` de `pandas`, almacenándolo en un DataFrame llamado `df`. Este archivo contiene el histórico de campañas realizadas por el negocio, incluyendo variables categóricas como la segmentación del público, el tipo de anuncio, el dispositivo objetivo, entre otros, así como métricas de desempeño como número de clientes potenciales, alcance y conversaciones iniciadas.

```
# Cargar el archivo original  
file_path = "/Users/marianahernandez/TESINA/TESINA FINAL.xlsx"  
df = pd.read_excel(file_path)
```

Posteriormente, se realiza una transformación inicial sobre las columnas del DataFrame, renombrando las variables “Clientes potenciales”, “Conversaciones con mensajes iniciadas” y “Alcance” por los nombres “leads”, “mensajes_recibidos” y “alcance”, respectivamente. Esta

renombración tiene como objetivo estandarizar los nombres de las columnas para facilitar su manipulación en las siguientes etapas del procesamiento, asegurando mayor claridad y consistencia en el análisis posterior.

```
# Renombrar columnas para evaluación de KPI
df = df.rename(columns={
    'Clientes potenciales': 'leads',
    'Conversaciones con mensajes iniciadas': 'mensajes_recibidos',
    'Alcance': 'alcance'
})
```

2. 5.2 Definición de Éxito por Residencial

Esta sección del código tiene como finalidad construir una variable objetivo que clasifique si una campaña fue exitosa o no, con base en criterios comerciales previamente definidos. Para ello, se establece un diccionario denominado **kpi_por_residencial**, el cual contiene las metas mínimas (KPI) que cada residencial debe cumplir para considerarse como exitosa. Estas metas incluyen valores mínimos de leads (clientes potenciales), mensajes recibidos (conversaciones iniciadas), alcance (personas alcanzadas con la campaña), y CTR (click-through rate). Cada residencial por ejemplo, ANTEA, REFORMA o IRIA tiene valores personalizados de estos KPIs, lo que refleja diferencias en sus objetivos de marketing según su ubicación, mercado meta o presupuesto asignado.

```
# -----

# 2. Generar etiqueta 'éxito'

# -----

kpi_por_residencial = {

    'ADARA': {'leads': 35, 'mensajes': 25, 'alcance': 50000, 'ctr': 0.012},

    'REFORMA': {'leads': 30, 'mensajes': 20, 'alcance': 40000, 'ctr': 0.010},
```

```

'ANTEA': {'leads': 40, 'mensajes': 30, 'alcance': 60000, 'ctr': 0.015},
'PORTAGO': {'leads': 25, 'mensajes': 15, 'alcance': 35000, 'ctr': 0.010},
'IRIA': {'leads': 20, 'mensajes': 10, 'alcance': 30000, 'ctr': 0.009},
'CONDESA': {'leads': 30, 'mensajes': 20, 'alcance': 40000, 'ctr': 0.010},
}

```

Posteriormente, se define la función **calcular_exito**, que recibe una fila del DataFrame (representando una campaña específica) y compara sus valores con las metas establecidas en el diccionario. Primero, recupera las reglas correspondientes a la residencial de la campaña. Si no hay reglas definidas para esa residencial, la campaña se considera automáticamente como no exitosa (retornando un 0). En caso contrario, la función evalúa cuántos de los cuatro KPIs se cumplen utilizando condiciones lógicas. Por cada KPI cumplido, se suma 1 al contador **kpis_cumplidos**.

```

def calcular_exito(row):

    reglas = kpi_por_residencial.get(row['residencial'])

    if not reglas:

        return 0

    kpis_cumplidos = sum([

        row.get('leads', 0) >= reglas['leads'],

        row.get('mensajes_recibidos', 0) >= reglas['mensajes'],

        row.get('alcance', 0) >= reglas['alcance'],

        row.get('CTR', 0) >= reglas['ctr']

    ])

    return 1 if kpis_cumplidos >= 2 else 0

```

```
df['éxito'] = df.apply(calcular_exito, axis=1)
```

Finalmente, si al menos dos de los cuatro KPIs han sido alcanzados, la campaña se etiqueta como exitosa (**1**), de lo contrario, se etiqueta como no exitosa (**0**). Esta etiqueta binaria se asigna a una nueva columna en el DataFrame llamada **éxito**, utilizando la función **apply()** de **pandas** aplicada a cada fila del conjunto de datos. De esta forma, se obtiene una clasificación automática de todas las campañas históricas, que será utilizada como la variable dependiente para entrenar el modelo predictivo.

3. 5.3 Pre Entrenamiento y Entrenamiento de Modelo Regresión Logística

En esta etapa se construye y entrena un modelo de clasificación binaria, específicamente una regresión logística, cuyo **objetivo** es **predecir si una campaña publicitaria será exitosa o no** con base únicamente en las características previas a su lanzamiento. Primero, se definen las variables categóricas que serán utilizadas como predictores (**cat_features**), las cuales incluyen atributos de segmentación como edad y género, características del anuncio como su formato y tipo de mensaje, así como atributos de contexto como el dispositivo objetivo, el llamado a la acción y la temporada del año.

```
# -----
```

```
# 3. Preparación del modelo
```

```
# -----
```

```
cat_features = [
```

```
    'residencial', 'segmentación_edad', 'segmentación_género',
```

```
    'formato_anuncio', 'tipo_mensaje', 'call_to_action',
```

```
    'dispositivo_objetivo', 'temporada'
```

```
]
```

```
df_model = df.dropna(subset=cat_features + ['éxito'])
```

A continuación, se filtran del conjunto de datos original (**df**) aquellas filas que no contengan valores nulos en ninguna de las variables predictoras ni en la columna objetivo **éxito**. Esto garantiza que el modelo se entrene únicamente con datos completos. Luego, se aplica una estrategia de balanceo de clases para evitar que el modelo aprenda un sesgo hacia la clase mayoritaria. Para ello, se identifica la clase minoritaria (campañas no exitosas) y se realiza un submuestreo aleatorio de la clase mayoritaria (campañas exitosas) con el mismo número de observaciones, utilizando la función **resample** de **sklearn.utils**. El conjunto balanceado resultante se combina en un nuevo DataFrame llamado **df_balanced**.

```
# Separar clases para balanceo
```

```
minority = df_model[df_model['éxito'] == 0]
```

```
majority = resample(
```

```
    df_model[df_model['éxito'] == 1],
```

```
    replace=False,
```

```
    n_samples=len(minority),
```

```
    random_state=42
```

```
)
```

```
df_balanced = pd.concat([minority, majority])
```

```
X = df_balanced[cat_features]
```

```
y = df_balanced['éxito']
```

Posteriormente, se definen las variables independientes (**X**) y la variable objetivo (**y**) a partir del conjunto balanceado, y se realiza una partición en

conjunto de entrenamiento y prueba con una proporción de 70/30, respetando la aleatoriedad mediante una semilla fija (**random_state=42**).

```
# Split train/test
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42)
```

El preprocesamiento y el modelo se integran en un solo flujo mediante el uso de **Pipeline** de **sklearn.pipeline**. En este flujo se incluye una transformación de variables categóricas a variables numéricas mediante **OneHotEncoder**, que convierte cada categoría en una columna binaria, ignorando cualquier valor desconocido que pudiera aparecer en la predicción futura. Finalmente, el modelo de regresión logística es entrenado con este flujo sobre los datos de entrenamiento. El parámetro **class_weight= 'balanced'** se especifica para que el modelo compense automáticamente el desequilibrio original de clases al ajustar los pesos internos, reforzando su capacidad para detectar campañas exitosas y no exitosas con igual relevancia.

```
# Pipeline con codificador y regresión logística
```

```
preprocessor = ColumnTransformer([\n\n    ('cat', OneHotEncoder(sparse_output=False, handle_unknown='ignore'),\n    cat_features)\n\n])
```

```
model_balanced = Pipeline([\n\n    ('preprocessor', preprocessor),\n\n    ('classifier', LogisticRegression(max_iter=1000, class_weight='balanced'))\n\n])
```

```
# Entrenamiento
```

```
model_balanced.fit(X_train, y_train)
```

4. 5.4 Evaluación de Modelo

Una vez entrenado el modelo de regresión logística, es esencial validar su desempeño utilizando el conjunto de prueba previamente separado. Para ello, se generan las predicciones (**y_pred**) con los datos de prueba (**X_test**) y se comparan con las etiquetas verdaderas (**y_test**). Este proceso permite determinar qué tan bien generaliza el modelo a datos no vistos.

Como primera métrica de evaluación se presenta la **matriz de confusión**, la cual permite visualizar el número de verdaderos positivos, verdaderos negativos, falsos positivos y falsos negativos. Esta matriz resulta clave para entender los tipos de errores que comete el modelo, especialmente en contextos donde un tipo de error (como predecir una campaña no exitosa cuando en realidad lo es) puede tener mayores consecuencias estratégicas.

Además, se genera el **informe de clasificación** mediante **classification_report**, el cual resume métricas fundamentales como la **precisión**, **recuperación (recall)**, **f1-score** y **exactitud (accuracy)** para cada una de las clases. Estas métricas ofrecen una visión más completa del rendimiento del modelo, permitiendo evaluar no sólo cuántas predicciones fueron correctas, sino también cómo de equilibrado es el rendimiento entre las clases. Esta etapa es crucial para verificar que el modelo no esté sesgado hacia una sola clase debido al desequilibrio original de los datos y que puede identificar de forma confiable campañas tanto exitosas como no exitosas.

```
# -----
```

```
# 4. Evaluación del modelo
```

```
# -----
```

```

y_pred = model_balanced.predict(X_test)

print(confusion_matrix(y_test, y_pred))

print(classification_report(y_test, y_pred))

```

Resultado de Código:

```

[[111 105]
 [111 128]]

```

	precision	recall	f1-score	support
0	0.50	0.51	0.51	216
1	0.55	0.54	0.54	239
accuracy			0.53	455
macro avg	0.52	0.52	0.52	455
weighted avg	0.53	0.53	0.53	455

Figura 1.1

Los resultados obtenidos en la etapa de evaluación se ilustran en la imagen correspondiente al reporte de métricas de desempeño. En la **matriz de confusión (Figura 1.1)**, se observa que el modelo clasificó correctamente 111 campañas exitosas (valor “1”) y 111 campañas no exitosas (valor “0”), mientras que cometió 105 y 128 errores, respectivamente, al confundir campañas entre ambas clases. Este balance sugiere que, a pesar de haber entrenado un modelo en datos balanceados, aún persisten ciertos errores en la clasificación.

En el **reporte de clasificación**, el modelo alcanza una **precisión del 50% para la clase 0** (no éxito) y **55% para la clase 1** (éxito), lo cual significa que más de la mitad de las predicciones positivas fueron correctas, pero con margen de mejora. Los valores de **recall** (capacidad para identificar correctamente los positivos reales) son similares: 0.51 para clase 0 y 0.54 para clase 1. El **F1-score**, que equilibra precisión y recall, también se mantiene en torno al 0.52–0.54, lo que indica un rendimiento aceptable pero modesto. Finalmente, la **exactitud general (accuracy)** es del 53%, lo que refleja la proporción global de predicciones correctas sobre el total de observaciones.

Estos resultados sugieren que el modelo tiene un desempeño ligeramente superior al azar, pero puede optimizarse aún más. Esto puede implicar ajustes en el preprocesamiento, exploración de nuevos modelos, o incorporación de variables adicionales que mejoren su capacidad predictiva.

5. 5.5 Creación de Diccionario de Categorías

Este bloque de código tiene como objetivo construir un diccionario de categorías válidas para cada variable categórica utilizada en el modelo predictivo, lo cual es crucial para validar nuevas campañas antes de su evaluación o predicción.

Primero, se define la lista **features_previas**, que contiene los nombres de las variables categóricas que fueron incluidas en la preparación del modelo: residencial, segmentación de edad, género, formato del anuncio, tipo de mensaje, llamada a la acción, dispositivo objetivo y temporada. Estas son las variables clave que se codificaron mediante **OneHotEncoder** y que influyen directamente la predicción del modelo.

```
# -----  
  
# 5. Diccionario de Categorías  
  
# -----  
  
# Revisar diccionario de categorías válidas para cada variable previa al  
lanzamiento  
  
features_previas = [  
    'residencial',  
    'segmentación_edad',  
    'segmentación_género',
```

```

'formato_anuncio',

'tipo_mensaje',

'call_to_action',

'dispositivo_objetivo',

'temporada'

]

```

Posteriormente, se construye el diccionario **categorias_por_variable** utilizando comprensión de diccionarios. Este recorre cada columna en **features_previas** y extrae todas las categorías únicas presentes en esa variable dentro del DataFrame **df**. El propósito de esto es capturar los valores válidos previamente observados en los datos originales.

```

categorias_por_variable = {col: list(df[col].unique()) for col in features_previas}

```

Finalmente, se imprime una salida organizada (**Figura 1.2**) , mostrando en consola una lista clara de las categorías aceptadas para cada variable. Este paso es útil como referencia para el analista o como validación previa al introducir una nueva campaña en el modelo: garantiza que los valores propuestos estén dentro del rango permitido y no provoquen errores durante la codificación o predicción.

```

print(" 📖 Diccionario de categorías válidas por variable:\n")

for var, categorias in categorias_por_variable.items():

    print(f"{var}:")

    for cat in categorias:

        print(f" - {cat}")

    print()

```

Figura 1.2

■ Diccionario de categorías válidas por variable:

```
residencial:  
- ADARA  
- IRIA  
- ANTEA  
- REFORMA  
- PORTAGO  
- CONDESA
```

```
segmentación_edad:  
- 45-54  
- 55+  
- 35-44  
- 25-34  
- 18-24
```

```
segmentación_género:  
- Ambos  
- Hombre  
- Mujer
```

```
formato_anuncio:  
- Carrusel  
- Video  
- Imagen
```

```
tipo_mensaje:  
- Economico  
- Familiar  
- Emocional
```

```
call_to_action:  
- Mas Informacion  
- Registrarse  
- Enviar mensaje
```

```
dispositivo_objetivo:  
- Ambos  
- Celular  
- Computadora
```

```
temporada:  
- Primavera  
- Invierno  
- Otoño  
- Verano
```

6. 5.6 Creando Campañas Nuevas

Este bloque de código está diseñado para realizar la predicción del éxito de una nueva campaña publicitaria utilizando el modelo de clasificación previamente entrenado. Esta etapa pone a prueba la utilidad del modelo en un escenario real o simulado, proporcionando una evaluación cuantitativa sobre si la campaña tiene o no alta probabilidad de éxito.

Primero, se crea un nuevo **DataFrame** llamado **nueva_campaña**, que contiene los valores seleccionados para cada una de las variables

categorías que el modelo considera: residencial, edad, género, tipo de formato de anuncio, tipo de mensaje, tipo de llamada a la acción, dispositivo objetivo y temporada. Este conjunto representa una configuración realista o hipotética de una campaña que se desea evaluar.

```
# -----  
  
# 6. Predicción de Datos Prueba  
  
# -----  
  
nueva_campaña = pd.DataFrame({  
    'residencial': 'ANTEA',  
    'segmentación_edad': '25-34',  
    'segmentación_género': 'Mujer',  
    'formato_anuncio': 'Video',  
    'tipo_mensaje': 'Familiar',  
    'call_to_action': 'Registrarse',  
    'dispositivo_objetivo': 'Celular',  
    'temporada': 'Invierno'  
})
```

Luego, se utiliza el método `.predict_proba()` del modelo `model_balanced` para calcular la **probabilidad de éxito**, específicamente accediendo al valor correspondiente a la clase positiva (`[0][1]`, es decir, la probabilidad de clase 1 o éxito). Esta probabilidad se almacena en `proba` y se convierte a porcentaje multiplicándola por 100 en etapas posteriores si se desea.

Simultáneamente, se usa `.predict()` para obtener una **clasificación binaria** (éxito o no) basada en el umbral por defecto del modelo (típicamente 0.5). Este valor se almacena en `prediccion`, que es un 1 si el modelo predice éxito y 0 en caso contrario.

```
proba = model_balanced.predict_proba(nueva_campaña)[0][1]
```

```
prediccion = model_balanced.predict(nueva_campaña)[0]
```

Finalmente, se imprime un mensaje indicando si la campaña sería exitosa o no (**Figura 1.3**), acompañado de la probabilidad exacta que el modelo le asignó a dicha predicción. Esta presentación es especialmente útil para la toma de decisiones dentro del área de marketing, ya que permite interpretar con claridad no solo la decisión del modelo, sino también la **confianza asociada** a esa predicción.

```
print(f"\n¿Será exitosa? {'Sí' if prediccion else 'No'} (probabilidad: {proba:.2%})")
```

Figura 1.3

```
¿Será exitosa? Sí (probabilidad: 51.38%)
```

7. 5.7 Justificación de Resultados

Este bloque de código se titula Justificación de Resultados y representa una parte fundamental del análisis interpretativo en un proyecto de ciencia de datos aplicado al marketing. No solo se limita a predecir si una campaña tendrá éxito, sino que va más allá: explica por qué el modelo llegó a esa conclusión. A continuación se desarrolla detalladamente cada paso que se ejecuta. Primero, se transforma la campaña nueva (`nueva_campaña`) con el OneHotEncoder (`ohc`), lo que convierte las variables categóricas en una matriz binaria de presencia/ausencia para cada posible categoría conocida

por el modelo. Esto se guarda como **variables_codificadas**, y sus columnas representan las variables dummy (por ejemplo: **temporada_Invierno**, **formato_anuncio_Video**, etc.).

```
# -----
```

```
# 7. Justificación de Resultados
```

```
# -----
```

Aquí se aplica el codificador OneHotEncoder (**ohe**) a la campaña nueva (**nueva_campaña**). Esto convierte todas las variables categóricas en una representación binaria (one-hot), necesaria para que el modelo pueda interpretarlas correctamente.

```
transformada = ohe.transform(nueva_campaña)
```

El resultado de la transformación se convierte en un **DataFrame** con los nombres de las columnas codificadas (**feature_names**). Esto permite visualizar fácilmente qué variables dummy están activas para esta campaña.

```
variables_codificadas=pd.DataFrame(transformada,columns=feature_names)
```

Se calcula el impacto que tiene cada variable activa codificada multiplicando sus valores binarios por los coeficientes aprendidos por el modelo (**coefs**). Este producto nos da una medida del peso o influencia de cada variable en la predicción

```
importancia_nueva = variables_codificadas.iloc[0] * coefs
```

Se organiza esta información en un nuevo **DataFrame** con tres columnas: el nombre de la variable dummy, su valor (0 o 1), y el impacto correspondiente. Este **DataFrame** será la base del análisis interpretativo.

```
importancia_df = pd.DataFrame({  
    'Variable': variables_codificadas.columns,  
  
    'Valor': variables_codificadas.iloc[0],
```

```
'Impacto': importancia_nueva
})

# Mostrar variables activas y las inactivas que restan éxito (excluyendo
residencial)
```

Aquí se filtran solo las variables **activas** (con valor 1) que **no sean de tipo residencial**, ya que esta última se excluye del análisis para evitar distorsiones (por tratarse de una variable fija para esta campaña).

```
activas = importancia_df[(importancia_df['Valor'] == 1) &
(~importancia_df['Variable'].str.startswith("residencial_"))]
```

Se combina la lista de variables activas con **inactivas_negativas** (que debería contener las variables inactivas que tienen impacto negativo), y luego se ordena el conjunto por impacto descendente. Esto permite identificar tanto las variables que suman como las que restan al éxito.

```
impacto_total_relevante = pd.concat([activas, inactivas_negativas])

impacto_total_relevante = impacto_total_relevante.sort_values(by='Impacto',
ascending=False)
```

Se imprime en pantalla la tabla ordenada con todos los impactos para facilitar su interpretación. Se usa **display()** para visualizarla de forma bonita en Jupyter Notebook, y se formatean los valores de impacto con seis decimales.

```
from IPython.display import display

print("\n VARIABLES QUE IMPACTAN EN LA CAMPAÑA")

display(impacto_total_relevante.style.format({"Impacto": "{:.6f}"))

# Variable con menor impacto entre las activas
```

Aquí se identifica cuál es la variable activa que tuvo el menor impacto positivo. Esta información es clave para identificar qué atributo podría reemplazarse o mejorarse en una futura versión de la campaña.

```
impacto_minimo = activas.sort_values(by='Impacto').iloc[0]
```

```
print(f"\nLa variable activa con menor impacto es:  
{impacto_minimo['Variable']} con un impacto de  
{impacto_minimo['Impacto']:.6f}")
```

```
# Mostrar impacto actual
```

Se recalcula la probabilidad total de éxito, esta vez con el modelo ya entrenado, para colocarla al final del análisis como conclusión de lo que se ha interpretado.

```
prob_exito = model_balanced.predict_proba(nueva_campaña)[0][1] * 100
```

```
print(f"\nActualmente, la campaña tiene una probabilidad de éxito de  
{prob_exito:.2f}%.") (Figura 1.4)
```

Figura 1.4

VARIABLES QUE IMPACTAN EN LA CAMPAÑA			
	Variable	Valor	Impacto
temporada_Invierno	temporada_Invierno	1.000000	0.165967
tipo_mensaje_Familiar	tipo_mensaje_Familiar	1.000000	0.107071
segmentación_edad_25-34	segmentación_edad_25-34	1.000000	0.054906
call_to_action_Registrarse	call_to_action_Registrarse	1.000000	0.042037
formato_anuncio_Video	formato_anuncio_Video	1.000000	0.041026
dispositivo_objetivo_Celular	dispositivo_objetivo_Celular	1.000000	0.015149
segmentación_género_Mujer	segmentación_género_Mujer	1.000000	0.011024

La variable activa con menor impacto es: segmentación_género_Mujer con un impacto de 0.011024
Actualmente, la campaña tiene una probabilidad de éxito de 51.38%.

8. 5.8 Visualización tipo Anillo de Resultados de Modelo

Esta sección del código tiene como propósito visualizar de manera gráfica y atractiva cuáles son las variables activas de la campaña que contribuyen al éxito, y qué tanto impacto tienen en la predicción final.

```
# -----
```

8. Visualización tipo anillo del impacto de las variables

Primero, se extraen las **etiquetas (labels)** correspondientes a las variables activas y los valores de su **impacto (sizes)**, los cuales han sido previamente calculados en el análisis de la predicción. La suma de esos impactos representa el total de aportación positiva de las variables al éxito de la campaña.

```
import matplotlib.pyplot as plt
labels = activas['Variable']
sizes = activas['Impacto']
total = sizes.sum()
```

Luego, se define una paleta de colores degradada en tonos verdes usando **matplotlib.cm.Greens**, lo que ayuda a distinguir visualmente cuáles variables tienen mayor peso (color más oscuro) y cuáles menor (color más claro). Este diseño también comunica de forma intuitiva la jerarquía de impacto.

```
# Colores automáticos
cmap = mpl.cm.Greens
colors = [cmap(i / (len(sizes) - 1)) for i in range(len(sizes))]
```

El gráfico que se genera es de tipo **anillo (Figura 1.5)**, que es una variante del gráfico de pastel (pie chart), pero con un círculo hueco en el centro. En ese espacio central se muestra la **probabilidad final de éxito** de la campaña como un porcentaje, lo cual permite al analista ver en una sola imagen tanto el resultado global como las variables que más lo impulsaron.

```
# Crear el gráfico de anillo
fig, ax = plt.subplots(figsize=(5, 5))
wedges, texts = ax.pie(sizes, labels=labels, startangle=90, colors=colors,
wedgeprops=dict(width=0.35, edgecolor='w'))
```

```
# Texto en el centro con la probabilidad de éxito
ax.text(0, 0, f'{prob_exito:.1f}%', ha='center', va='center', fontsize=17,
fontweight='bold')
```

```
ax.set_title('Variables que contribuyen al éxito de la nueva campaña',
fontsize=10)
```

Finalmente, se ajusta el diseño para presentación (**tight_layout**) y se muestra la figura. Esta visualización no solo cumple una función descriptiva, sino también explicativa: permite ver con claridad qué factores fueron decisivos en el pronóstico de éxito, lo cual es muy valioso para la toma de decisiones estratégicas en marketing.

```
plt.tight_layout()
```

```
plt.show()
```

Figura 1.5



9. 5.9 Visualización Mapa de Calor, Correlación entre Variables

Esta sección del código tiene como objetivo **visualizar de forma global y comparativa cómo diferentes configuraciones de variables afectan la probabilidad de éxito de una campaña**, utilizando un **mapa de calor (heatmap)**.

```
# -----
```

9. Mapa de Calor: Impacto de Variables en Campaña

```
# -----
```

Primero, se construye un **DataFrame** de tipo *tabla dinámica* (**pivot_df**) a partir del conjunto de simulaciones ya realizado previamente. En este DataFrame:

- Las **filas representan** las distintas variables que se modificaron (por ejemplo: **formato_anuncio**, **segmentación_edad**, etc.).
- Las **columnas representan** los nuevos valores asignados a esas variables en cada simulación (por ejemplo: **Carrusel**, **Imagen**, **Ambos**, etc.).
- El **valor en cada celda** es la **probabilidad de éxito** que el modelo estimó para esa configuración específica.

Luego, se genera el **heatmap** (**Figura 1.6**) utilizando la biblioteca **seaborn**. Este tipo de gráfico utiliza color para representar la magnitud de los valores (en este caso, probabilidades de éxito). La paleta **Greens** hace que las configuraciones con mayor impacto positivo se vean más intensamente verdes, y aquellas con menor impacto, más pálidas.

```
import seaborn as sns

pivot_df = df_simulaciones.pivot(index="Variable_modificada",
                                  columns="Nueva_configuracion",
                                  values="Probabilidad_exito")

plt.figure(figsize=(10, 6))

sns.heatmap(pivot_df, annot=True, cmap="Greens", fmt=".1f")

plt.title("Mapa de Calor: Impacto de Cambios en la Campaña")

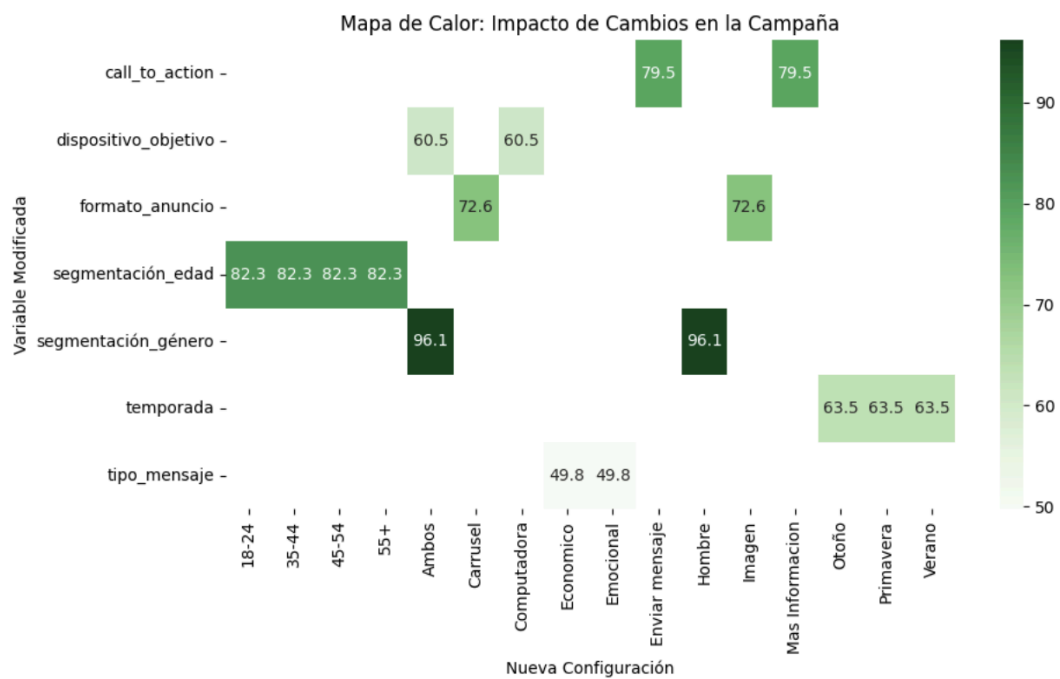
plt.ylabel("Variable Modificada")
```

```
plt.xlabel("Nueva Configuración")
```

```
plt.tight_layout()
```

```
plt.show()
```

Figura 1.6



Este gráfico es muy valioso para un analista porque:

Permite comparar rápidamente **qué cambios concretos** en variables específicas podrían mejorar (o no afectar tanto) el rendimiento de una campaña. Ayuda a **identificar patrones** de combinaciones exitosas en distintas dimensiones. Facilita la **toma de decisiones estratégicas** al sugerir ajustes de alto impacto.

10.5.10 Recomendaciones del Modelo en Prueba

Esta sección del código tiene como propósito **simular el resultado final** de la campaña que previamente simulamos después de aplicar cambios estratégicos sugeridos previamente por el modelo.

```
# -----
```

```
# 10. Cambio Aplicado, ¿Resultado Exitoso?
```

```
# -----
```

Primero, se define una nueva campaña dentro de un **DataFrame** con características ajustadas. Estas características son el resultado de simulaciones anteriores, que identificaron qué combinaciones de valores podrían mejorar la probabilidad de éxito (**Figura 1.7**). Por ejemplo, se elige una audiencia más amplia (**Ambos** géneros), un grupo etario distinto (**55+**), o un formato que tuvo mayor impacto (**Video**). Todos estos cambios fueron seleccionados por su potencial para aumentar el rendimiento de la campaña.

```
nueva_campaña = pd.DataFrame([  
  
    'residencial': 'ANTEA',  
  
    'segmentación_edad': '55+',  
  
    'segmentación_género': 'Ambos',  
  
    'formato_anuncio': 'Video',  
  
    'tipo_mensaje': 'Familiar',  
  
    'call_to_action': 'Registrarse',  
  
    'dispositivo_objetivo': 'Celular',  
  
    'temporada': 'Invierno'  
  
    ])
```

Luego, se utiliza el modelo de regresión logística entrenado (**model_balanced**) para hacer dos predicciones clave sobre esta campaña:

predict_proba estima la **probabilidad de éxito** en un rango de 0 a 1. Este valor se convierte a porcentaje para facilitar la interpretación.

predict devuelve una **clasificación binaria** (éxito o no éxito), basada en si la probabilidad supera el umbral por defecto (generalmente 0.5).

```
proba = model_balanced.predict_proba(nueva_campaña)[0][1]
```

```
prediccion = model_balanced.predict(nueva_campaña)[0]
```

```
print(f"\n¿Será exitosa? {'Sí' if prediccion else 'No'} (probabilidad: {proba:.2%})")
```

El resultado a el código fue el siguiente:

Figura 1.7

```
¿Será exitosa? Sí (probabilidad: 53.70%)
```

Esta respuesta es útil porque ofrece al usuario una confirmación clara y directa de si los cambios propuestos efectivamente transformaron la campaña en una opción viable. En otras palabras, se responde a la pregunta crítica: *¿valió la pena modificar mi campaña?* Este paso es clave para tomar decisiones en entornos reales, pues no solo diagnostica qué no funciona, sino que evalúa en tiempo real si las acciones correctivas fueron eficaces.

6. Conclusiones y Recomendaciones

El modelo predictivo desarrollado tiene un alto potencial estratégico para la toma de decisiones en el área de marketing de EXE Viviendas. Su principal función es anticipar, con base en datos históricos, la probabilidad de que una campaña publicitaria sea exitosa antes de ser lanzada. Esto permite transformar la planeación de campañas en un proceso más basado en evidencia, reduciendo la incertidumbre y maximizando el retorno sobre la inversión publicitaria.

Se recomienda su integración en el flujo de trabajo de planificación de campañas de la siguiente forma:

1. **Evaluación previa al lanzamiento:** antes de ejecutar una campaña, el equipo de marketing puede introducir las características clave (residencial, segmentación, tipo de mensaje, etc.) en el modelo para obtener una predicción del éxito. Esta predicción no solo incluye una probabilidad, sino también una descomposición de cuáles variables contribuyen positiva o negativamente al resultado esperado.
2. **Optimización iterativa de campañas:** al simular distintos escenarios (cambios de formato, mensaje, público objetivo, etc.), el modelo permite identificar combinaciones más prometedoras sin necesidad de experimentar en producción. Esto resulta en campañas mejor dirigidas desde el inicio.
3. **Toma de decisiones informadas por residencial:** dado que el modelo considera los KPIs diferenciados por proyecto inmobiliario, los resultados tienen alta relevancia local. El equipo podrá saber, por ejemplo, qué configuraciones funcionan mejor específicamente en ANTEA o IRIA, evitando generalizaciones que no aplican a todos los casos.
4. **Retroalimentación y mejora continua:** el modelo puede reentrenarse con cada nuevo lote de campañas ejecutadas, fortaleciendo su precisión a medida que se acumulan datos. Este aprendizaje continuo permite adaptar las estrategias al comportamiento cambiante del mercado digital.

5. **Visualización e interpretación ejecutiva:** las salidas del modelo están acompañadas de visualizaciones intuitivas (anillos de impacto, mapas de calor, etc.) que permiten a tomadores de decisiones no técnicos comprender fácilmente qué variables están detrás de una predicción. Esto facilita la comunicación entre los equipos de marketing, análisis y dirección.

El modelo no debe verse solo como una herramienta técnica, sino como un **asistente de planeación estratégica**, que potencia la inteligencia comercial de EXE Viviendas. Implementarlo de forma sistemática puede llevar a una mejora sostenible en la eficacia de las campañas y en el uso inteligente del presupuesto publicitario.

Por otro lado, el desarrollo de este modelo predictivo representó un proceso integral que combinó análisis de datos, lógica de negocio y aprendizaje automático para abordar un reto real del entorno corporativo: anticipar el éxito de campañas publicitarias antes de ser lanzadas. A través de este proyecto, no solo se logró construir un modelo robusto y explicable, sino que también se generaron herramientas visuales, simulaciones y recomendaciones estratégicas que agregan un valor tangible a la toma de decisiones en marketing digital.

Desde una perspectiva personal, este proyecto me permitió aplicar conocimientos teóricos de machine learning en un caso práctico, donde fue esencial comprender la lógica detrás de los indicadores clave de rendimiento (KPI), el tratamiento de variables categóricas, y la importancia de evaluar modelos más allá de la simple precisión, incluyendo interpretabilidad, impacto de variables y simulaciones de escenarios. Aprendí cómo transformar un conjunto de datos crudo en una solución funcional, validada y alineada con los objetivos de negocio.

En términos de valor profesional, el modelo demuestra cómo la ciencia de datos puede servir como un puente entre el análisis técnico y la estrategia comercial. El hecho de que ahora sea posible predecir el desempeño de una campaña antes de su implementación permite reducir riesgos, optimizar presupuesto y replicar prácticas exitosas en futuras ejecuciones. Además, el modelo no solo predice, sino que explica el *por qué* detrás de sus decisiones, fomentando confianza y

adoptabilidad por parte de usuarios no técnicos.

Finalmente, para la empresa, esta solución representa una herramienta de inteligencia comercial que puede integrarse en flujos de trabajo de marketing para generar campañas más eficientes y alineadas con los objetivos del negocio. Su enfoque explicativo, sus simulaciones estratégicas y sus visualizaciones intuitivas permiten que los equipos creativos y ejecutivos trabajen con una visión basada en datos, maximizando el retorno sobre inversión publicitaria.

Anexos

Lista de Tablas

Comparación y justificación del modelo elegido

Tabla 1.1 Compara los principales modelos supervisados utilizados en marketing digital:

Figura 1.2 Ejemplo de interpretación del gráfico de importancia de variables

Figura 1.1 Matriz de confusión

Figura 1.2 Diccionario de categorías válidas por variable

Figura 1.3 Predicción de Datos Prueba

Figura 1.4 Impacto de Variables

Figura 1.5 Visualización Gráfica tipo Anillo

Figura 1.6 Visualización Matriz tipo Heatmap

Figura 1.7 Campaña con mayor probabilidad de éxito

Referencias Bibliográficas

Boulding, W., Staelin, R., Ehret, M., & Johnston, W. J. (2005). *A customer relationship management roadmap: What is known, potential pitfalls, and where to go*. *Journal of Marketing*, 69(4), 155–166. <https://doi.org/10.1509/jmkg.2005.69.4.155>

Buttle, F., & Maklan, S. (2015). *Customer Relationship Management: Concepts and Technologies* (3rd ed.). Routledge.

Chaffey, D., & Ellis-Chadwick, F. (2019). *Digital Marketing* (7th ed.). Pearson Education.

Chen, H., Chiang, R. H. L., & Storey, V. C. (2012). *Business Intelligence and Analytics: From Big Data to Big Impact*. *MIS Quarterly*, 36(4), 1165–1188.

Fernández Loayza, A. (2021). *Predicción de velocidad de ventas en proyectos inmobiliarios utilizando técnicas de machine learning*. Tesis de Maestría, Universidad Peruana de CiKotler, P., Kartajaya, H., & Setiawan, I. (2017). *Marketing 4.0: Moving from traditional to digital*. Wiley.

López-Morales, C., Romero-Torres, L. J., & Ríos-Carmenado, I. (2020). *Diagnóstico del uso del big data y la inteligencia artificial en América Latina: oportunidades y retos*. *Revista Iberoamericana de Ciencia, Tecnología y Sociedad*, 15(45), 33–50.

Marr, B. (2016). *Big Data in Practice: How 45 Successful Companies Used Big Data Analytics to Deliver Extraordinary Results*. Wiley.

OCDE. (2022). *Perspectivas económicas de América Latina 2022: Transición digital para una mejor reconstrucción*. <https://doi.org/10.1787/e2be75be-es>

Payne, A., & Frow, P. (2005). *A Strategic Framework for Customer Relationship Management*. *Journal of Marketing*, 69(4), 167–176. <https://doi.org/10.1509/jmkg.2005.69.4.167>

Power, D. J., Heavin, C., McDermott, J., & Daly, M. (2018). *Data-Based Decision Making and Digital Transformation*. Springer.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know about Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Ryan, D. (2016). *Understanding Digital Marketing: Marketing Strategies for Engaging the Digital Generation* (4th ed.). Kogan Page.

Sharda, R., Delen, D., & Turban, E. (2018). *Business Intelligence, Analytics, and Data Science: A Managerial Perspective* (4th ed.). Pearson.

Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). *Digital marketing and social media: Why bother?* *Business Horizons*, 57(6), 703–708. <https://doi.org/10.1016/j.bushor.2014.07.002>

Turban, E., Sharda, R., & Delen, D. (2020). *Decision Support and Business Intelligence Systems* (11th ed.). Pearson.

Wixom, B. H., & Watson, H. J. (2010). *The BI-Based Organization*. *International Journal of Business Intelligence Research*, 1(1), 13–28. <https://doi.org/10.4018/jbir.2010071702>

Woodcock, N., Stone, M., & Wilson, M. (2005). *Managing the Customer Experience*. Kogan Page.

Witten, I. H., Frank, E., & Hall, M. A. (2016). *Data Mining: Practical Machine Learning Tools and Techniques* (4th ed.). Morgan Kaufmann.

Provost, F., & Fawcett, T. (2013). *Data Science for Business: What You Need to Know About Data Mining and Data-Analytic Thinking*. O'Reilly Media.

Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

Kuhn, M., & Johnson, K. (2013). *Applied Predictive Modeling*. Springer. <https://doi.org/10.1007/978-1-4614-6849-3>

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.).

draft). Stanford University. Recuperado de: <https://web.stanford.edu/~jurafsky/slp3/>

Field, A. (2013). *Discovering Statistics Using IBM SPSS Statistics* (4th ed.). SAGE Publications.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addison-Wesley.

Kelleher, J. D., Mac Namee, B., & D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. MIT Press.