ProspectAI: Modelos de Machine Learning y Preprocesamiento de Lenguaje natural para la Clasificación Efectiva de Clientes

José Fernando García Vidal

Asesor: Miguel Ángel Vargas Valencia

Universidad Nacional Abierta y a Distancia – UNAD

Escuela de Ciencias Básicas, Tecnología e Ingeniería – ECBTI

Especialización en Ciencia de Datos y Analítica

Marzo de 2024

Dedicatoria

Con inmensa gratitud y humildad, dedico este esfuerzo y logro a Dios, quien ha sido mi refugio y fortaleza a lo largo de este desafiante viaje. Su presencia invisible pero palpable me ha brindado la sabiduría y el aliento necesarios para superar cada obstáculo. En los momentos de incertidumbre, ha sido mi guía constante, y en los tiempos de esfuerzo, la fuente de mi perseverancia.

Agradezco sinceramente por cada día de inspiración, por las fuerzas renovadas en los momentos de cansancio, y por la paz en los momentos de turbulencia. Este logro es un testimonio de Su gracia inagotable y de su amor incondicional que me ha permitido desarrollar y completar este proyecto con éxito

Agradecimientos

Quiero comenzar expresando mi profunda gratitud hacia mis abuelos, Beatriz Enríquez y José de la Luz Vidal. Su deseo de verme triunfar académicamente ha sido una inspiración constante y un motivo de profundo agradecimiento.

A mis padres, Marta Rebecca Vidal Enríquez y Luis Fernando García Ordóñez, les debo todo. Han sido los pilares fundamentales de mi vida, brindándome amor incondicional, orientación y apoyo en cada paso de mi camino. Su sacrificio y confianza en mis capacidades me han llevado a alcanzar este logro académico.

A mi tía Susana Mariela Vidal Enríquez, agradezco especialmente por su amor constante y su apoyo resiliente. Su presencia ha sido un pilar de fuerza en mi vida.

Mi gratitud se extiende a mi asesor científico, Miguel Ángel Vargas Valencia, cuya experta guía y dedicación han enriquecido inmensamente mi trabajo. Sus valiosas enseñanzas y consejos han sido fundamentales para el desarrollo de este proyecto.

Finalmente, a todos mis familiares y amigos cercanos que han estado conmigo en cada paso del camino, gracias por sus palabras de aliento y por su compañía, que han hecho que esta travesía académica sea mucho más significativa. Gracias a cada uno de ustedes por ser parte esencial de esta etapa de mi vida y por contribuir de manera significativa a mi desarrollo personal y profesional.

Resumen

Este proyecto investiga cómo optimizar la clasificación de clientes potenciales para MAVV mediante técnicas avanzadas de procesamiento de datos y aprendizaje automático. La investigación comienza con un análisis exploratorio de los datos, seguido de la implementación de técnicas de procesamiento de lenguaje natural (NLP) para extraer características relevantes de los mensajes de los clientes.

Se evaluaron varios modelos de aprendizaje supervisado, incluyendo Regresión Logística, Árboles de Decisión, Bosques Aleatorios, k-Nearest Neighbors y Máquinas de Soporte Vectorial, utilizando métricas como precisión, recall y el puntaje F1. En particular, la Regresión Logística, después de un ajuste meticuloso, alcanzó una precisión de prueba de 0.962, destacando por su eficacia junto con el modelo K-Nearest Neighbors, que logró una precisión impresionante de 0.991.

El ajuste de hiperparámetros fue fundamental para afinar la precisión y el rendimiento de los modelos. Este proceso, apoyado por la validación cruzada, mejoró significativamente su capacidad para generalizar sobre nuevos datos. La Regresión Logística mostró no solo alta precisión sino también una excelente regularización, evidenciada por una destacada curva ROC con un área bajo la curva (AUC) de 0.91.

El estudio destaca la importancia de combinar análisis de datos, NLP y aprendizaje automático para mejorar la clasificación de leads en marketing digital, proporcionando un marco sólido para futuras investigaciones y aplicaciones prácticas en entornos empresariales.

Palabras clave: Aprendizaje automático, Modelos supervisados, Calibración de hiperparámetros, Machine learning, Técnicas de NLP (Procesamiento de Lenguaje Natural)

Abstract

This project investigates how to optimize lead classification for MAVV using advanced data processing and machine learning techniques. The research starts with an exploratory analysis of the data, followed by the implementation of natural language processing (NLP) techniques to extract relevant features from customer messages.

Several supervised learning models, including Logistic Regression, Decision Trees, Random Forests, k-Nearest Neighbors and Support Vector Machines, were evaluated using metrics such as accuracy, recall and F1 score. In particular, Logistic Regression, after meticulous tuning, achieved a test accuracy of 0.962, standing out for its effectiveness along with the K-Nearest Neighbors model, which achieved an impressive accuracy of 0.991.

Hyperparameter tuning was instrumental in refining the accuracy and performance of the models. This process, supported by cross-validation, significantly improved their ability to generalize to new data. Logistic Regression showed not only high accuracy but also excellent regularization, evidenced by an outstanding ROC curve with an area under the curve (AUC) of 0.91. The study highlights the importance of combining data analytics, NLP and machine learning to improve lead classification in digital marketing, providing a solid framework for future research and practical applications in enterprise environments.

***Keywords*:** Machine learning, Supervised models, Hyperparameter calibration, Machine learning, NLP (Natural Language Processing) techniques.

Tabla de Contenidos

Lista de Tablas

Lista de Figuras

Glosario

Aprendizaje Supervisado: Método de aprendizaje automático que utiliza datos etiquetados para entrenar algoritmos, que luego pueden hacer predicciones o clasificaciones sobre nuevos conjuntos de datos no vistos.

Vectorización: Proceso de convertir texto en una representación numérica que los modelos de aprendizaje automático pueden entender y procesar.

CountVectorizer: Técnica de vectorización que convierte un texto en una representación de la frecuencia de palabras dentro de un documento.

TF-IDF (Term Frequency-Inverse Document Frequency): Técnica de vectorización que evalúa la importancia de una palabra en un documento en relación con una colección de documentos o corpus.

Regresión Logística: Modelo de clasificación que estima probabilidades utilizando una función logística.

Árbol de Decisión: Modelo predictivo que mapea características (árboles de decisión) de observaciones sobre un elemento para concluir sobre el valor objetivo.

Random Forest: Método de ensamble que opera construyendo una multitud de árboles de decisión durante el entrenamiento y entregando la clase que es el modo de las clasificaciones de los árboles individuales.

SVM (Support Vector Machines): Clasificador que encuentra el hiperplano en un espacio N-dimensional (N — el número de características) que clasifica claramente los puntos de datos.

K-Nearest Neighbors (KNN): Algoritmo que clasifica un dato basándose en cómo están clasificados sus vecinos más cercanos.

Pipeline: Herramienta de Sklearn que ayuda a secuenciar transformaciones de datos y aplicar modelos de manera coherente y eficiente.

Hiperparámetros: Parámetros de un modelo que se definen antes del aprendizaje y que influyen en la operación del modelo de aprendizaje automático.

Matriz de Confusión: Tabla que se utiliza para describir el rendimiento de un modelo de clasificación en un conjunto de datos para los cuales los valores verdaderos son conocidos.

Precisión: Métrica que mide la proporción de identificaciones positivas que fueron efectivamente correctas.

Recall (Sensibilidad o Exhaustividad): Métrica que mide la capacidad del modelo para identificar todos los casos relevantes dentro de un conjunto de datos.

F1 Score: Promedio armónico de la precisión y la exhaustividad, donde un F1 Score alcanza su mejor valor en 1 (precisión y recall perfectos) y peor en 0.

Introducción

Este proyecto aborda el desafío de mejorar la identificación y clasificación de clientes potenciales para MAVV Smart Optimization, una empresa especializada en brindar servicios de RPA y desarrollo de soluciones tecnológicas, con el objetivo de afinar su proceso de conversión de leads en clientes reales. A lo largo de este trabajo, se despliegan diversas técnicas de análisis de datos y aprendizaje automático, empezando por un exhaustivo análisis exploratorio de los datos disponibles.

Se implementan métodos avanzados de preprocesamiento, donde el lenguaje empleado por los clientes es procesado para extraer la información más relevante. Esto incluye la transformación del texto a través de CountVectorizer y TF-IDFVectorizer, que preparan los datos para ser interpretados por los modelos predictivos.

La esencia del proyecto reside en la selección, implementación y evaluación de una variedad de modelos de aprendizaje supervisado. Cada modelo se somete a un proceso de optimización y se evalúa rigurosamente utilizando métricas como la precisión, el recall y el puntaje F1 para determinar su efectividad en la tarea de clasificación.

La investigación concluye con la selección de un modelo que exhibe un rendimiento destacado, manteniendo un delicado balance entre la precisión de la clasificación y la minimización de errores, particularmente aquellos que podrían llevar a la descalificación errónea de leads valiosos. Este trabajo no solo revela técnicas aplicables para MAVV, sino que también establece un precedente para futuras investigaciones y desarrollos en el campo del aprendizaje automático en marketing y gestión de clientes. Aunque se logran avances significativos, queda

claro que el proyecto es un escalón en un proceso continuo de mejora y exploración en el ámbito de la ciencia de datos.

Planteamiento del problema

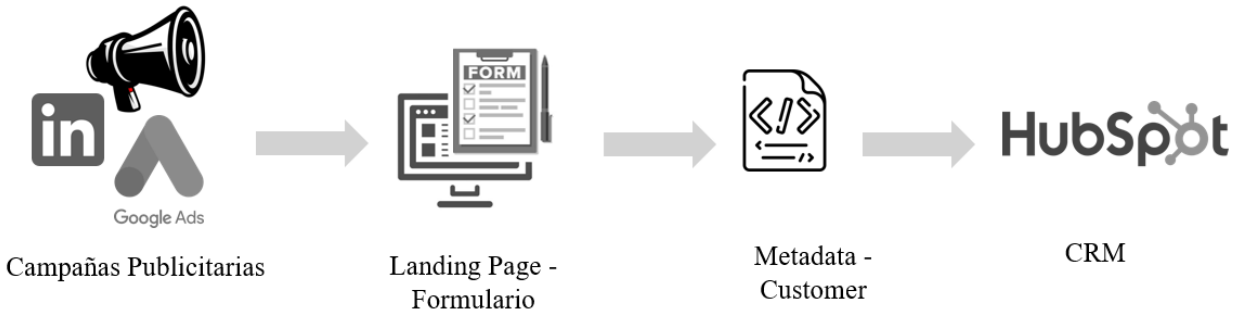
En el mundo actual, saturado de información y con un mercado en constante cambio, la eficacia en la adquisición y clasificación de leads (clientes potenciales) es esencial para cualquier estrategia de marketing exitosa. Los leads, como posibles clientes, representan la base sobre la cual las empresas construyen su crecimiento (Barredo, 2018). Sin embargo, no todos los leads son creados iguales. Aquí es donde entra el concepto de MQL, o "Marketing Qualified Lead". Un MQL se refiere a un lead que ha sido considerado como más probable de convertirse en un cliente en base a ciertos criterios definidos por la empresa (Doyle, 2016). La identificación precisa de un MQL puede aumentar drásticamente la eficiencia de los equipos de ventas y marketing, permitiéndoles concentrar sus esfuerzos en los prospectos con mayor probabilidad de conversión (Järvinen & Taiminen, 2016).

Sin embargo, para la empresa MAVV Smart Optimización, la transición de un prospecto general a un MQL se ha identificado como un punto problemático. Aunque se hacen inversiones sustanciales en campañas publicitarias, no siempre se obtiene el retorno deseado en términos de conversiones efectivas, lo cual tiene un impacto indirecto en el Retorno de Inversión (ROI). Si bien el ROI es una métrica financiera crucial, en este contexto, es sólo un reflejo del desafío subyacente: la correcta clasificación de leads (Chu Rubio, 2020). El proceso típico que sigue MAVV para adquirir un lead comienza con campañas publicitarias, generalmente a través de estrategias de Marketing en Motores de Búsqueda (SEM). Estas campañas dirigen a los posibles clientes a una página de destino o 'Landing Page', donde se les invita a llenar un formulario proporcionando datos y, a menudo, un mensaje. Es en este punto donde reside el desafío: procesar y analizar esta información para determinar qué leads tienen un potencial real para

convertirse en clientes, como se muestra en la Figura 1. Este proceso de análisis es crucial para la identificación eficaz de leads cualificados y representa una oportunidad significativa para la aplicación de un modelo de machine learning especializado en procesamiento de lenguaje natural (NLP).

Figura 1.

Proceso de adquisición de clientes



Fuente. Elaboración Propia

La era de la digitalización ha llevado a que diversas organizaciones y entidades B2B inviertan en plataformas y herramientas analíticas avanzadas (Poecze et. al, 2018). Estas tecnologías les permiten monitorear en tiempo real la trayectoria de sus leads, desde el primer contacto hasta la culminación de una transacción (Sabnis et. al, 2013). El poderoso auge de estas herramientas ha tenido como propósito no solo aumentar la eficiencia de los procesos internos, sino también optimizar las estrategias de marketing y ventas en función de la data recopilada (Siavichay & Eduardo, 2023). Sin embargo, a pesar del notable avance tecnológico y de la disponibilidad de datos, muchas empresas aún enfrentan desafíos en el traspaso de leads a clientes efectivos (Hernández, 2020). Esta brecha sugiere que la adquisición de tecnologías avanzadas, por sí sola, no garantiza una conversión exitosa; más bien, requiere de una estrategia y ejecución adecuadas para lograrlo (Sánchez, 2011).

En este contexto, la inteligencia artificial (IA) se presenta como una herramienta prometedora. Según Bravo (2021), herramientas como IBM Cognos han demostrado la capacidad de la IA para acelerar la preparación y análisis de datos, incrementando la productividad y eficiencia. Este avance es crucial, ya que la IA no solo ahorra tiempo y mejora el rendimiento, sino que también proporciona una comprensión más profunda de los datos, recomendando acciones y prediciendo resultados con mayor eficacia. Además, en una encuesta realizada por Evergage, se reportó que el 61% de los especialistas en marketing experimentaron una mejora en la experiencia del cliente tras la implementación de IA en sus estrategias (Bravo, 2021).

Por lo tanto, el problema central a abordar es: ¿Cómo puede MAVV, mediante la aplicación de técnicas de inteligencia artificial, mejorar la clasificación y priorización de sus leads para identificar de manera más efectiva a aquellos que califican como MQL?

Justificación

El marketing moderno enfrenta un desafío sin precedentes. En un mundo sobresaturado de datos, la capacidad para identificar y priorizar prospectos adecuados conocidos como leads es un componente crucial para garantizar la supervivencia y el crecimiento de las empresas (Tavira & Estrada, 2015). La eficiencia en este proceso no solo garantiza una mejor asignación de recursos, sino que también asegura una mayor conversión, lo que se traduce directamente en mayores ingresos y un retorno de inversión (ROI) optimizado (Turletti, 2018). Para empresas como MAVV, la transición entre un prospecto y un 'Marketing Qualified Lead' (MQL) ha demostrado ser problemática. El dilema no radica en la falta de herramientas tecnológicas o datos, sino en la capacidad de interpretar y utilizar esta información de manera efectiva. La justificación para la implementación de técnicas avanzadas, como la inteligencia artificial, en este proceso radica en su potencial para transformar la manera en que las empresas interpretan y actúan sobre sus datos (Rouhiainen, 2018).

La innovación en marketing no se trata solo de adoptar las últimas tecnologías, sino de entender cómo estas pueden ser aplicadas para resolver problemas reales y entregar valor a las organizaciones (Shevtsova et. al, 2021). Las técnicas de inteligencia artificial, por ejemplo, ofrecen oportunidades sin precedentes para analizar grandes volúmenes de datos y extraer insights que antes eran inalcanzables (Jarek & Mazurek, 2019). Estas técnicas requieren no solo de una inversión tecnológica, sino también de una reconfiguración en la forma en que las organizaciones abordan sus estrategias y operaciones de marketing. La propuesta es crear o mejorar un sistema que permita a MAVV y, potencialmente, a otras empresas, clasificar y priorizar leads con mayor precisión y eficiencia.

Además, este enfoque se alinea con el objetivo de fomentar el desarrollo productivo, tecnológico y social empresarial. Se busca no solo diseñar e implementar sistemas avanzados, sino también asegurar que estos sistemas estén correctamente planificados, dirigidos, controlados, evaluados y realimentados, de acuerdo con las actividades procedimentales de la empresa. La relevancia de este proyecto radica en su potencial para transformar la manera en que las empresas enfrentan uno de sus desafíos más fundamentales. Mediante la mejora del proceso de clasificación y priorización de leads con la aplicación de inteligencia artificial, este proyecto busca optimizar el retorno de inversión (ROI) y sentar las bases para un crecimiento sostenido y eficiente (Sciammarelli, 2023).

Sin embargo, es importante destacar que el alcance de este proyecto se limitará a la *fase de diseño y evaluación de un modelo de inteligencia artificial para la clasificación de leads*, sin incluir la implementación de este modelo en un entorno de producción real. La evaluación del modelo seleccionado como el más eficaz será exhaustiva, pero el proyecto concluirá antes de la fase de producción, En conclusión, este proyecto representa una oportunidad para avanzar en la integración de tecnologías emergentes en el ámbito empresarial, alineándose con los objetivos académicos y prácticos de mejorar la eficiencia y eficacia en la clasificación de leads. La combinación de teoría y práctica, junto con una clara orientación hacia la solución de problemas concretos, asegura que este proyecto sea no solo relevante, sino también esencial en el contexto actual del marketing y la gestión empresarial.

Objetivos

Objetivo general

Entrenar un modelo de machine learning basado en técnicas de procesamiento de lenguaje natural para la clasificación de clientes potenciales.

Objetivos específicos

Realizar un análisis exploratorio de la base de datos "Customer Data" para la identificación de patrones y tendencias que informen sobre el proceso de clasificación efectiva.

Implementar técnicas de NLP para el procesamiento de los datos en los mensajes enviados por los clientes

Entrenar un modelo de machine learning con los datos procesados para la clasificación de los leads, utilizando la información extraída de la base de datos "Customer Data".

Evaluar el rendimiento del modelo, utilizando métricas de machine learning identificando áreas de mejora en su capacidad de clasificación de leads como clientes potenciales (MQL).

Marco conceptual

Figura 2.

Conceptos claves de marketing

Fuente. Elaboración Propia

Marco teórico

El mundo del marketing ha experimentado una revolución sin precedentes con la irrupción de las tecnologías digitales. Si bien el marketing digital es una subcategoría del marketing en general, su impacto y relevancia en el mundo empresarial moderno es innegable. No obstante, la eficacia del marketing digital no radica únicamente en su capacidad para llegar a una audiencia más amplia o diversa; es fundamental clasificar eficientemente a los leads para maximizar el retorno de inversión.

Barredo (2018) introduce al lector en el vasto mundo del marketing digital, subrayando la interconexión entre sus diversos elementos. Más que una simple introducción, este manual destaca la importancia de ver y entender la relación existente entre todos los elementos que conforman el marketing digital, sugiriendo implícitamente la importancia de clasificar y priorizar los leads de manera eficiente para garantizar que cada elemento funcione en conjunto y ofrezca un rendimiento óptimo.

Sin embargo, donde realmente se destaca la importancia de una adecuada gestión y análisis es en la obra de Chu Rubio (2020). Rubio va más allá de la mera conceptualización y se adentra en la rentabilidad de las estrategias de marketing. Enfatiza que las empresas que no miden el retorno de inversión de sus campañas corren el riesgo de desperdiciar recursos valiosos. Una parte integral de este análisis es la clasificación eficiente de leads, ya que dirigir recursos hacia leads mal clasificados o de baja calidad puede resultar en una baja rentabilidad.

Por otro lado, Järvinen & Taiminen (2016) nos ofrecen una visión sobre cómo las empresas B2B están utilizando la automatización del marketing para abordar este desafío. La

generación de leads de alta calidad a través de la personalización de contenidos es una estrategia efectiva. Pero la clave aquí es "alta calidad". La mera generación de leads sin una clasificación y seguimiento adecuados puede no ser beneficiosa. Los autores destacan cómo la automatización puede ser una herramienta crucial en este proceso, garantizando que los leads se clasifiquen, prioricen y gestionen de manera efectiva.

En resumen, el marketing digital, a pesar de sus numerosas ventajas y capacidades ampliadas, presenta desafíos inherentes. Uno de los más significativos es la necesidad de clasificar eficientemente los leads para garantizar un retorno de inversión óptimo. Las fuentes discutidas aquí ofrecen diversas perspectivas y soluciones a este problema, destacando la importancia de la adaptación, la innovación y la estrategia informada en el ámbito del marketing moderno.

A continuación, se encuentra la tabla de resumen del marco teórico:

Tabla 1.

Resumen del Marco teórico

AUTOR Y AÑO	OBJETIVO	PRINCIPALES HALLAZGOS
Barredo, 2018	Explorar la interconexión entre los elementos del marketing digital.	Destaca la importancia de clasificar y priorizar leads para mejorar el rendimiento del marketing digital.
Chu Rubio, 2020	Analizar la rentabilidad de las estrategias de marketing.	Subraya la importancia de medir el ROI y gestionar eficientemente los leads para evitar el desperdicio de recursos.
Järvinen & Taiminen, 2016	Estudiar el uso de la automatización del marketing en empresas B2B.	Enfatiza que la personalización y la calidad de los leads son cruciales para el éxito de la automatización del marketing.

Nota. Esta tabla resume perspectivas académicas clave sobre marketing digital y clasificación de leads, destacando objetivos y hallazgos principales de las fuentes consultadas para enfatizar la necesidad de eficiencia y optimización en las estrategias de marketing.

Metodología

En respuesta a la creciente necesidad de análisis efectivo en el ámbito de grandes volúmenes de datos, este trabajo se centra en la aplicación de la metodología Cross Industry Standard Process for Data Mining (CRISP-DM). Esta metodología, destacada por su eficacia en el estudio de Galán Cortina (2016), se implementará para analizar y extraer valor de los datos. CRISP-DM nos permitirá no solo interpretar los datos a nivel superficial, sino también descubrir patrones y relaciones significativas entre las distintas variables.

Comprensión del Negocio:

Planteamiento del Problema: MAVV Smart Optimization busca mejorar su capacidad para identificar y clasificar eficientemente los leads como MQLs, lo cual es crucial para afinar sus estrategias de marketing y ventas.

Justificación: La implementación de un modelo de machine learning, integrado con técnicas de procesamiento de lenguaje natural (NLP), es fundamental para el crecimiento sostenido de la empresa, alineándose con el objetivo general y los objetivos específicos del proyecto.

Comprensión de los Datos:

Análisis del Dataset: Se analizará el dataset actualizado, que contiene 7077 filas y 16 columnas, incluyendo variables críticas como 'Fecha_creación', 'id_Cliente', 'Pais', 'Mensaje', 'Calificacion', entre otros, para comprender mejor las características y patrones que pueden indicar un MQL.

Preparación de los Datos:

Limpieza y Procesamiento: A partir del dataset inicial, se realizará una limpieza profunda, eliminando filas con valores nulos en columnas críticas y aplicando técnicas de NLP en la columna 'Mensaje' para preparar los datos para el modelado. Esto incluye la eliminación de stopwords y la tokenización.

Modelado:

Se diseñará y entrenará un modelo de machine learning utilizando los datos procesados. Este modelo se centrará en clasificar los leads en MQLs con mayor precisión. Se explorarán y validarán varios algoritmos de aprendizaje automático para encontrar el más adecuado, alineándose con los objetivos específicos del proyecto.

Evaluación:

Los modelos serán evaluados usando métricas como precisión, recall y F1-score, comparando su rendimiento con métodos de clasificación anteriores para validar su eficacia. Esta fase es esencial para asegurar que el modelo cumpla con los objetivos de identificar patrones y tendencias que informen sobre el proceso de clasificación efectiva y su capacidad para clasificar leads como MQLs.

Análisis exploratorio para la identificación de patrones y tendencias en clasificación de leads

Este capítulo se centra en el análisis detallado de los datos recopilados de los clientes potenciales de MAVV Smart Optimization. Al desarrollar este segmento, se ha tenido en cuenta la importancia de comprender a fondo cómo la empresa atrae a nuevos clientes y cómo recoge sus datos. MAVV ofrece servicios especializados en el desarrollo de software y en soluciones que automatizan tareas repetitivas, conocidas como RPA por sus siglas en inglés.

La manera en que la empresa consigue la atención de posibles clientes es a través de anuncios en internet que los llevan a una página específica, conocida como "landing page". En esta página, los clientes interesados dejan su información y un mensaje que a menudo indica cuán interesados están en los servicios de MAVV. Estos mensajes se convierten en un recurso valioso, ya que aportan pistas sobre lo que el cliente busca o necesita. Esta información se registra en un sistema llamado CRM, que es una herramienta que la empresa usa para mantener toda la información de los clientes organizada y accesible.

Es importante resaltar que toda la información que se extrajo para ser analizada en este estudio se ha tratado con cuidado, asegurándonos de respetar la privacidad de los clientes. Se ha trabajado solo con datos que ya están disponibles para el público y que no exponen detalles personales sensibles de los clientes. Esto garantiza que el análisis se realice de manera ética y responsable, respetando los derechos de privacidad de cada persona involucrada.

Tabla 2.

Resumen de Datos de Leads de MAVV

METADATA	DESCRIPCIÓN
Fecha de creación	Fecha y hora en que se agregó cada registro al sistema, útil para análisis de tendencias temporales.
Id Cliente	Identificador numérico único para cada cliente potencial, clave para seguimientos individuales.
País	Origen geográfico del lead, importante para segmentación y análisis regional.
Mensaje	Incluye el contenido textual de los mensajes dejados por los leads, generalmente en formularios o correos electrónicos. Como variable de objeto, puede contener información cualitativa que, mediante procesamiento de lenguaje natural (NLP), revelará intenciones o intereses específicos de los clientes.
Calificación	Estado del lead como 'Calificado' o no, esencial para evaluar la calidad y proximidad al perfil deseado.
Canceló la suscripción a las notificaciones de correo	Indica si el lead optó por no recibir notificaciones por correo, reflejando su interés en mantener comunicación.
Canceló suscripción a todos los correos	Muestra si el lead se des inscribió de todos los correos, proporcionando perspectivas sobre su voluntad de interacción.
Correos de marketing entregados	Total, de correos de marketing exitosamente entregados al lead, indicador de la frecuencia de contacto.
Correos de marketing rebotados	Cuenta los correos de marketing que no fueron recibidos, útil para evaluar problemas de entrega.
Correos electrónicos de marketing respondidos	Número de correos de marketing a los que el cliente respondió, medida directa de interacción.
Correos de marketing con clic	Veces que se hizo clic en enlaces dentro de correos de marketing, indicador temprano de interés en conversión.
Correos de marketing abiertos	Cantidad de correos abiertos por el cliente, reflejo del interés y compromiso con la marca.
Cantidad de Sesiones	Número de interacciones del cliente con el sitio web, útil para entender la frecuencia de compromiso en línea.

Cantidad de Páginas Visitadas	Total, de páginas web visitadas, indicador del nivel de interés y exploración del cliente.
Year	Año de registro del lead, permite identificar tendencias y cambios anuales.
Month	Mes de registro, facilita el análisis de tendencias y efectos estacionales en la adquisición de leads.

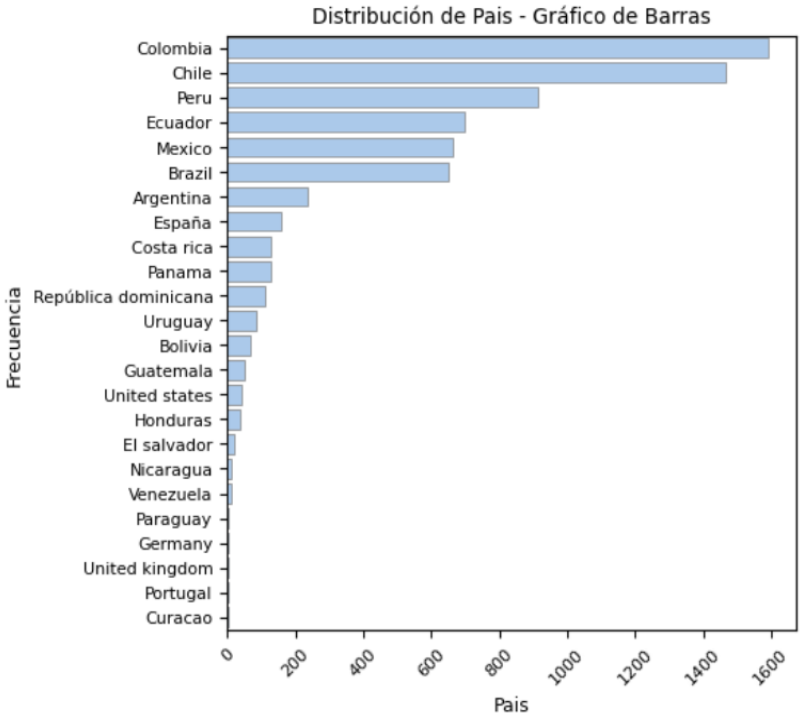
Nota. La tabla proporciona un resumen detallado de las metadata asociadas con cada registro de cliente potencial en la base de datos

El conjunto de datos consiste en 7,077 registros de clientes potenciales con 16 variables, que incluyen datos demográficos, comportamientos de interacción por correo electrónico y actividad web. Entre estas variables, tres son categóricas ‘País', 'Mensaje', 'Calificación’ y el resto son numéricas, proporcionando una matriz de información para el análisis.

Al realizar el análisis exploratorio, emergieron descubrimientos notables, particularmente con relación a los países donde MAVV tiende a atraer a la mayoría de sus clientes. Este análisis detallado de cómo se distribuyen los clientes geográficamente mostró que Colombia, Chile, Perú, Ecuador y México, se destacan como los países con el mayor número de clientes registrados en nuestra base de datos, la cual comprende un total de 7,077 clientes. Estos cinco países, como se ilustra en la figura 3, representan conjuntamente el 75.3% del total de la base de datos. Este hallazgo resalta la significativa influencia que tienen estos mercados en la operación de la empresa y enfatiza su valor estratégico y su potencial como focos principales de adquisición de clientes.

Figura 3.

Distribuciones de país



Fuente. Elaboración Propia

Con respecto a la calificación de clientes, el 54.5% han sido clasificados como 'Calificados', es decir, como Marketing Qualified Leads (MQL). Esto indica que más de la mitad de los prospectos cumplen con los criterios establecidos por la empresa para ser considerados potenciales clientes efectivos. Por otro lado, el 45.5% de los clientes se encuentran en la categoría de 'No Calificados', representando una porción considerable de prospectos que no alcanzan los estándares requeridos para ser MQL y esto se logra evidenciar en la tabla 3.

Tabla 3.

Distribución de Calificaciones de Leads

CALIFICACIÓN	FRECUENCIA ABSOLUTA	FRECUENCIA RELATIVA	PORCENTUAL
Calificado	3859	0.545288	54.528755

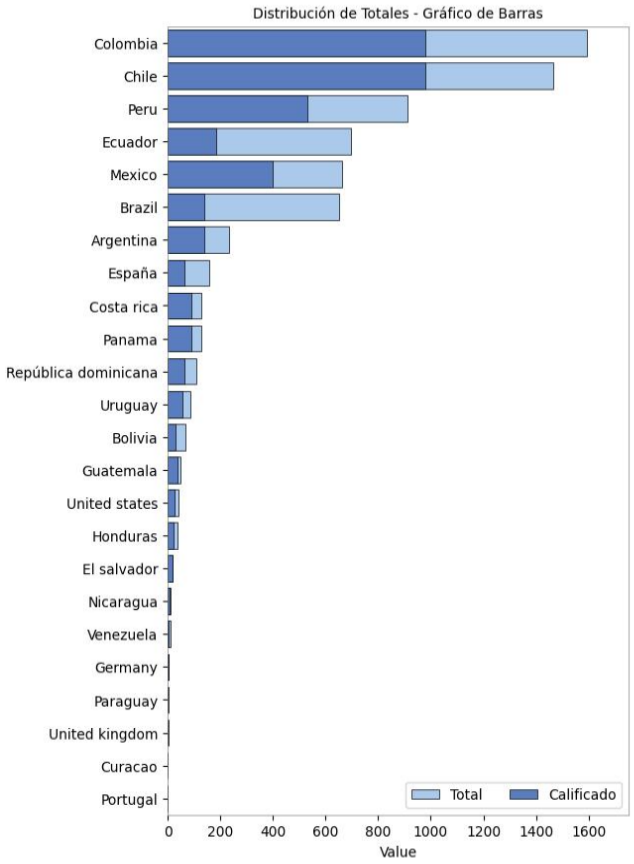
No calificado	3218	0.454712	45.471245

Nota. La Tabla 3 clasifica los leads en 'Calificados' (54.53%) y 'No Calificados' (45.47%), mostrando la proporción de prospectos que cumplen o no con los criterios de selección como clientes potenciales

En la figura 4, se logra apreciar que a pesar de que Ecuador, se encuentre entre los cinco países con más clientes, muestra un alto porcentaje de leads no calificados. Esto sugiere la oportunidad de reevaluar las estrategias de adquisición de clientes y la aplicación de criterios de calificación en esta región. La calidad de los leads y la eficiencia en los procesos de calificación son cruciales para la optimización de la conversión de prospectos en clientes y para un enfoque más efectivo en las campañas de marketing y ventas.

Figura 4.

Distribución y clasificación de leads por país



Pais	Calificado	No calificado	Totales
Colombia	978	615	1593
Chile	979	487	1466
Peru	531	381	912
Ecuador	185	512	697
Mexico	398	266	664
Brazil	139	513	652
Argentina	138	96	234
España	63	95	158
Costa rica	90	39	129
Panama	89	37	126
República dominicana	63	46	109
Uruguay	56	30	86
Bolivia	30	38	68
Guatemala	38	11	49
United states	25	16	41
Honduras	23	16	39
El salvador	17	3	20
Nicaragua	7	5	12
Venezuela	5	6	11
Germany	2	1	3
Paraguay	1	2	3
United kingdom	1	2	3
Curacao	1	0	1
Portugal	0	1	1

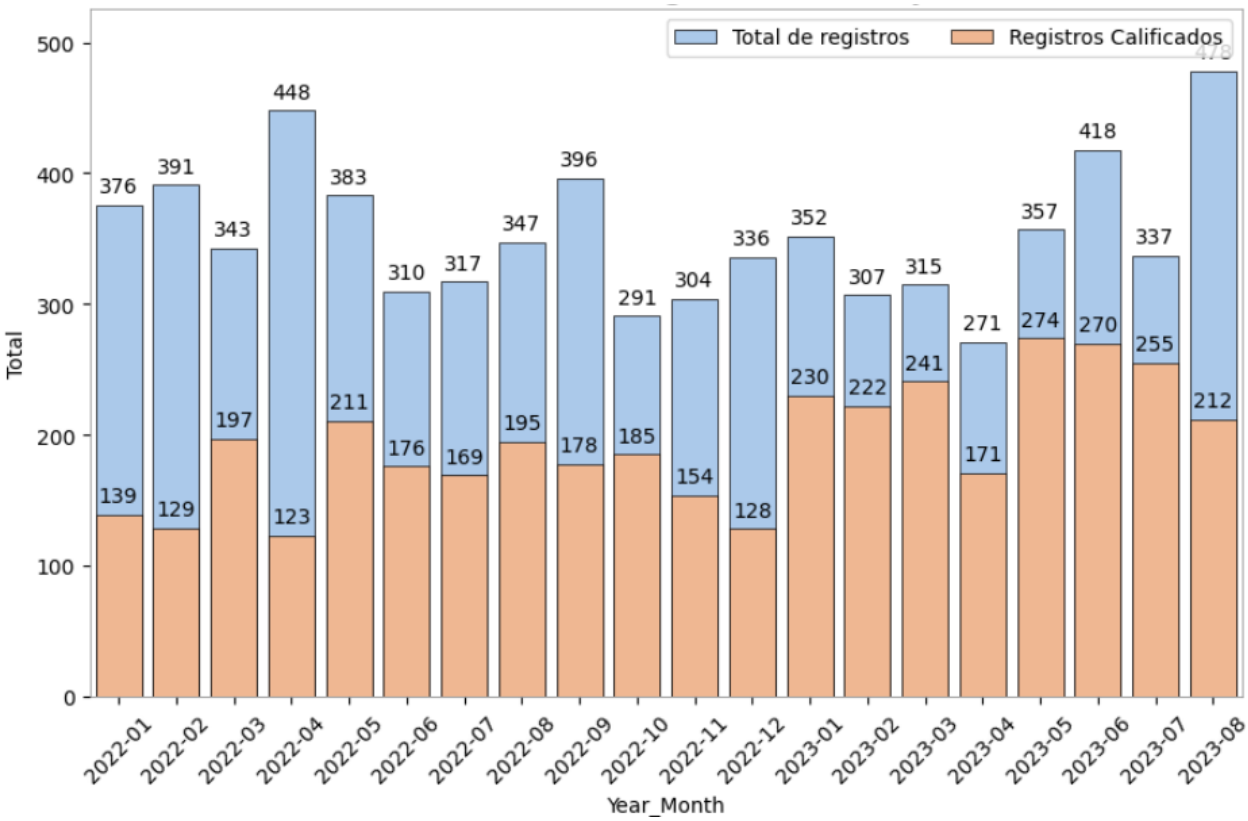
Fuente. Elaboración Propia

Este análisis indica que la implementación de estrategias de marketing personalizadas para cada país podría ser ventajosa, en especial para incrementar la cantidad de leads calificados en mercados como el ecuatoriano. Asimismo, una segmentación más exhaustiva de los clientes calificados podría ser útil para comprender a fondo sus perfiles y cómo estos influyen en la conversión a ventas. Este conocimiento es esencial para el éxito de cualquier estrategia de crecimiento y retención de clientes, siendo un componente clave en la expansión y fidelización dentro del ámbito comercial.

Continuando con esta línea de investigación en el desarrollo de este proyecto, se examinó el comportamiento mensual en la adquisición de clientes. La figura 5, titulada "Distribución Mensual de Registros Totales y Calificados", desglosa la información por mes mostrando el número total de nuevos interesados en el servicio o producto y aquellos clasificados como 'Calificados'. Esta representación gráfica evidencia que una porción significativa de los nuevos registros mensuales cumple con los criterios de calificación. Tal consistencia señala la efectividad de las tácticas empleadas para captar leads propensos a convertirse en clientes activos. Notablemente, desde el año 2022 al 2023, se percibe un crecimiento moderado en el número de clientes, destacando especialmente en meses como marzo y mayo, donde la adquisición de nuevos clientes alcanza picos significativos.

Figura 5.

Distribución Mensual de registros totales y calificados



Fuente. Elaboración Propia

Esta secuencia de observaciones y análisis refuerza la idea de que una estrategia de marketing bien segmentada y adaptada no solo mejora la calidad de los leads captados, sino que también afina el proceso de calificación, asegurando que se identifiquen con precisión aquellos prospectos con un mayor potencial de conversión. La capacidad de la empresa para ajustar y orientar sus esfuerzos de marketing de manera efectiva es crucial para su crecimiento sostenido y la expansión de su base de clientes en el competitivo mercado actual.

Prueba de Hipótesis de variables categóricas

Durante el desarrollo de este proyecto, se optó por realizar una prueba de hipótesis utilizando el método del chi cuadrado (χ^2) con el objetivo de investigar la posible existencia de una relación entre variables específicas: el país de origen y la calificación de los leads, así como el contenido del mensaje y su calificación. La elección de esta técnica estadística se basa en la recomendación de Carl Pearson, quien propuso la prueba de chi cuadrado como un método para evaluar la adecuación de los datos observados a una distribución teórica esperada. Según Hernández de la Rosa et al. (2019), esta se ha convertido en la técnica preferida para el análisis comparativo entre dos o más grupos y variables. Este enfoque permite discernir si variables como el país de origen del lead o el contenido del mensaje influyen en la probabilidad de que un lead sea considerado como calificado.

Una prueba de hipótesis es un procedimiento estadístico que permite tomar una decisión sobre el valor de un parámetro de la población basada en los datos muestrales. En el contexto de este estudio, se realizaron las siguientes pruebas:

Tabla 4.

Tabla de Prueba de Hipótesis - Chi cuadrado

HIPÓTESIS NULA (H0)	VARIABLE 1	VARIABLE 2	CHI-CUADRADO	P-VALOR	DECISIÓN
Ho: Las variables no tienen relación	País	Calificación	712.9	8.4e-36	Rechazar H0
Ho: Las variables no tienen relación	Mensaje	Calificación	7013.7	4.4e-62	Rechazar H0

Nota. Esta tabla resume los resultados de las pruebas de Chi-cuadrado, evaluando la relación entre diferentes variables. Los valores de Chi-cuadrado y p-valor indican que se rechaza la hipótesis nula, demostrando una relación significativa entre las variables analizadas.

Los resultados obtenidos de estas pruebas proporcionan evidencia suficiente para rechazar la hipótesis nula en ambos casos, demostrando así la existencia de una relación significativa tanto entre el país y la calificación del lead, como entre el mensaje proporcionado por el lead y su calificación. Esto indica que tanto el país de origen como el contenido del mensaje tienen un impacto en la calificación de los leads como calificados o no calificados.

Análisis Cuantitativo de la Interacción Cliente-Marca

A medida que profundizamos en el estudio, se realizó un análisis detallado de las interacciones de los usuarios con las campañas de correo electrónico, identificando patrones clave que pueden influir significativamente en la efectividad de estas estrategias de comunicación. Según los datos recopilados y presentados en la Tabla 5, cada cliente, en promedio, ha recibido alrededor de 5.06 correos de marketing, con casos que alcanzan hasta 129 correos enviados. Este amplio rango sugiere la existencia de campañas intensivas o una prolongada duración de suscripción por parte de algunos clientes. Además, se observa que la tasa media de clics es de 1.92 y la de apertura de correos de 2.08, lo cual plantea una oportunidad para mejorar el engagement significativo a través de los contenidos enviados por correo.

Tabla 5.

Resumen Estadístico de Interacciones en Campañas de Marketing por Correo Electrónico

ESTADÍSTICO	CORREOS DE MARKETING ENTREGADOS	CORREOS DE MARKETING CON CLIC	CORREOS DE MARKETING ABIERTOS	CANTIDAD DE SESIONES
Cantidad	7077	7077	7077	7077
Media	5.06	1.92	2.08	2.03

Desviación Estandar	5.09	2.81	2.88	9.20
Mínimo	0	0	0	1
25%	1	0	0	1
50%	4	1	1	1
75%	7	3	3	2
Máximo	129	28	101	461

Nota. Esta tabla presenta estadísticas esenciales de campañas de marketing por correo electrónico, destacando la entrega, interacción y apertura de correos, así como la actividad del sitio web. Refleja la variabilidad en el comportamiento del cliente, subrayando la importancia de adaptar y optimizar continuamente las estrategias de marketing.

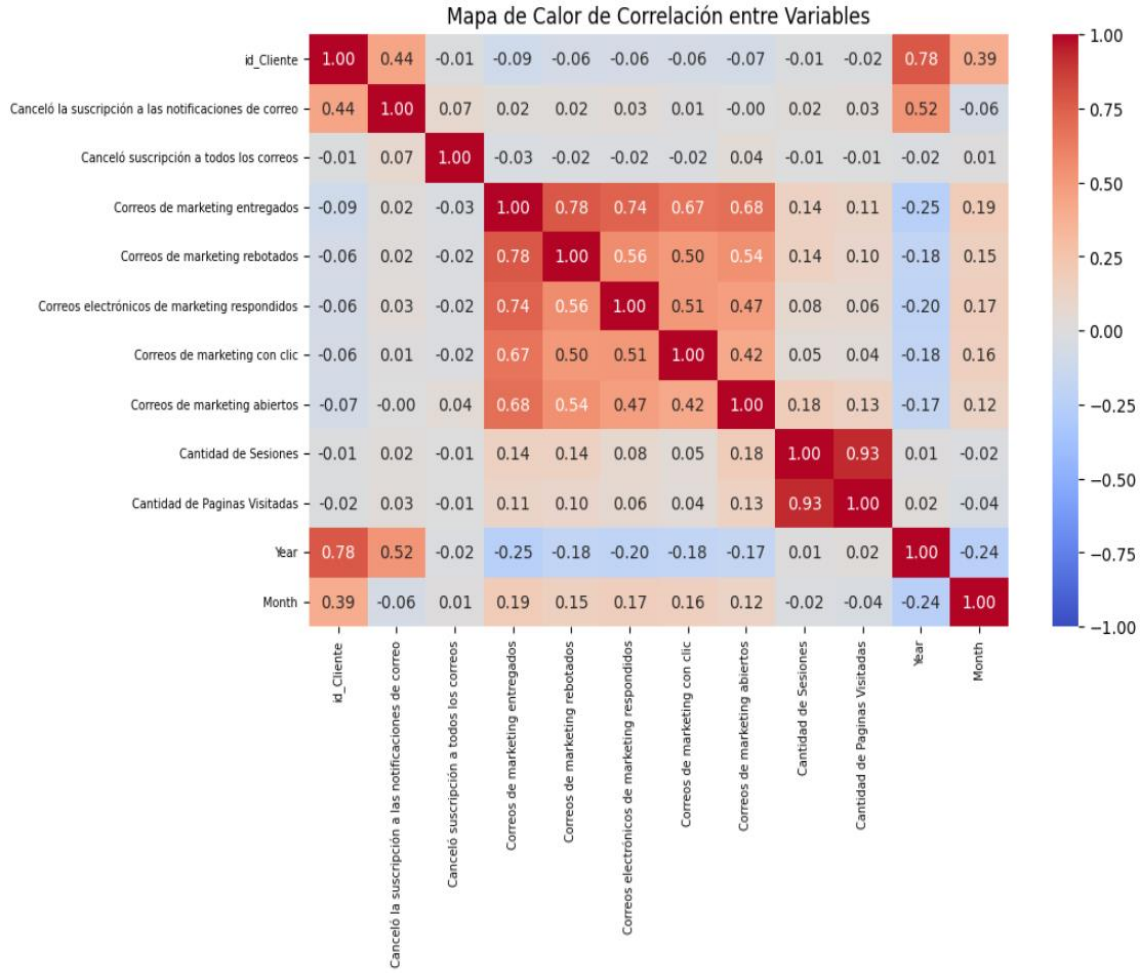
La tasa de apertura promedio se sitúa en el 41.07%, indicando que más de cuatro de cada diez correos son abiertos por los destinatarios, reflejando un nivel de compromiso activo. No obstante, la posibilidad de aumentar la interacción efectiva, específicamente a través de clics en el contenido, sugiere que hay margen para optimizar aún más estas campañas.

La variabilidad observada en el comportamiento de los clientes, como lo indica la desviación estándar en las métricas de correos entregados y sesiones en el sitio web, respalda la necesidad de adoptar una estrategia de marketing más personalizada. Adaptar las campañas para reflejar las preferencias y comportamientos específicos de los clientes no solo podría mejorar la eficacia de estas sino también incrementar el grado de interacción y compromiso. Estos hallazgos, resumidos en la Tabla 5, resaltan la importancia crítica de evaluar y ajustar continuamente las estrategias de marketing por correo electrónico para asegurar que estén alineadas con las preferencias y comportamientos de los clientes. Mediante la optimización del contenido y la personalización de las comunicaciones, se puede aspirar a una mayor participación del cliente, lo cual es esencial para la conversión efectiva de los leads de marketing en ventas confirmadas.

Continuando con el rigor analítico presentado hasta ahora, nos enfocaremos en una evaluación cuantitativa de las interacciones entre los clientes y las campañas de marketing de MAVV Smart Optimization. La Figura 6, nuestro mapa de calor permite examinar la intensidad de la relación entre distintas métricas numéricas de comunicación.

Figura 6.

Mapa de Calor de las Correlaciones entre las variables numéricas



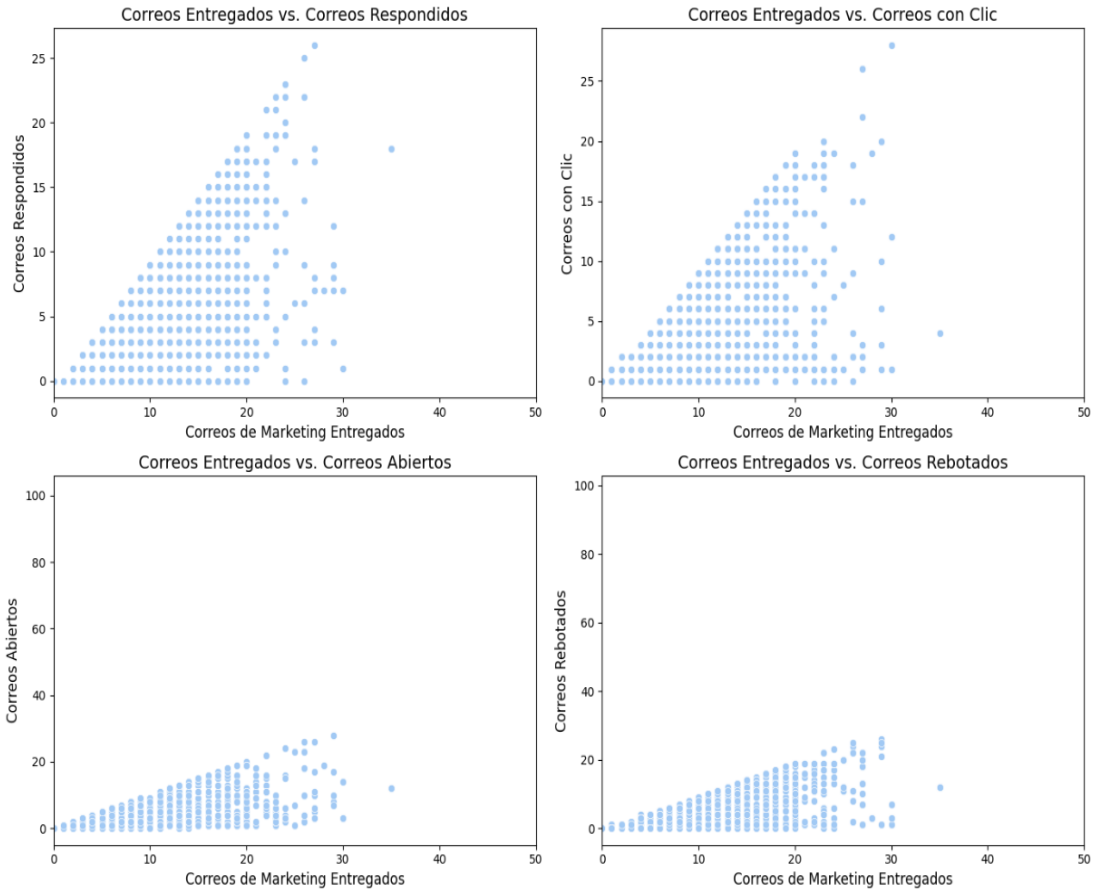
Fuente. Elaboración Propia

Al examinar el mapa de calor, observamos que existe una correlación positiva moderada entre la cantidad de correos entregados y las métricas de correos rebotados, respondidos, abiertos y con clic. Estos valores indican que hay una tendencia en la que un aumento en los correos

entregados está asociado con un incremento en las mencionadas interacciones de los usuarios. No obstante, es importante reconocer que estos resultados no necesariamente establecen una relación causal directa; otros factores no identificados podrían influir en la interacción de los clientes con los correos.

Figura 7.

Análisis de Interacción con los correos de marketing



Fuente. Elaboración Propia

Los diagramas de dispersión de la figura 7, se observa visualmente la mencionada correlación. Los clientes parecen interactuar más con los correos a medida que aumenta la

cantidad recibida. No obstante, la relación no es lineal, y la alta densidad de puntos en los valores más bajos indica una amplia variabilidad en la interacción cliente-correo.

Este análisis cuantitativo también revela una correlación fuerte y directa entre la cantidad de páginas web visitadas y el número de sesiones en el sitio, lo cual es un comportamiento esperado: usuarios que visitan con frecuencia tienden a explorar más contenido, lo que señala un mayor compromiso con la marca.

En síntesis, la Figura 7 proporciona una visión clara de cómo las tácticas de marketing por correo electrónico y la interacción en la web de MAVV se relacionan con el compromiso del cliente. Este análisis cuantitativo soporta la necesidad de una estrategia de marketing que se ajuste a los patrones de comportamiento de los clientes, con el objetivo de profundizar su compromiso y, en última instancia, mejorar las tasas de conversión.

La integración de este análisis con los hallazgos anteriores refuerza la necesidad de una evaluación continua y de la personalización de las estrategias de marketing para que resuenen con eficacia en la base de clientes de la empresa. La optimización basada en datos claros y cuantificables es esencial para ajustar las campañas de marketing y maximizar el rendimiento de la inversión en la adquisición y retención de clientes.

A partir del análisis exhaustivo que se llevó a cabo, queda claro que la ***calificación de los leads, su país de origen y el contenido de los mensajes*** son variables interconectadas que influirán decisivamente en la elaboración del modelo de clasificación. La importancia de integrar las técnicas de procesamiento de lenguaje natural (NLP) para desentrañar el significado de los mensajes de los clientes es fundamental, considerando su potencial para revelar patrones subyacentes en sus intenciones y preferencias.

Reflexionando sobre el análisis:

- La calificación de los leads emerge como un indicador crucial de su potencial para convertirse en clientes activos, lo que justifica su papel central en el modelo.
- La variable geográfica ofrece un contexto esencial para una segmentación más precisa y adaptada a la diversidad de los mercados.
- La riqueza de información contenida en los mensajes de los clientes subraya la necesidad de aplicar NLP para capturar aspectos cualitativos que los datos puros no pueden proporcionar.

Con el terreno preparado por el análisis previo, nos dirigimos ahora a la etapa de procesamiento preliminar de los datos de texto. En la siguiente sección, exploraremos en detalle el proceso de tokenización y cómo esta técnica prepara el camino para una interpretación más profunda de las interacciones de los clientes.

Aplicación de Técnicas de Procesamiento de Lenguaje Natural (NLP) en Mensajes de Clientes

Esta sección del estudio se enfoca en el procesamiento de lenguaje natural (NLP), una técnica que nos permite preparar y analizar texto, en este caso, los mensajes de los clientes, para descifrar su intención subyacente. Es crucial llevar a cabo un preprocesamiento detallado de estos mensajes antes de poder modelarlos, ya que el modelo necesita 'entender' el lenguaje humano de una manera que se alinee con los objetivos del análisis.

El NLP es una disciplina que combina ciencias de la computación y lingüística para permitir que las máquinas comprendan e interpreten el lenguaje humano. En el pasado, los sistemas de NLP se basaban en reglas específicas para el procesamiento de texto. Sin embargo, con el tiempo, se ha transitado hacia enfoques más sofisticados que emplean redes neuronales, las cuales simulan la manera en que los humanos aprendemos y entendemos el lenguaje. Estos sistemas modernos 'aprenden' de grandes cantidades de datos textuales para poder realizar inferencias y reconocer patrones contextuales, en lugar de depender de reglas fijas (Alias & Cassanelli, 2019).

Las tareas que lleva a cabo el NLP abarcan desde el reconocimiento de voz, que transforma palabras habladas en texto, hasta la resolución de correferencias, donde el sistema identifica a qué o a quién se refieren los pronombres en un texto. Estas tareas incluyen el etiquetado de partes del discurso, donde se identifica la función gramatical de cada palabra según su contexto; la desambiguación de palabras con múltiples significados, seleccionando el sentido más adecuado en base al análisis semántico; y el análisis de opinión, que busca extraer matices como la emoción o el sarcasmo de los textos (Kavlakoglu, 2020).

A través de la aplicación de NLP, el objetivo es desvelar los matices y significados ocultos en las comunicaciones de los clientes, lo que nos permite obtener una comprensión más profunda de sus necesidades y percepciones. Con esta herramienta, podemos construir un modelo de clasificación robusto y sensible a las sutilezas del lenguaje natural, que sea capaz de identificar con precisión a los clientes más valiosos para MAVV y mejorar la eficacia de nuestras estrategias de interacción y crecimiento.

Para aplicar las técnicas de procesamiento de lenguaje natural (NLP), se eligió ***Python***, un lenguaje de programación versátil y de alto nivel que destaca por su flexibilidad y amplia adopción en diferentes campos (Llerena, 2020). En este proyecto, se seleccionó la biblioteca Spacy, reconocida por ser una de las herramientas más potentes y utilizadas en NLP junto con NLTK. Desarrollada por Matt Honnibal, Spacy es accesible y de código abierto, con su código disponible en GitHub desde su lanzamiento en 2015. Spacy sobresale por sus características avanzadas:

- Ofrece soporte para más de 70 idiomas.
- Dispone de 80 modelos procesales para 24 idiomas.
- Incluye versiones preentrenadas de BERT, una de las arquitecturas de deep learning más avanzadas, basada en transformers.
- Cuenta con vectores de palabras ya entrenados.
- Realiza una tokenización precisa a nivel lingüístico.

Implementación de Spacy para la eliminación de Stop Words

Con la ayuda de Spacy se realizó el proceso de eliminación de ***palabras vacías*** (Stop Words). Tale, como preposiciones, artículos y conjunciones ("es", "el", "un", "y", "en", "a", entre

otras), a menudo se eliminan del procesamiento lingüístico porque contribuyen poco al significado contextual de las comunicaciones. Estos términos, frecuentes en el lenguaje cotidiano, pueden considerarse como ruido en el análisis de texto debido a su baja carga informativa (Mudasir, 2024).

Al excluir estas palabras de bajo valor semántico, la atención se centra en los términos sustantivos y verbos que llevan el peso de la comunicación, permitiendo que el análisis se enfoque en las palabras que realmente capturan las ideas centrales y los conceptos significativos expresados por los clientes. Esta aproximación mejora la calidad del procesamiento de datos y facilita la extracción de información valiosa de los mensajes, contribuyendo así a la construcción de un modelo analítico más preciso y efectivo

A continuación, en la figura 8, se puede apreciar el proceso de eliminación de Stop Words:

Figura 8.

Proceso de eliminación de palabras vacías

Fuente. Elaboración Propia

Después de concluir la eliminación de palabras vacías se procede a iniciar el proceso de tokenización de los mensajes. Para este proceso se emplea la librería scikit-learn, un marco de trabajo de Python de gran aceptación para la implementación eficiente de algoritmos de machine learning. Este conjunto de herramientas es reconocido por su versatilidad en el análisis predictivo

y su facilidad de uso, lo que lo convierte en una elección predilecta en el campo de la ciencia de datos (Bobadilla, 2021).

Métodos de Tokenización

Dentro de scikit-learn, se seleccionan dos métodos de vectorización: ***CountVectorizer*** y ***TfidfVectorizer***. El primero convierte texto en una matriz de recuentos de tokens, generando una representación matricial que refleja la frecuencia de las palabras (Alvi & Talukder, 2021). El segundo, TfidfVectorizer, construye una matriz de características TF-IDF que valora no solo la frecuencia de las palabras sino también su relevancia en todo el conjunto de datos, ofreciendo una perspectiva más matizada de la importancia del término (Carrera et al, 2008).

La distinción entre ambos métodos reside en su enfoque: mientras CountVectorizer se centra exclusivamente en la frecuencia de aparición de los términos, TfidfVectorizer ajusta esta frecuencia con respecto a la importancia de la palabra en el conjunto completo de documentos, otorgando una puntuación más alta a los términos que son distintivos del documento en cuestión (Shandeep, 2022).

En la fase subsecuente del proyecto, se evaluarán ambos métodos para determinar cuál se alinea mejor con los datos y objetivos específicos de la investigación. Mediante el análisis comparativo basado en métricas de rendimiento apropiadas, se seleccionará el método de vectorización que optimice la eficacia del modelo de clasificación propuesto, facilitando así una interpretación más precisa y detallada de los mensajes de los clientes.

Entrenamiento de un Modelo de Machine Learning para Clasificación Eficiente de Leads

En esta sección, se abordará el proceso de modelado, destacando que, dentro del amplio espectro del machine learning, este proyecto se centra en la aplicación del aprendizaje supervisado. Este enfoque requiere de datos previamente etiquetados para entrenar al modelo, permitiéndole así aprender y hacer predicciones o clasificaciones con nuevos datos no vistos. Los modelos seleccionados para desarrollar dentro de este paradigma incluyen:

Random Forest (Bosque Aleatorio): Mejora y generaliza el método de árboles de decisión mediante la creación de un 'bosque' de múltiples árboles. Esta técnica es valorada por su capacidad para reducir el sobreajuste, un problema común donde el modelo se especializa demasiado en los datos de entrenamiento (Tusar et al., 2017).

k-Nearest Neighbors: comúnmente abreviado como KNN o k-NN, es un método de aprendizaje supervisado que se clasifica dentro de los clasificadores no paramétricos. Este enfoque depende de la cercanía entre los puntos de datos para realizar clasificaciones o estimaciones sobre la categoría a la que pertenece un punto específico.

Aunque KNN se puede aplicar tanto a tareas de clasificación como de regresión, su uso más frecuente es en clasificación. La idea central detrás de KNN es que los puntos de datos con características similares tienden a agruparse. Por lo tanto, para determinar la clasificación de un nuevo punto de datos, el algoritmo busca los k puntos más cercanos (vecinos) y basa su predicción en la categoría predominante dentro de este grupo cercano (Géron, 2020).

Árbol de Decisión: Utiliza una estructura en forma de árbol para representar una serie de decisiones y sus posibles consecuencias. Es ampliamente reconocido por su claridad

interpretativa y adaptabilidad a distintos tipos de datos, siendo particularmente útil para tareas de clasificación (Theodoridis, 2015).

Regresión Logística: Este modelo estadístico es capaz de predecir el resultado de una variable categórica basándose en otras variables independientes. Es especialmente efectivo en clasificaciones binarias, como puede ser la identificación de mensajes de spam o la detección de enfermedades (Dreiseitl y Ohno-Machado, 2002).

Máquinas de Vectores de Soporte (SVM): Esta metodología es efectiva para clasificar datos en múltiples categorías creando un hiperplano o 'vector' que maximiza la distancia entre las clases. En situaciones donde las clases no son linealmente separables, se utilizan funciones de kernel para facilitar la clasificación (Dreiseitl y Ohno-Machado, 2002).

La Matriz de Confusión: se presenta como una herramienta evaluativa clave, proporcionando una visualización comprensible del rendimiento de los modelos al comparar las predicciones generadas con los valores reales, facilitando la identificación precisa de verdaderos y falsos positivos, así como verdaderos y falsos negativos (Robalino, et. al, 2020).

Tabla 6.

Matriz de confusión

		PREDICCIÓN	
		Positivos	Negativos
OBSERVACION	Positivos	Verdaderos Positivos (TP)	Falsos Negativos (FN)
	Negativos	Falsos Positivos (FP)	Verdaderos Negativos (TN)

Nota. La Matriz de Confusión compara las predicciones de un modelo con los valores reales, organizándose en Verdaderos Positivos (TP), Falsos Negativos (FN), Falsos Positivos (FP), y Verdaderos Negativos (TN). Esta estructura permite evaluar la precisión del modelo en la clasificación, destacando su eficacia y las áreas de mejora.

El empleo de *pipelines* en este contexto busca simplificar y estandarizar el proceso de modelado mediante la automatización de las secuencias de pasos necesarios para la preparación de datos y la aplicación de los modelos. Esta metodología es indispensable para asegurar la coherencia y eficiencia a lo largo de todo el proceso de análisis.

Durante la fase de evaluación, se implementan métricas detalladas para juzgar la precisión de los modelos en la tarea de clasificar MQL. Estas métricas incluyen:

- Precisión: Evalúa cuán precisas son las clasificaciones positivas del modelo. Si un modelo identifica 100 leads como MQL y 80 de ellos realmente lo son, la precisión es del 80%.

$$\text{PRECISION} = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ POSITIVE}$$

- Recall: Mide la habilidad del modelo para detectar todos los casos reales de MQL. Por ejemplo, si de 150 MQL reales, el modelo identifica 120, entonces el recall es del 80%.

$$\text{RECALL} = \frac{TRUE\ POSITIVE}{TRUE\ POSITIVE + FALSE\ NEGATIVE}$$

- F1 Score: Combina precisión y recall en un único indicador para ofrecer un balance entre ambas medidas. Un modelo con una precisión y un recall del 80% tendría un F1 Score del 80%.

$$F1 = \frac{2x(PRECISION\ X\ RECALL)}{(PRECISION + RECALL)}$$

- Exactitud: Calcula el porcentaje total de predicciones correctas, aunque debe usarse con precaución en datos desbalanceados.

$$ACCURACY = \frac{TP+TN}{TP+TN+FP+FN}$$

Para agilizar el proceso de evaluación se desarrolló una función personalizada fit_evaluate_pipeline que permite automatizar el flujo de trabajo aplicando los pipelines preparados a cada modelo y generando un análisis comparativo de su rendimiento en base a las métricas mencionadas. Así, se garantiza una evaluación uniforme y eficiente, crucial para identificar el enfoque de vectorización y el modelo de clasificación más efectivos para distinguir entre MQL y otros leads.

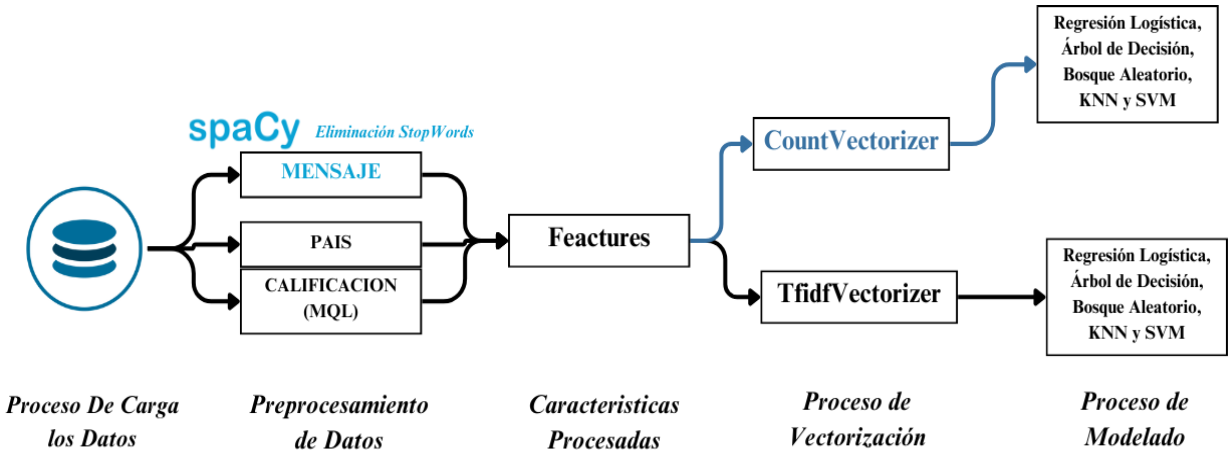
Antes de sumergirnos en el detallado proceso de modelado, es importante reconocer la cuidadosa preparación de los datos, utilizando las capacidades de Sklearn para segmentar el conjunto de datos en una proporción de 75% para entrenamiento y 25% para pruebas. Esta segmentación estratégica permitió validar la precisión del modelo en datos no observados, un paso crucial en el desarrollo de un modelo robusto.

Durante el preprocesamiento, se empleó la biblioteca SpaCy para limpiar los mensajes de los clientes, eliminando palabras no significativas y optimizando así el contenido para el análisis. Este refinamiento de los datos fue fundamental para la precisión y eficacia de los modelos posteriores.

Los modelos de machine learning utilizados, incluyendo la Regresión Logística, Árbol de Decisión, Bosque Aleatorio, KNN y SVM, fueron seleccionados por su relevancia y probada eficacia en tareas de clasificación. Se experimentó con técnicas de vectorización como CountVectorizer y TF-IDFVectorizer para transformar los mensajes textuales en vectores numéricos, permitiendo así que los algoritmos procesaran y aprendieran de los datos de manera efectiva.

Figura 9.

Proceso de Modelado



Fuente: Elaboración Propia

Nota: Si desea examinar o consultar el código utilizado en este proyecto, puede visitar el repositorio en GitHub: García Vidal, J. (2024). ProspectAI. Recuperado de https://github.com/JoseFernandoGV/ProspectAI.

Análisis Comparativo de Modelos

Para evaluar los modelos, se utilizó una variedad de métricas, incluyendo la precisión, el recall y el puntaje F1, que se calcularon tanto en los conjuntos de entrenamiento como de prueba. Este enfoque proporcionó una visión integral del rendimiento de cada modelo, destacando su capacidad para identificar verdaderos leads calificados frente a los no calificados. Tras concluir

el proceso de modelado, se evaluaron y compararon dos estrategias de vectorización: CountVectorizer y TF-IDFVectorizer, aplicadas sobre distintos modelos en el ámbito del aprendizaje supervisado. Entre los modelos examinados se encontraban la Regresión Logística, el Árbol de Decisión, el Bosque Aleatorio, k-Nearest Neighbors (KNN) y las Máquinas de Soporte Vectorial (SVM). Se puso especial énfasis en la necesidad de minimizar los falsos negativos, considerando que clasificar erróneamente a un lead como no calificado podría ocasionar la pérdida de una valiosa oportunidad de conversión. A continuación, se presentan los resultados obtenidos con ambas técnicas de vectorización y sus respectivas interpretaciones:

Tabla 7.

Matriz de resultados del proceso de modelado

COUNTVECTORIZER					**TFIDFVECTORIZER**				
Modelo	Calificación	precision	recall	f1-score	Modelo	Calificación	precision	recall	f1-score
Logistic Regression *(Accuracy: 0.9119)*	0	0.88	0.94	0.91	Logistic Regression *(Accuracy: 0.9136)*	0	0.88	0.94	0.91
	1	0.95	0.89	0.92		1	0.94	0.89	0.92
Decision Tree *(Accuracy: 0.8966)*	0	0.86	0.93	0.89	Decision Tree *(Accuracy: 0.9073)*	0	0.89	0.92	0.90
	1	0.93	0.87	0.90		1	0.93	0.90	0.91
Random Forest *(Accuracy: 0.9186)*	0	0.89	0.93	0.91	Random Forest *(Accuracy: 0.9186)*	0	0.90	0.93	0.91
	1	0.94	0.90	0.92		1	0.94	0.91	0.92
KNN *(Accuracy: 0.7458)*	0	0.65	0.99	0.78	KNN *(Accuracy: 0.5831)*	0	0.53	0.97	0.68
	1	0.98	0.54	0.69		1	0.90	0.25	0.39
SVM *(Accuracy: 0.9260)*	0	0.90	0.89	0.90	SVM *(Accuracy: 0.9130)*	0	0.92	0.89	0.90
	1	0.95	0.91	0.93		1	0.91	0.93	0.92

Nota. La tabla muestra el rendimiento de modelos de machine learning utilizando CountVectorizer y TfidfVectorizer. Se evalúan precisión, recall y puntuación F1, indicando la exactitud de las predicciones.

CountVectorizer:

- *Regresión Logística:* Muestra un alto balance entre precisión y recall, con una precisión del 95% para la clase 1, lo que indica una alta capacidad para identificar correctamente los MQL. Sin embargo, el recall de 89% sugiere que hay un pequeño margen de mejora en capturar todos los MQL potenciales.
- *Árbol de Decisión:* Ofrece un rendimiento ligeramente inferior al de la regresión logística, con una precisión del 93% y un recall del 87% para la clase 1. Esto indica una buena capacidad de clasificación, pero con un riesgo moderado de omitir algunos MQL.
- *Random Forest:* Exhibe un rendimiento comparable al de la regresión logística, con una precisión del 94% y un recall del 90% para la clase 1, lo que lo posiciona como un modelo robusto para identificar MQL con un bajo riesgo de falsos negativos.
- *KNN:* Presenta un rendimiento significativamente inferior, especialmente en el recall para la clase 1 (54%), lo que indica una alta proporción de MQL no detectados, a pesar de una precisión del 98%
- *SVM:* Es el modelo con mejor equilibrio, alcanzando la mayor precisión (95%) y un alto recall (91%) para la clase 1, sugiriendo ser el más eficaz en clasificar MQL minimizando tanto falsos positivos como negativos.

TF-IDFVectorizer:

- *Regresión Logística, Árbol de Decisión y Bosque Aleatorio:* Aunque cada uno mostró mejoras en sus métricas con TF-IDF en comparación con CountVectorizer, aún no superan el rendimiento global de SVM con

CountVectorizer, especialmente en el aspecto crucial del recall para leads calificados.

- *KNN:* Con TF-IDF, este modelo mostró una disminución significativa en su eficacia, subrayando su inadecuación para el conjunto de datos y el objetivo de minimizar falsos negativos.
- *SVM:* Aunque su desempeño con TF-IDF fue levemente inferior a su versión con CountVectorizer, sigue siendo excepcionalmente alto, reforzando su posición como el modelo más apto para la clasificación precisa de leads MQL.

Selección del Mejor Modelo:

Al considerar la importancia de minimizar los falsos negativos y la eficacia en la clasificación precisa de leads como MQL, el modelo SVM utilizando CountVectorizer se destaca como el más adecuado. Su rendimiento superior en precisión, recall, y especialmente en el equilibrio entre minimizar falsos negativos y maximizar la identificación correcta de leads calificados, lo convierte en la elección óptima para la clasificación de leads en este contexto. Este análisis detallado asegura una base sólida para futuras acciones y estrategias de marketing, orientadas hacia la maximización de oportunidades de conversión.

El análisis previo indicó que el método de tokenización CountVectorizer se ajusta adecuadamente a los datos, y se seleccionó el modelo SVM por su desempeño destacado. Es importante señalar que en la próxima sección se llevarán a cabo pruebas adicionales para evaluar más a fondo el rendimiento del modelo. Aunque los resultados iniciales son prometedores, es crucial realizar estas pruebas adicionales para verificar la presencia de sobreajuste y asegurar que el modelo mantiene su eficacia al generalizar a nuevos datos. Esta etapa es vital para confirmar

la idoneidad del modelo SVM en la clasificación precisa de los clientes potenciales y garantizar la robustez de la solución propuesta.

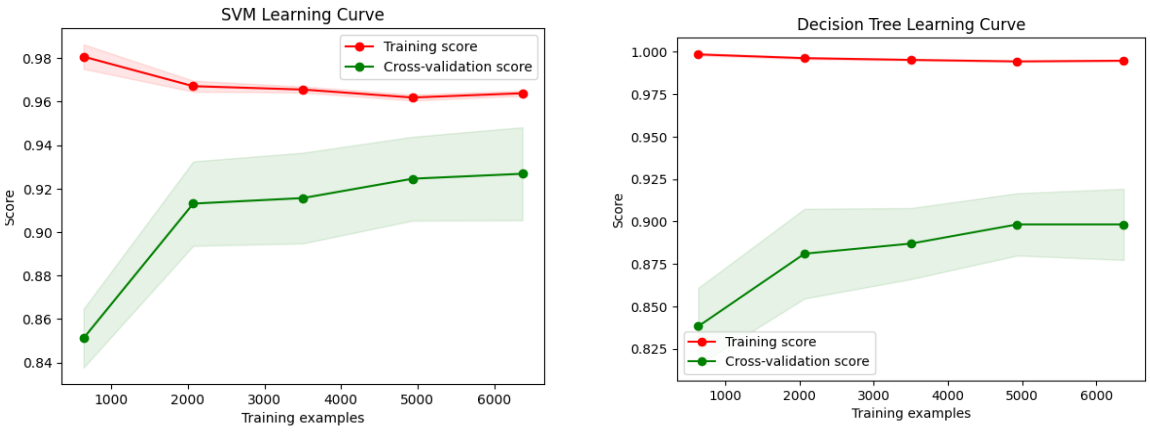
Evaluar el rendimiento del modelo, utilizando métricas de machine learning identificando áreas de mejora en su capacidad de clasificación de leads como clientes potenciales (MQL)

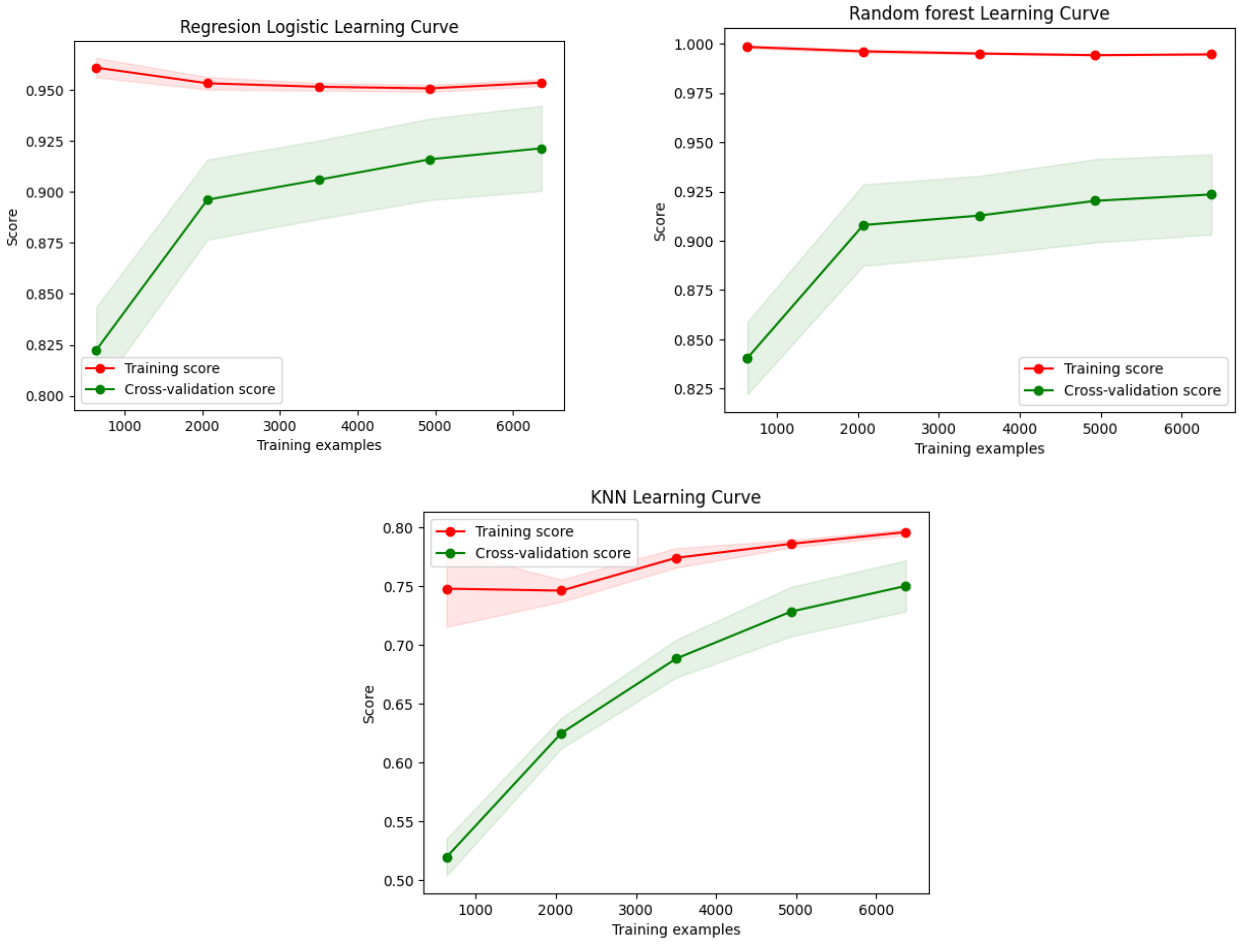
Después de definir el método de tokenización CountVectorizer, se avanzó con la implementación de las curvas de aprendizaje para analizar la regularización del modelo. Una curva de aprendizaje representa gráficamente cómo evoluciona la eficiencia de un modelo de aprendizaje automático con la acumulación de experiencia o a lo largo del tiempo. Estas curvas son esenciales para identificar y diagnosticar los distintos comportamientos del modelo durante su entrenamiento, permitiendo determinar si se requieren ajustes o si el modelo se ajusta correctamente al tamaño y a la complejidad de los datos de entrenamiento y validación (Perlich, 2010). El análisis de estas curvas ayuda a reconocer si hay problemas tales como sobreajuste o subajuste y a evaluar si los datos utilizados son representativos y adecuados para el entrenamiento del modelo. La figura 10 presenta las curvas de aprendizaje correspondientes a

todos los modelos examinados, ofreciendo una perspectiva visual de su rendimiento y evolución.:

Figura 10.

Curvas de Aprendizaje de los distintos modelos





Fuente: Elaboración Propia

Durante la evaluación de los modelos mediante curvas de aprendizaje, se identificó que el modelo SVM, a pesar de su destacado rendimiento en términos de precisión y sensibilidad, exhibía señales de sobreajuste. Esto implica que el modelo podría estar excesivamente ajustado a los datos de entrenamiento y, por tanto, con limitada capacidad para predecir con precisión nuevos datos. Aunque el modelo KNN no obtuvo métricas tan elevadas, mostró una regularización más balanceada, indicando una mejor adaptabilidad a datos no vistos anteriormente.

Tuning Parameters

Frente a la tendencia al sobreajuste observada en la mayoría de los modelos, se optó por una técnica avanzada denominada 'ajuste de hiperparámetros' o 'tuning'. Este proceso de optimización busca los valores óptimos para los hiperparámetros del modelo, que son configuraciones clave que guían su aprendizaje (Hossain & Timmer, 2021). Para ello, se utilizó la validación cruzada, un método de evaluación que incrementa la fiabilidad de los resultados del modelo. La validación cruzada implica dividir los datos en subconjuntos múltiples y realizar series iterativas de entrenamiento y validación, en las que cada subconjunto se utiliza tanto para entrenar como para validar el modelo en diferentes ciclos. Esto asegura que el modelo sea probado contra diversas porciones de los datos, mitigando así el riesgo de un ajuste excesivo a un único conjunto de entrenamiento y promoviendo una evaluación más robusta y representativa del rendimiento del modelo

El fin último de incorporar el ajuste de hiperparámetros junto con la validación cruzada es refinar los modelos, para que se desempeñen de manera óptima y sean capaces de generalizar mejor sus predicciones más allá de los datos sobre los cuales fueron entrenados, a continuación, se encuentran los parámetros óptimos que se obtuvieron del tuning parameter con sus respectivas curvas de aprendizaje.

Tabla 8.

Resultados del Tunnig Parameters

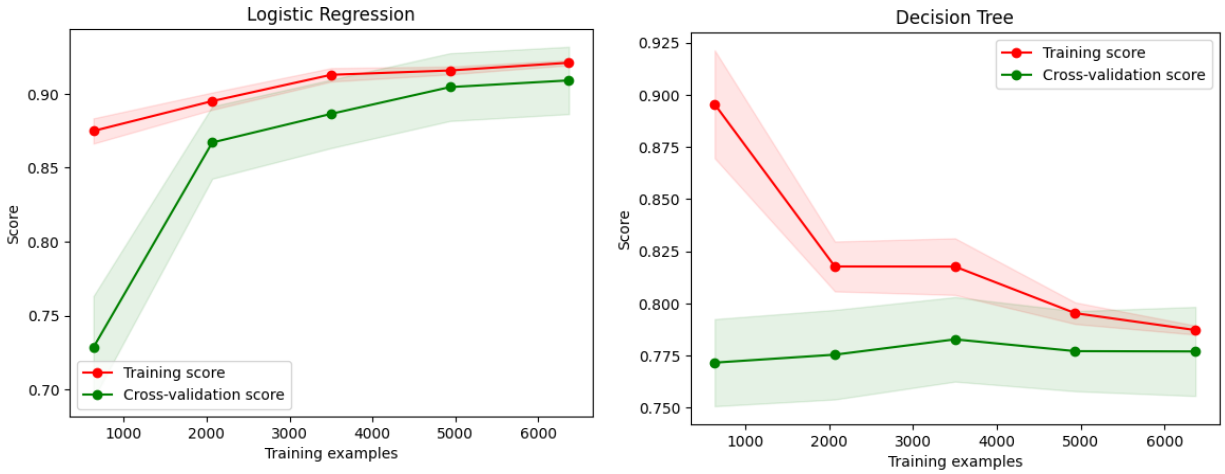
MODELO	PRECISIÓN DE PRUEBA	PARÁMETROS ÓPTIMOS
Logistic Regression	0,962	C: 0.1, Penalty (Pen): l2, Solver: liblinear

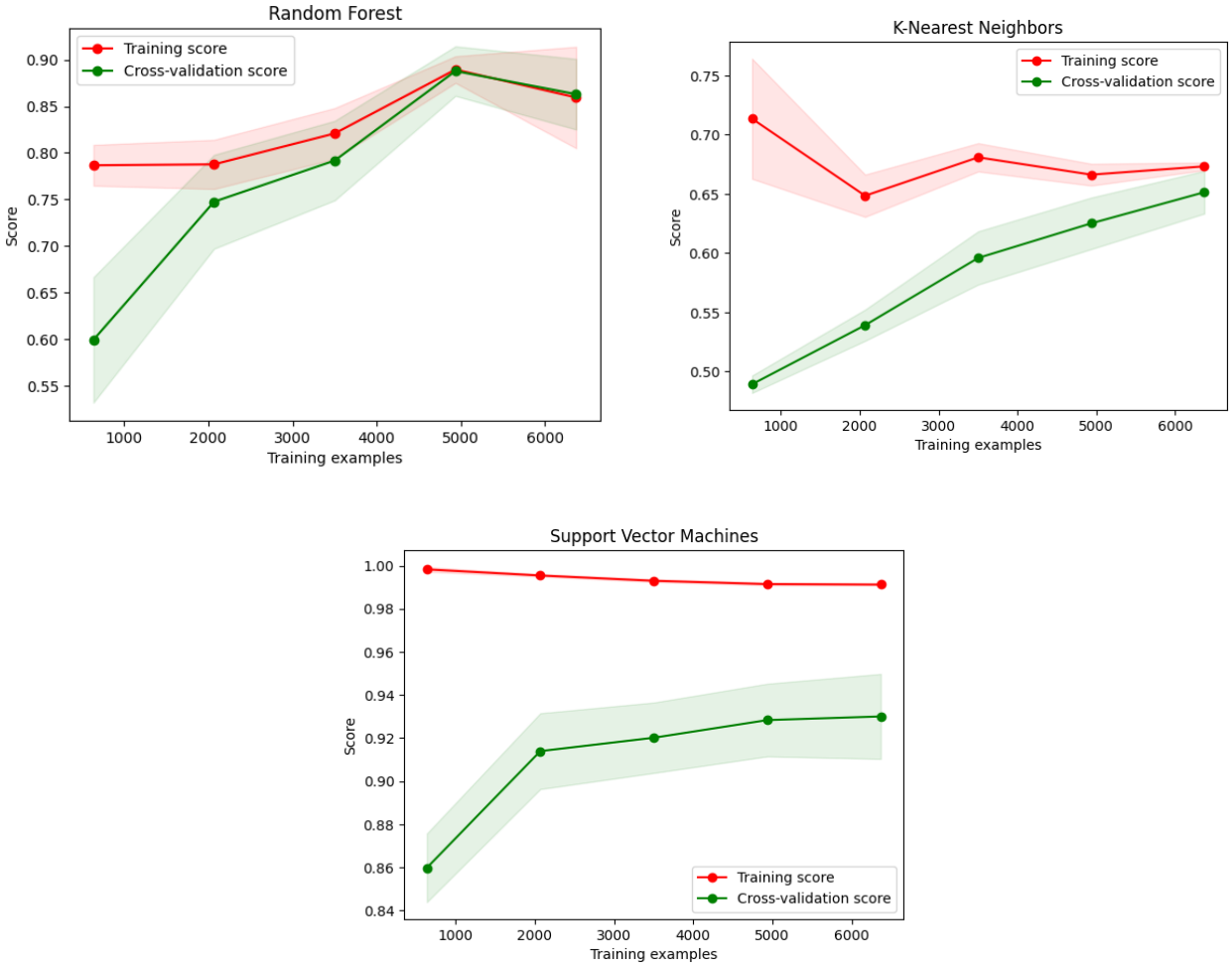
Decision Tree	0,981	Criterio: gini, Profundidad máxima (Prof. máx.): 6, Mín. muestras hoja: 1, Mín. muestras división: 4
Random Forest	0,889	Profundidad máxima (Prof. máx.): 6, Mín. muestras hoja: 3, Mín. muestras división: 2
K-Nearest Neighbors	0,991	Métrica: manhattan, Número de vecinos (Vecinos): 5, Pesos: uniforme
Support Vector Machines	0,948	C: 5, Kernel: rbf

Nota. Esta tabla presenta los resultados del ajuste de hiperparámetros para diversos modelos de machine learning. Las precisiones de prueba indican la efectividad de cada modelo bajo la configuración óptima de parámetros, resaltando la importancia de una configuración meticulosa para mejorar la precisión en la clasificación de leads.

Figura 11.

Curvas de aprendizaje con Hiperparametros Optimizados





Fuente. Elaboración Propia

Después de aplicar el ajuste de hiperparámetros, se observó que, aunque el modelo SVM había mostrado un rendimiento destacado en las métricas con el método CountVectorizer, aún presentaba signos de sobreajuste en comparación con otros modelos. Para evaluar la capacidad de generalización y regularización, se generaron curvas de aprendizaje para cada modelo.

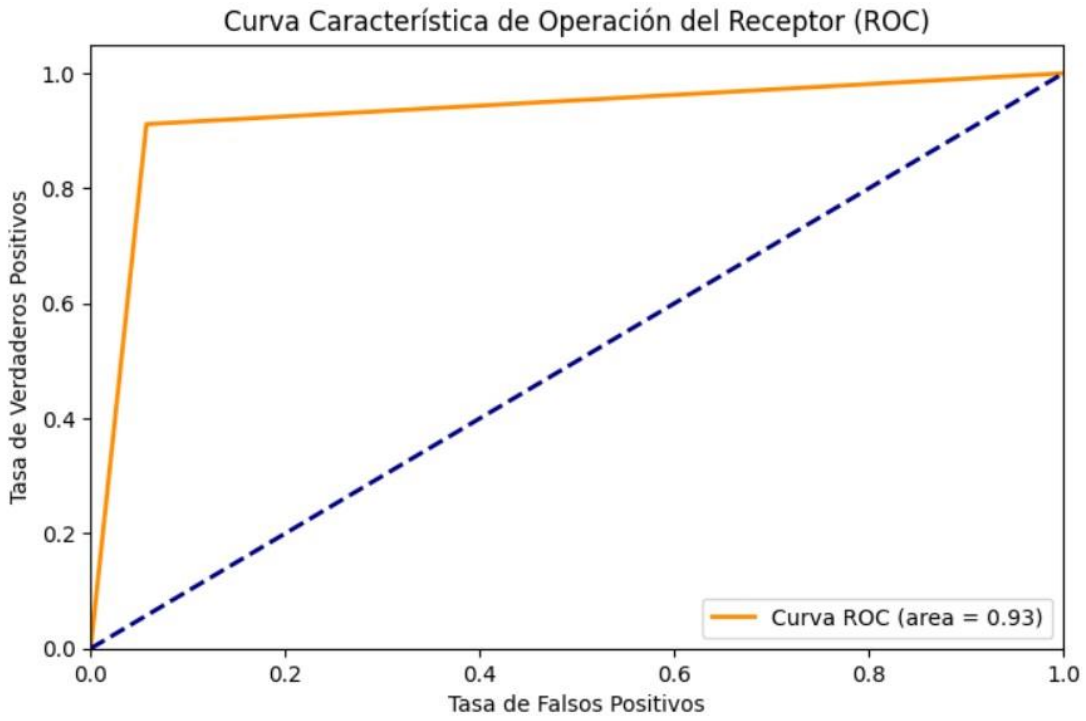
En la interpretación de estas curvas, se hizo evidente que, aunque el SVM seguía siendo fuerte en términos de precisión, la Regresión Logística mostró una regularización más robusta después del ajuste de hiperparámetros. En términos de las métricas de clasificación (precisión, recall y f1-score), la Regresión Logística demostró estar a la par con el SVM, pero con una mejor capacidad de generalizar a datos no vistos, lo que indica una reducción significativa en el riesgo

de sobreajuste. Por lo tanto, el análisis concluye que la Regresión Logística, tras la optimización de hiperparámetros, emerge como el modelo más equilibrado. Presenta no solo una alta precisión, crucial para asegurar que los leads identificados como MQL sean verdaderos, sino también una regularización adecuada, lo que la convierte en la elección preferente para la clasificación de MQLs en este escenario particular

Curva Característica de Operación del Receptor (ROC)

Figura 12.

Curva ROC del modelo Regresión logística



Fuente. Elaboración Propia

Tras aplicar el ajuste de hiperparámetros, la Regresión Logística no solo mostró una regularización efectiva, sino también resultados prometedores en la Curva Característica de Operación del Receptor (ROC). Esta curva es un gráfico de la tasa de verdaderos positivos frente a la tasa de falsos positivos (Prati, et. al, 2008). En el caso de la Regresión Logística, la curva

ROC alcanza un área bajo la curva (AUC) de 0.91, lo que es significativamente superior a un clasificador aleatorio, indicado por la línea punteada diagonal con un AUC de 0.5.

El alto valor del AUC refleja la excelente capacidad del modelo de Regresión Logística para diferenciar entre las clases de MQL y no MQL. Un modelo perfecto tendría un AUC de 1, mientras que el alto AUC obtenido sugiere que la Regresión Logística posee un excelente equilibrio entre sensibilidad (identificar verdaderos MQL) y especificidad (rechazar no MQL), con una tasa reducida de falsos positivos.

Este análisis robusto y la capacidad demostrada de generalización justifican la elección de la Regresión Logística como el modelo óptimo en este proyecto. No solo proporciona una alta precisión en la clasificación de MQLs, sino que también garantiza una confianza considerable en su capacidad predictiva, como lo demuestra la curva ROC.

Conclusiones

La investigación desarrollada en este proyecto fue impulsada por la necesidad de optimizar el proceso de identificación de clientes potenciales cualificados para la empresa MAVV. A través de un enfoque sistemático y metódico, se abordaron distintas fases que contribuyeron a la consecución de los objetivos específicos planteados inicialmente.

Se comenzó con un análisis exploratorio profundo de la base de datos "Customer Data", donde se extrajeron patrones y tendencias significativas que esclarecieron aspectos cruciales para una clasificación efectiva. Este análisis no solo sirvió para entender mejor los datos con los que se trabajaba, sino que también fundamentó las decisiones metodológicas posteriores.

Se procedió a implementar técnicas avanzadas de procesamiento de lenguaje natural (NLP) para tratar los mensajes enviados por los clientes. Esta etapa fue esencial para transformar datos textuales en un formato que los modelos de machine learning pudieran interpretar eficazmente, extrayendo características que son críticas para la precisión del modelo.

La selección y entrenamiento de modelos de machine learning se centraron en identificar la herramienta más eficaz para clasificar los leads. Después de evaluar varias alternativas, el modelo de Regresión Logística emergió como el más prometedor, equilibrando adecuadamente la sensibilidad y especificidad necesarias para el contexto de MAVV. Este modelo no solo cumplió con los criterios de rendimiento establecidos, sino que también mostró un equilibrio robusto en las métricas de evaluación clave.

Finalmente, el rendimiento del modelo seleccionado fue rigurosamente evaluado a través de métricas estándar, incluyendo la precisión, el recall y la curva ROC. Estas evaluaciones

confirmaron la capacidad del modelo para clasificar efectivamente los leads como clientes potenciales cualificados (MQL). A pesar de estos resultados positivos, se tomó la decisión de no avanzar hacia la implementación en producción, considerando las limitaciones prácticas y los desafíos de integración que podrían comprometer la efectividad operacional del modelo.

Este proyecto ha proporcionado no solo una visión detallada de la clasificación de leads en MAVV, sino también un marco valioso para futuras investigaciones y aplicaciones prácticas en el ámbito de la ciencia de datos. A través de este trabajo, se ha destacado la importancia de abordar los proyectos de modelado con un pensamiento crítico, preparados para adaptar y evolucionar frente a los desafíos prácticos y teóricos que surgen.

Referencias

Alias, G., Cassanelli, R. (2019). *NLP aplicado a análisis de texto* (Doctoral dissertation, Universidad Nacional de Mar del Plata. Facultad de Ingeniería. Argentina).

Alvi, N., Talukder, K. (2021). *Sentiment Analysis of Bengali Text using CountVectorizer with Logistic Regression. 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), Computing Communication and Networking Technologies (ICCCNT), 2021 12th International Conference On*, 01–05. https://doi-org.bibliotecavirtual.unad.edu.co/10.1109/ICCCNT51525.2021.9580017.

Bobadilla, J. (2021). *Machine learning y deep learning: usando Python, Scikit y Keras. Ediciones de la U*.

Barredo, I. C. (2018). *Marketing digital: Mide, analiza y mejora*. ESIC Editorial.

Bravo, J. (2021). *La influencia de la inteligencia artificial en el futuro del marketing*. Big Bang Faustiniano, 10(3).

Carrera, E., García, M., Pasmay, F. (2008). *Un algoritmo simple y eficiente para la clasificación automática de páginas Web*.

Chu Rubio, M. (2020). *El ROI de las decisiones del marketing. Universidad Peruana de Ciencias Aplicadas*.

Doyle, C. (2016). *A dictionary of marketing. Oxford University Press*.

Dreiseitl, S., Machado, L. (2002). *Logistic regression and artificial neural network classification models: a methodology review. Journal of Biomedical Informatics*, 35(5), 352-359. doi: 10.1016/S1532-0464(03)00034-0.

Hernández, Rosa., Hernández, V., Batista, N., Castañeda, E. (2017). *¿Chi cuadrado o Ji cuadrado? Medicentro Electrónica*, 21(4), 294-295.

Hernández, S. (2020). *La fidelización del cliente y retención del cliente: Tendencia que se exige hoy en día. Gestión en el tercer milenio,* 23(45), 5-13.

Hossain, M., Timmer, D. (2021). *Machine learning model optimization with hyper parameter tuning approach. Global Journal of Computer Science and Technology*, 21(D2), 7-13.

Jarek, K., Mazurek, G. (2019). *Marketing and Artificial Intelligence. Central European Business Review*, 8(2).

Järvinen, J., Taiminen, H. (2016). *Harnessing marketing automation for B2B content marketing. Industrial Marketing Management*, 54, 164-175. https://doi.org/10.1016/j.indmarman.2015.07.002.

Kavlakoglu, E. (2020). *NLP vs. NLU vs. NLG: the differences between three natural language processing concepts*. https://www.ibm.com/blog/nlp-vs-nlu-vs-nlg-the-differences-between-three-natural-language-processing-concepts/.

Llerena, J. (2020). *Codifica en Python*.

Mudasir, Y. (2024). *How to remove stop words using spaCy in Python*. https://www.educative.io/answers/how-to-remove-stop-words-using-spacy-in-python.

Perlich, C. (2010). *Learning Curves in Machine Learning*.

Prati, C., Batista, G., Monard, M. (2008). *Evaluating classifiers using ROC curves*. IEEE Latin America Transactions, 6(2), 215-222.

Poecze, F., Ebster, C., Strauss, C. (2018). *Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts*. Procedia Computer Science, 130, 660-666. https://doi:10.1016/j.procs.2018.04.117.

Rouhiainen, L. (2018). *Inteligencia artificial*. Madrid: Alienta Editorial.

Robalino, R., Getino, A., Rodellar, J. (2020). *Estandarización de métricas de rendimiento para clasificadores Machine y Deep Learning*. Revista Ibérica de Sistemas e Tecnologias de Informação, (30), 184-196.

Shandeep, S. (2022). *CountVectorizer vs TfidfVectorizer* https://medium.com/@shandeep92/countvectorizer-vs-tfidfvectorizer-cf62d0a54fa4.

Theodoridis, S. (2015). *Machine Learning: A Bayesian and Optimization Perspective*. Estados Unidos, Massachusetts: Academic Press.

Sabnis, G., Chatterjee, S., Grewal, R., Lilien, G. (2013). *The Sales Lead Black Hole: On Sales Reps' Follow-Up of Marketing Leads*. Journal of Marketing, 77(1), 52–67. https://doi-org.bibliotecavirtual.unad.edu.co/10.1509/jm.10.0047.

Siavichay, C., Eduardo, A. (2023*). La transformación digital comercial de empresas B2B: incidencia en la satisfacción de los clientes; La Transformación Digital Comercial de Empresas B2B: Incidencia en la Satisfacción de los Clientes*. Repositorio de la Universidad de Cuenca.

Sánchez, J. (2011). *La innovación: una revisión teórica desde la perspectiva de marketing*. Perspectivas, (27), 47-71.

Shevtsova, Z., Roman, V., Sokolova, B. (2021). *Digital innovation as a driving force for development of modern companies. European Reforms Bulletin*, 95.

Sciammarelli, J. (2023*). El caso de uso de la inteligencia artificial en marketing y ventas. Revista Foco (Revista de Estudios Interdisciplinarios)*, 16 (5), 1–24. https://doi-org.bibliotecavirtual.unad.edu.co/10.54751/revistafoco.v16n5-142

Tavira, E., Estrada, E. (2015). *Marketing relacional: valor, satisfacción, lealtad y retención del cliente. Análisis y reflexión teórica*. Ciencia y sociedad, 40(2), 307-340.

Turletti, P. (2018). *El ROI de marketing y ventas: Cálculo y utilidad*. Nuevo estarndar de rendimiento. ESIC Editorial.