Universidad de Buenos Aires
Facultad de Ciencias Económicas
Escuela de Negocios y Administración Pública

CARRERA DE ESPECIALIZACIÓN EN MÉTODOS CUANTITATIVOS PARA LA GESTIÓN Y ANÁLISIS DE DATOS EN ORGANIZACIONES

TRABAJO FINAL INTEGRADOR DE ESPECIALIZACIÓN

Optimización Estratégica de Marketing en Empresas de eLearning. *Procesamiento de Datos en Python y Modelos Predictivos para la Conversión de Clientes*

AUTORA: SOFIA ANA WAISSMANN

TUTOR: PABLO MATIAS HERRERA

JUNIO 2024

Resumen

El presente trabajo de investigación aborda la problemática crítica que enfrentan las empresas de e-learning en la actualidad: la optimización de estrategias de marketing para convertir prospectos en clientes. En un entorno educativo digital diverso y en constante evolución, la toma de decisiones estratégicas en marketing se ha vuelto cada vez más compleja debido a la rápida transformación de las preferencias estudiantiles, las tendencias tecnológicas y las dinámicas del mercado.

El objetivo principal de esta investigación es analizar cómo la implementación del procesamiento de datos en Python y la aplicación de modelos de análisis predictivo pueden optimizar las estrategias de marketing en el contexto de empresas de e-learning, aumentando la conversión de leads y mejorando la eficacia de los esfuerzos de conversión. Para ello, se estructura el estudio en varias secciones. En primer lugar, se aborda la consideración de datos en contextos organizacionales. En segundo lugar, se detalla la selección y análisis de datos. En tercer lugar, se exploran y comparan modelos predictivos. Finalmente, se presentan conclusiones y recomendaciones basadas en los hallazgos obtenidos.

Para llevar a cabo la investigación, se utilizaron diversas técnicas y herramientas. Se empleó Python para el procesamiento y análisis de datos, incluyendo la limpieza de datos, eliminación de valores faltantes y outliers, imputación de datos y análisis exploratorio de datos. Se comprendió la distribución y relaciones entre las variables y se realizaron visualizaciones de datos para identificar patrones y tendencias. La selección de variables clave se realizó mediante análisis de correlación y el uso de la métrica de Cramer's V para variables categóricas. Se evaluaron varios modelos de aprendizaje automático, incluyendo Regresión Logística, Árbol de Decisión, Máquina de Vectores de Soporte (SVM), K-Nearest Neighbors (KNN), Gradient Boosting y Random Forest. De todos ellos, Gradient Boosting se destacó por su rendimiento superior, alcanzando una exactitud del 88%, un F1-Score del 85% y un AUC-ROC del 95%.

La visualización de resultados incluyó matrices de confusión para identificar verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos, así como curvas ROC para evaluar el rendimiento de los modelos en la clasificación de casos positivos y negativos. Se realizó un análisis de errores para identificar patrones en los datos donde los modelos fallan en

la clasificación correcta. Además, se llevó a cabo un análisis de la importancia de las características para entender el impacto de cada variable en las predicciones.

Palabras clave: Empresas de e-learning, Python, Modelos predictivos, Marketing, Clientes

Introducción

La educación en línea, también conocida como e-learning, ha experimentado una evolución significativa en respuesta a la creciente demanda de formación digital. Las empresas del sector enfrentan desafíos en la conversión de prospectos en clientes debido a la diversidad de la audiencia y la rápida evolución del entorno educativo digital. Este trabajo tiene como objetivo optimizar las estrategias de marketing para empresas de e-learning mediante el procesamiento de datos en Python y la aplicación de modelos predictivos.

El objetivo general de esta investigación es analizar cómo el procesamiento de datos en Python y los modelos predictivos pueden mejorar las estrategias de marketing en empresas de e-learning, aumentando la conversión de leads y mejorando la eficacia de los esfuerzos de conversión. Para cumplir este objetivo, se propone la siguiente estructura: en primer lugar, abordar la consideración de datos en contextos organizacionales y empresas de e-learning, analizando el ecosistema de datos y su utilización estratégica. En segundo lugar, centrarse en variables y procesamiento en estrategias predictivas de e-learning, detallando el proceso de recolección y procesamiento de datos. En tercer lugar, explorar y analizar el rendimiento de los modelos predictivos en empresas de e-learning. Finalmente, presentar conclusiones y recomendaciones basadas en los hallazgos obtenidos.

Para llevar a cabo la investigación, se utilizarán diversas técnicas y herramientas. Se empleará Python para el procesamiento y análisis de datos, incluyendo la limpieza de datos, eliminación de valores faltantes y outliers, imputación de datos y análisis exploratorio de datos (EDA). Se comprenderá la distribución y relaciones entre las variables y se realizarán visualizaciones de datos para identificar patrones y tendencias. La selección de variables clave se realizará mediante análisis de correlación y el uso de la métrica de Cramer's V para variables categóricas. Se evaluarán varios modelos de aprendizaje automático, incluyendo Regresión Logística, Árbol de Decisión, Máquina de Vectores de Soporte (SVM), K-Nearest Neighbors (KNN), Gradient Boosting y Random Forest. De todos ellos, Gradient Boosting se destacará por su rendimiento superior, alcanzando una exactitud del 88%, un F1-Score del 85% y un AUC-ROC del 95%.

1.Consideración de Datos en Contextos Organizacionales y Empresas de eLearning

En el ámbito del marketing para empresas de e-learning, considerar datos en contextos organizacionales es esencial para tomar decisiones estratégicas. La segmentación de audiencia se destaca como un componente clave. Recopilar y analizar datos permite comprender en profundidad las características demográficas y las preferencias de los usuarios. Este enfoque facilita personalizar mensajes y ofertas, adaptándolos a las necesidades específicas de cada segmento.

Personalizar contenidos es otro aspecto crucial. Comprender el comportamiento de los estudiantes en plataformas de e-learning desempeña un papel central. Analizar cómo los usuarios interactúan con el material educativo proporciona información valiosa para ajustar estrategias de marketing. Este análisis también mejora la experiencia de aprendizaje.

Adaptarse a las tendencias del mercado es un tercer elemento clave. Analizar datos sobre nuevas tecnologías educativas, preferencias de los estudiantes y cambios en la competencia es fundamental. Esta habilidad permite a las empresas anticipar y ajustarse a las tendencias emergentes. De esta manera, las empresas pueden mantenerse relevantes en un entorno dinámico.

Evaluar el rendimiento de las campañas de marketing es esencial para una estrategia efectiva. Medir conversiones, tasas de clics y otros indicadores clave proporciona información valiosa. Estos datos sobre la efectividad de las estrategias actuales orientan ajustes para futuras iniciativas. Esta práctica permite una mejora continua.

Considerar datos en el contexto del marketing para empresas de e-learning no solo optimiza la eficacia de las campañas. También proporciona una base sólida para la adaptabilidad, la personalización y la anticipación de tendencias. Estos elementos son esenciales en un sector en constante evolución.

1.1 Análisis del Ecosistema de Datos en el Marco Organizacional de Empresas de eLearning

En el contexto dinámico y evolutivo de las empresas de e-learning, analizar el ecosistema de datos es esencial para comprender la complejidad y las oportunidades en el ámbito educativo

digital. Este análisis abarca la interacción y gestión de diversos elementos que facilitan la entrega efectiva de servicios educativos en entornos virtuales. "El uso de técnicas de minería de datos para predecir el rendimiento académico ha sido ampliamente investigado, destacando su eficacia en diversos contextos educativos" (Ahmed & Elaraby, 2014; Mueen et al., 2016).

Las empresas de e-learning dependen del aprovechamiento de datos para ofrecer experiencias de aprendizaje personalizadas y efectivas. En un mundo donde la información es un activo clave, analizar el ecosistema de datos influye en la toma de decisiones y en la optimización de los recursos. Construir sistemas capaces de encontrar patrones en los datos y aprender de ellos sin programación explícita es fundamental para predecir el abandono de clientes. "En el contexto de la predicción de abandono de clientes, estas son características de comportamiento en línea que indican una disminución de la satisfacción del cliente al usar los servicios y/o productos de la compañía" (Arango, 2021).

El ecosistema de datos en empresas de e-learning abarca una variedad de componentes interrelacionados. Las plataformas de gestión del aprendizaje (LMS) sirven como núcleo central para la entrega de contenidos, la interacción estudiante-instructor y la evaluación del progreso. Los sistemas de gestión de datos estudiantiles (SMS) almacenan y administran información relacionada con la inscripción, el rendimiento académico y otros datos estudiantiles.

La infraestructura tecnológica, incluyendo servidores, redes y sistemas de seguridad, constituye otra parte integral del ecosistema. Integrar herramientas analíticas y de inteligencia empresarial permite extraer conocimientos significativos a partir de los datos recopilados. Evaluar datos relacionados con la participación del estudiante, tasas de retención, preferencias de contenido y desempeño en evaluaciones proporciona información valiosa para adaptar y mejorar continuamente la oferta educativa.

Tomar decisiones informadas por datos también se extiende a la gestión de recursos y la planificación a largo plazo. Comprender la demanda de cursos, identificar tendencias emergentes en el ámbito educativo y evaluar el rendimiento de instructores son aspectos cruciales derivados del análisis del ecosistema de datos. Este enfoque basado en datos permite a las empresas de e-learning adaptarse ágilmente a las demandas cambiantes del mercado.

Analizar el ecosistema de datos en empresas de e-learning es una práctica necesaria para la eficiencia operativa y la calidad de la experiencia educativa ofrecida. Este enfoque proporciona una base sólida para la innovación continua en el campo educativo digital. Las empresas pueden mantenerse relevantes en un entorno dinámico mediante la adaptación y mejora continua.

1.2 Utilización de Datos para Adaptarse a Cambios en la Demanda Educativa

En el dinámico entorno educativo, adaptarse a cambios en la demanda es crucial para el éxito de las empresas de e-learning. Utilizar estratégicamente los datos es un recurso invaluable en este contexto. Esto permite a las organizaciones anticipar, responder y evolucionar en función de las fluctuaciones en las necesidades educativas. "A través del aprendizaje automatizado podemos perfeccionar estos métodos para que cada vez proporcionen resultados más precisos" (Arango, 2021).

El análisis predictivo basado en datos demográficos y de inscripción es una herramienta fundamental. Recopilar información demográfica de los estudiantes facilita la identificación de patrones de inscripción. Utilizar análisis predictivo para anticipar cambios estacionales en la demanda educativa es esencial. Evaluar continuamente el desempeño del contenido educativo a través de datos detallados contribuye a prever la demanda de cursos específicos. Esto permite una adaptación proactiva.

La flexibilidad en la oferta y las modalidades educativas es un aspecto esencial. Utilizar datos de preferencias de los estudiantes permite a las empresas de e-learning ofrecer modelos de aprendizaje híbridos que se ajusten a diferentes estilos de aprendizaje. Evaluar constantemente la eficiencia de estas modalidades, respaldada por analíticas de rendimiento y satisfacción estudiantil, guía la adaptación continua de la oferta educativa.

Personalizar trayectorias educativas se facilita a través de la recopilación y análisis de datos sobre cómo los estudiantes eligen y completan cursos. Implementar sistemas de recomendación personalizados basados en estos patrones de trayectorias educativas anteriores garantiza una experiencia educativa más personalizada y relevante.

En la gestión de recursos y la escalabilidad, optimizar recursos educativos mediante sistemas de gestión basados en datos permite tomar decisiones informadas sobre la creación y distribución de nuevos recursos. Evaluar la capacidad de las plataformas tecnológicas para manejar picos de demanda respalda inversiones estratégicas en tecnología. Esto garantiza la escalabilidad y el rendimiento durante períodos de alta demanda.

Adaptarse efectivamente a cambios en la demanda educativa se fundamenta en una comprensión profunda del panorama educativo. Utilizar inteligentemente los datos emerge como un activo estratégico. Esto permite a las empresas de e-learning no solo mantenerse al día con las tendencias cambiantes, sino también liderar la evolución del sector. Así, se pueden ofrecer soluciones educativas que anticipen y satisfagan las demandas emergentes.

1.3 Enfoque de Datos en Estrategias de Marketing para Empresas de eLearning

En el dinámico y competitivo panorama de las empresas de e-learning, formular estrategias de marketing respaldadas por un enfoque robusto en datos es una necesidad imperante. "Resulta desafiante para un comercializador manipular un modelo de predicción para sus propios objetivos, de modo que puedan decidir la mejor estrategia de actividad de marketing para cada cliente individual o subconjunto de clientes" (Mathur et al., 2022). El papel crítico de estas estrategias radica en la capacidad de las empresas para comprender a fondo a su audiencia y adaptar sus mensajes de manera altamente personalizada.

El análisis de datos se convierte en un pilar fundamental para crear estrategias de marketing efectivas. Comprender detalladamente las características y preferencias específicas de la audiencia permite una personalización precisa de mensajes, canales de distribución y contenido promocional. Este enfoque maximiza el impacto de las campañas de marketing.

Variables como el origen y la fuente del lead, el comportamiento en el sitio web, la actividad reciente y la interacción con correos electrónicos y llamadas, proporcionan información valiosa sobre las preferencias y el comportamiento del usuario. Analizar estas variables permite identificar patrones y segmentar la audiencia de manera más efectiva, resultando en mensajes más personalizados y campañas más eficaces.

Comprender las actividades y ocupaciones de los clientes, así como sus intereses y motivos para elegir un curso, facilita la creación de contenido relevante y atractivo. Esto mejora la experiencia del usuario y aumenta la probabilidad de conversión de prospectos en clientes. La observación y análisis del comportamiento del usuario en plataformas digitales son esenciales para comprender las preferencias y patrones de navegación.

Implementar herramientas de análisis web proporciona información valiosa sobre cómo los usuarios interactúan con el sitio web. Esto permite identificar el contenido más atractivo y señalar posibles obstáculos en el proceso de registro o inscripción. Este enfoque basado en datos permite una retroalimentación en tiempo real, fundamental para ajustar la experiencia del usuario y optimizar la efectividad de las campañas de captación.

La implementación de estrategias de marketing basadas en datos también involucra el uso inteligente de herramientas de marketing automatizado. Este enfoque permite la personalización y secuenciación de mensajes basados en la interacción pasada del usuario. Los sistemas de gestión de relaciones con el cliente (CRM) impulsados por datos posibilitan un seguimiento detallado de los clientes potenciales, desde el primer contacto hasta la conversión.

Un enfoque de datos en las estrategias de marketing para empresas de e-learning no solo optimiza el gasto publicitario y mejora las tasas de conversión, sino que también fortalece la conexión entre la oferta educativa y las necesidades específicas del público objetivo. Esto establece las bases para relaciones duraderas y exitosas con los estudiantes, posicionando a la empresa de e-learning en un espacio destacado en el panorama educativo digital.

2.Metodología y Selección de Datos para la Optimización Predictiva

Este apartado detalla la metodología y los criterios de selección de datos utilizados para optimizar la predicción de la conversión de leads en empresas de eLearning. El proceso incluye la descripción del conjunto de datos, la limpieza y el análisis exploratorio, la selección de variables clave y la evaluación del rendimiento de los modelos predictivos. La primera fase consistió en la recopilación y descripción del conjunto de datos utilizado. Se empleó un conjunto de datos secundario que contiene información detallada sobre la interacción de clientes potenciales con un sitio web educativo y su posterior conversión en leads. Este conjunto de datos incluye múltiples variables que capturan diversos aspectos del comportamiento de los usuarios. La selección de un conjunto de datos robusto y representativo es crucial para desarrollar modelos predictivos efectivos. En la siguiente fase se realizó la limpieza y análisis exploratorio de datos. Este proceso implicó la identificación y tratamiento de valores faltantes y outliers. Se eliminaron las columnas con una alta proporción de valores faltantes y se imputaron aquellas con proporciones moderadas o bajas. Además, se identificaron y eliminaron los outliers utilizando técnicas estadísticas adecuadas. Este paso es esencial para asegurar que el conjunto de datos final sea representativo y libre de anomalías que podrían sesgar los resultados.

La selección de variables clave se llevó a cabo mediante análisis de correlación y evaluación de la importancia de las variables. Se identificaron las variables más relevantes para la predicción de la conversión de leads, incluyendo tanto datos numéricos como categóricos. La identificación de estas variables clave es crucial para el desarrollo de modelos precisos y fiables.

La evaluación del rendimiento del modelo se realizó utilizando varias métricas. Se seleccionaron métricas como la exactitud, precisión, sensibilidad, F1-Score y el Área Bajo la Curva ROC (AUC-ROC). Cada métrica proporciona una perspectiva diferente del rendimiento del modelo y se elige en función de los objetivos específicos del análisis. Considerar múltiples métricas permite una evaluación más completa y precisa del rendimiento del modelo. Para concluir, la metodología seguida en este estudio incluye la selección y limpieza de datos, el análisis exploratorio, la selección de variables clave y la evaluación del rendimiento del modelo. Este

enfoque integral asegura que los modelos predictivos desarrollados sean de alta calidad y efectividad. Este proceso metodológico garantiza que las estrategias de marketing basadas en estos modelos sean optimizadas y efectivas en la conversión de leads.

2.1 Descripción, Análisis Exploratorio y Limpieza de Base de Datos

En este estudio, se utiliza un conjunto de datos secundario proveniente de Kaggle, específicamente el "X Education Leads Management" dataset. Este conjunto de datos consta de alrededor de 9000 puntos. Ofrece información detallada sobre la interacción de clientes potenciales con el sitio web de X Education y su posterior conversión en leads.

X Education vende cursos en línea a profesionales de la industria. A diario, muchos profesionales interesados visitan su sitio web y navegan por los cursos disponibles. La empresa promociona sus cursos en varios sitios web y motores de búsqueda como Google. Una vez que las personas aterrizan en el sitio web, pueden navegar por los cursos, completar un formulario de inscripción o ver videos. Al completar un formulario con su dirección de correo electrónico o número de teléfono, se clasifican como leads.

La empresa también obtiene leads a través de referencias pasadas. Una vez adquiridos estos leads, los empleados del equipo de ventas comienzan a hacer llamadas y escribir correos electrónicos. Este dataset contiene leads del pasado con aproximadamente 9000 puntos de datos. El "X Education Leads Management" dataset incluye un total de 37 variables. Estas variables capturan diversos aspectos del comportamiento y las características de los clientes potenciales. Son fundamentales para desarrollar modelos predictivos efectivos que mejoren la conversión de leads y optimicen las estrategias de marketing.

Este conjunto de datos proporciona una base sólida para el análisis predictivo. Permite anticipar comportamientos, personalizar experiencias de aprendizaje y optimizar estrategias de marketing. Al analizar estas variables y aplicar modelos predictivos avanzados, es posible mejorar significativamente la conversión de leads y la retención de estudiantes.

El análisis exploratorio y la limpieza de la base de datos son pasos esenciales en cualquier proyecto de análisis de datos. En este contexto, se trabajó con un conjunto de datos compuesto

por 9240 filas y 37 columnas. Cada columna ofrece una perspectiva única sobre los clientes potenciales. Estas columnas se clasificaron en diferentes tipos de datos, incluyendo numéricos y categóricos, cada uno con su propia complejidad y desafíos asociados.

Los datos numéricos se dividen en discretos y continuos. Hay tres variables discretas y cuatro continuas. Los datos categóricos se dividen en nominales y binarios. Hay dieciocho variables nominales y dieciséis binarias.

El primer paso fue identificar y comprender la distribución de los valores faltantes en el conjunto de datos. Se encontró que varias columnas tenían un alto porcentaje de valores faltantes, planteando interrogantes sobre la integridad de los datos y la fiabilidad de los análisis posteriores. La Figura 1 muestra el porcentaje de valores faltantes.

Figura 1: Valores faltantes del DataFrame

Valores faltantes en el DataFrame:

	Valores Faltantes	Porcentaje
Lead Source	36	0.389610
TotalVisits	137	1.482684
Page Views Per Visit	137	1.482684
Last Activity	103	1.114719
Country	2461	26.634199
Specialization	1438	15.562771
How did you hear about X Education	2207	23.885281
What is your current occupation	2690	29.112554
What matters most to you in choosing a course	2709	29.318182
Tags	3353	36.287879
Lead Quality	4767	51.590909
Lead Profile	2709	29.318182
City	1420	15.367965
Asymmetrique Activity Index	4218	45.649351
Asymmetrique Profile Index	4218	45.649351
Asymmetrique Activity Score	4218	45.649351
Asymmetrique Profile Score	4218	45.649351

Fuente: Elaboración propia

Para abordar este desafío, se adoptó un enfoque estratégico. Se diferenciaron las columnas con valores faltantes significativos, moderados y bajos. Las columnas con una alta proporción de valores faltantes (>45%), se consideraron para eliminación. Para las columnas con proporciones moderadas (15%-30%), se propuso la imputación. Las columnas con baja proporción de valores faltantes (<5%) se imputaron con técnicas simples.

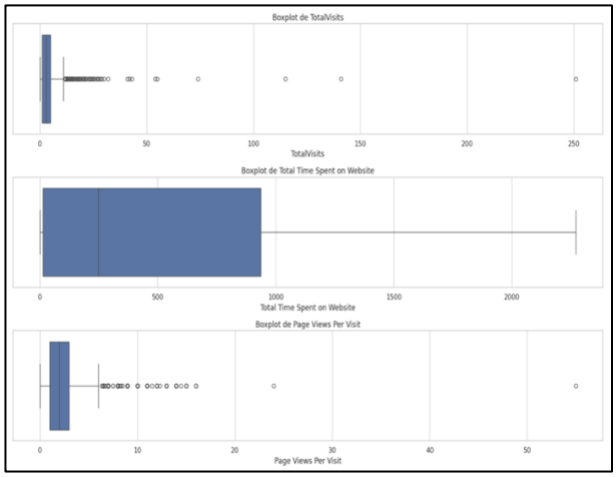
En el proceso de imputación de valores faltantes, se aplicaron diferentes enfoques según el tipo de datos. Para las columnas categóricas, los valores faltantes se imputaron con categorías como

"Unknown" o con la moda. Para las variables numéricas, los valores faltantes se imputaron utilizando la media.

En la identificación y tratamiento de outliers, se siguieron varios pasos. Se identificaron filas con valores atípicos en variables clave utilizando boxplots y el rango intercuartílico (IQR). Luego, se eliminaron las filas con outliers, resultando en un conjunto de datos final con 8679 filas.

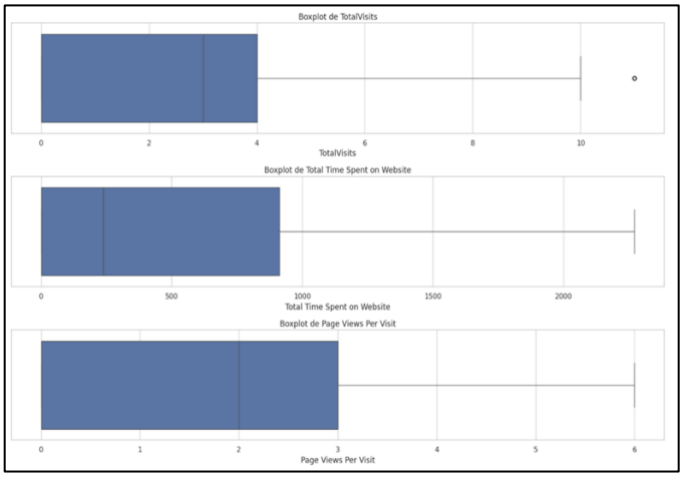
En la identificación y tratamiento de outliers, se siguieron varios pasos. Se identificaron filas con valores atípicos en variables clave utilizando boxplots y el rango intercuartílico (IQR). Luego, se eliminaron las filas con outliers en "TotalVisits" y "Page Views Per Visit", resultando en un conjunto de datos final con 8679 filas. La Figura 2 muestra los outliers antes de la limpieza y la Figura 3 muestra los boxplots después de la limpieza.

Figura 2: Boxplots



Fuente: Elaboración propia

Figura 3: Boxplots nuevos



Fuente: Elaboración propria

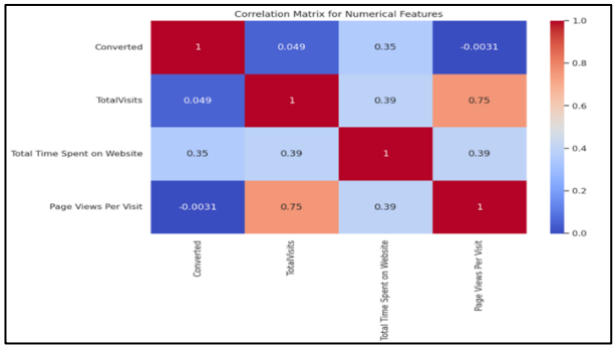
Se realizó una revisión exhaustiva en busca de valores duplicados, confirmando que no había duplicados. Se examinaron las frecuencias de las variables categóricas más relevantes. Se identificaron y eliminaron columnas con una alta proporción de valores nulos o que no aportaban información relevante. También se descartaron variables con una sola categoría.

Se convirtió el Prospect ID en el índice del DataFrame para facilitar el acceso a los datos. El conjunto de datos depurado y enfocado quedó compuesto por 8679 filas y 13 columnas. Este proceso de limpieza y preparación de datos estableció una base sólida para los análisis posteriores y la construcción de modelos predictivos.

2.2 Selección de Variables Clave para Predicción

El análisis de correlación y la importancia de las variables categóricas proporcionan una guía clara para la selección de variables clave para la predicción de la conversión de leads. "La identificación de variables clave es crucial para la construcción de modelos predictivos efectivos. Estudios previos han demostrado que características como el tiempo total en el sitio web y la interacción con el contenido son predictores significativos del rendimiento académico" (Martins et al., 2019). La Figura 4 muestra la matriz de correlación, mientras que la Figura 5 ilustra el análisis de Cramer's V.

Figura 4: Matriz de Correlación



Fuente: Elaboración propia

Figura 5: Análisis de cramers v

```
[('Tags', 0.5648016992949948),
 ('Last Activity', 0.40559296748366297),
 ('Last Notable Activity', 0.3871084055111251),
 ('Lead Source', 0.3463062139551914),
 ('Lead Origin', 0.33624656522081375),
 ('What is your current occupation', 0.31711602925835936),
 ('Do Not Email', 0.130943833441558),
 ('A free copy of Mastering The Interview', 0.03965465000722958)]
```

Fuente: Elaboración propia

La combinación de variables numéricas y categóricas seleccionadas ofrece una amplia gama de información fundamental para construir modelos predictivos efectivos. Entre las Variables Numéricas, Total Time Spent on Website, tiene una correlación moderada positiva con la conversión de leads. Sugiere que cuanto más tiempo pasa un cliente potencial en el sitio web, es más probable que esté interesado en los cursos. Este indicador refleja el compromiso del cliente con el contenido y las ofertas del sitio web, convirtiéndolo en un predictor valioso de la conversión.

Page Views Per Visit, aunque la correlación con la conversión de leads es baja, sigue siendo relevante. Proporciona información adicional sobre el comportamiento de navegación del cliente. Aunque su influencia en la conversión es limitada, es un aspecto para considerar en la experiencia del usuario.

Entre las variables Categóricas, Tags es la más fuertemente asociada con la conversión de leads. Refleja el estado actual del lead y sus interacciones previas con la empresa. Indica el nivel de interés, el grado de compromiso y las necesidades específicas del cliente potencial, lo que lo convierte en un predictor clave de la conversión.

Last Activity proporciona una instantánea del compromiso reciente del cliente con la empresa. Actividades como abrir un correo electrónico, visitar una página web o participar en una conversación indican un interés activo y reciente, haciendo que esta variable sea crucial para predecir la conversión. Last Notable Activity, similar a la variable anterior, proporciona información adicional sobre las interacciones más significativas del cliente con la empresa. Ayuda a priorizar leads y enfocar los esfuerzos de seguimiento en aquellos con un mayor potencial de conversión.

Lead Source indica el canal a través del cual el cliente potencial llegó al sitio web. Algunas fuentes pueden ser más efectivas para generar leads de alta calidad que otras. Conocer la fuente del lead proporciona información sobre la efectividad de las estrategias de marketing y la calidad de los leads generados. Lead Origin proporciona información sobre cómo se inició la interacción con el cliente potencial. Dependiendo del origen (página de destino, referencia, campaña de marketing, etc.), la probabilidad de conversión puede variar. Captura el punto de entrada del cliente potencial en el embudo de ventas, haciendo que sea relevante para predecir la conversión.

What is your current occupation, la ocupación actual del cliente potencial, afecta directamente su capacidad y disponibilidad para inscribirse en cursos. Esta variable demográfica proporciona información valiosa sobre el perfil y las circunstancias del cliente potencial. Do Not Email, la preferencia del cliente sobre recibir correos electrónicos, puede influir en las estrategias de comunicación y marketing utilizadas por la empresa. Aunque su asociación con la conversión es baja, es importante considerar en la personalización de las comunicaciones con los leads.

Total Visits tiene una correlación baja positiva con la conversión de leads. El número total de visitas al sitio web puede indicar el interés del cliente. Cuantas más visitas realice un cliente potencial, es más probable que esté explorando activamente los cursos ofrecidos. Aunque su influencia en la conversión es limitada, sigue siendo relevante para comprender el comportamiento del cliente potencial. Se eliminarán las variables con baja asociación con la conversión de leads, como Lead Number y A free copy of Mastering The Interview, para simplificar el modelo y centrar esfuerzos en las variables más predictivas.

Estas variables seleccionadas proporcionan una combinación equilibrada de información demográfica, comportamental y de interacción, esencial para predecir la conversión de leads. Al incluir estas variables en los modelos predictivos, es probable obtener resultados más precisos y diseñar estrategias de marketing más efectivas.

2.3 Métricas de Evaluación del Rendimiento del Modelo

El proceso de evaluación del rendimiento del modelo es fundamental en la construcción de sistemas de aprendizaje automático efectivos. En el contexto de un problema de clasificación binaria como la conversión de leads, donde la variable objetivo es "Converted", es esencial seleccionar las métricas de evaluación adecuadas. Estas métricas ayudan a comprender cómo se está desempeñando el modelo y qué tan bien está cumpliendo su propósito.

Una de las métricas más comunes es la exactitud (Accuracy), que calcula la proporción de predicciones correctas sobre el total de predicciones realizadas por el modelo. Aunque la exactitud es fácil de entender e interpretar, puede ser engañosa en conjuntos de datos desbalanceados, donde una clase es mucho más frecuente que la otra. En tales casos, una métrica más informativa es la precisión (Precision), que mide la proporción de verdaderos positivos sobre el total de predicciones positivas realizadas por el modelo.

La precisión es especialmente útil cuando el costo de los falsos positivos es alto, como en el caso de enviar correos electrónicos de seguimiento a leads que no se convertirán. Otra métrica importante es el recall (Sensibilidad o Tasa de Verdaderos Positivos), que indica la proporción de verdaderos positivos sobre el total de positivos reales en los datos. El recall es crítico cuando

el costo de los falsos negativos es alto, ya que se centra en la capacidad del modelo para capturar todos los positivos reales, incluso a costa de tener más falsos positivos.

El F1-Score combina la precisión y el recall en una sola métrica, calculando la media armónica de ambas. Esta métrica proporciona un equilibrio entre precisión y recall, y es útil cuando se necesita considerar ambas en la evaluación del modelo. El F1-Score es particularmente valioso cuando se requiere una medida única que refleje tanto la precisión como la capacidad del modelo para identificar correctamente los positivos.

Finalmente, el Área Bajo la Curva ROC (AUC-ROC) es una métrica que evalúa la capacidad del modelo para distinguir entre clases. La curva ROC representa la relación entre la tasa de verdaderos positivos (TPR) y la tasa de falsos positivos (FPR) para diferentes umbrales de clasificación. Un AUC-ROC de 1 indica un modelo perfecto, mientras que un valor de 0.5 indica un modelo que no tiene capacidad de discriminación entre las clases.

Las Figuras 6 a 10 ilustran las fórmulas de estas métricas clave. La exactitud (Figura 6), precisión (Figura 7), sensibilidad (Figura 8), F1-Score (Figura 9) y AUC-ROC (Figura 10) proporcionan una comprensión detallada del rendimiento del modelo en diversos aspectos. Cada métrica tiene su importancia y aplicación dependiendo del contexto y los objetivos del análisis.

Figura 6: Formula de Exactitud

$$\text{Exactitud} = \frac{\text{Número de predicciones correctas}}{\text{Número total de predicciones}}$$

Fuente: Elaboración propia

Figura 7: Formula de precisión

$$\text{Precisión} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Positivos (FP)}}$$

Fuente: Elaboración propia

Figura 8: Formula de sensibilidad

$$\text{Recall} = \frac{\text{Verdaderos Positivos (TP)}}{\text{Verdaderos Positivos (TP)} + \text{Falsos Negativos (FN)}}$$

Fuente: Elaboración propia

Figura 9: Formula de F1-Score

$$\text{F1-Score} = 2 \times \frac{\text{Precisión} \times \text{Recall}}{\text{Precisión} + \text{Recall}}$$

Fuente: Elaboración propia

Figura 10: Formula de AUC-ROC

$$\text{AUC-ROC} = \text{Área bajo la curva ROC}$$

Fuente: Elaboración propia

Al seleccionar las métricas de evaluación del rendimiento del modelo, es importante considerar el contexto del problema, los costos asociados con los diferentes tipos de errores y los objetivos específicos del análisis. Esto permite una evaluación más completa y precisa del rendimiento del modelo en la tarea de conversión de leads. Considerar múltiples métricas y su relevancia en el contexto del problema es crucial para tomar decisiones informadas sobre la efectividad del modelo y su aplicación práctica.más completa y precisa del rendimiento del modelo en la tarea de conversión de leads.

3.Preparación, Selección y Evaluación de Modelos Predictivos

Para lograr modelos predictivos efectivos y precisos en el análisis de conversión de leads, es fundamental seguir un proceso estructurado que abarca desde la preparación de datos hasta la selección y evaluación de los modelos. Este apartado se centra en tres subapartados principales. El primero será el Proceso de Preparación de Datos para Análisis Predictivo. Aquí se aborda la transformación de los datos brutos en un formato adecuado para los algoritmos de aprendizaje automático. Este proceso incluye la codificación de variables categóricas, el escalado de características numéricas y la división de los datos en conjuntos de entrenamiento y prueba.

Luego, en la Selección de Modelos de Aprendizaje Automático para la Predicción de Conversión de Clientes, se exploran y comparan varios modelos de aprendizaje automático. Se evaluarán modelos como la regresión logística, árboles de decisión, Random Forest, Gradient Boosting, Support Vector Machine (SVM) y K-Nearest Neighbors (KNN). Cada modelo será evaluado por su rendimiento en la tarea de predecir la conversión de leads.

También se realizará el Ajuste de Hiperparámetros de los Modelos Seleccionados. En esta fase, se optimizan los modelos seleccionados ajustando sus hiperparámetros mediante técnicas como la búsqueda en cuadrícula. El objetivo es mejorar su precisión y eficacia en la predicción. Esta optimización es crucial para maximizar el rendimiento de los modelos en escenarios reales.

Finalmente, se presenta el Análisis de Errores de los Modelos Gradient Boosting y Random Forest. Este análisis busca identificar áreas de mejora y optimizar las estrategias de marketing. Comprender los errores del modelo permite ajustar y mejorar las técnicas predictivas, asegurando una mayor precisión en futuras predicciones.

3.1 Proceso de Preparación de Datos para Análisis Predictivo

El proceso de preparación de datos para el análisis predictivo es esencial para crear modelos de aprendizaje automático efectivos y precisos. Este proceso incluye varias etapas para garantizar que los datos estén en la forma óptima para los algoritmos de predicción.

Las variables categóricas representan características no numéricas, como el género, el estado civil o la profesión. Los algoritmos de aprendizaje automático requieren que los datos de entrada sean numéricos. Por lo tanto, es necesario convertir estas variables categóricas en valores numéricos. En este estudio, las variables categóricas 'Lead Origin', 'Lead Source', 'Do Not Email', 'Last Activity', 'What is your current occupation', 'Tags' y 'Last Notable Activity' fueron codificadas usando Label Encoding. El Label Encoding asigna a cada categoría un valor numérico único. Esto facilita la comprensión de los algoritmos y captura las relaciones entre las variables de manera adecuada.

Las variables numéricas pueden tener escalas diferentes, afectando el rendimiento de algunos algoritmos de aprendizaje automático. Algoritmos sensibles a la escala de los datos, como K-NN o SVM, pueden producir resultados subóptimos si las características no están en el mismo rango. Para abordar este problema, se aplican técnicas de normalización o estandarización para escalar las características numéricas a un rango común. En este estudio, las variables numéricas 'TotalVisits', 'Total Time Spent on Website' y 'Page Views Per Visit' fueron escaladas utilizando StandardScaler. La normalización ajusta los valores de las características a un rango entre 0 y 1, mientras que la estandarización transforma los datos para que tengan una media de 0 y una desviación estándar de 1.

La división de los datos en conjuntos de entrenamiento y prueba es esencial para evaluar la capacidad predictiva del modelo. Esta división permite entrenar el modelo en un conjunto de datos y luego evaluar su rendimiento en datos no vistos. Generalmente, se asigna una parte de los datos para el entrenamiento y otra para la prueba, comúnmente en una proporción de 70-30 o 80-20. Esto ayuda a validar el modelo y ajustar sus parámetros según sea necesario. Además, la validación cruzada puede ser implementada para asegurar que el modelo generalice bien a datos nuevos, dividiendo los datos en múltiples subconjuntos y entrenando el modelo varias veces con diferentes combinaciones de datos de entrenamiento y prueba.

El proceso de preparación de datos para el análisis predictivo incluye la codificación de variables categóricas, el escalado de características numéricas y la división de los datos en conjuntos de entrenamiento y prueba. Estas tareas garantizan que los datos estén en la forma adecuada para los algoritmos de aprendizaje automático, contribuyendo al desarrollo de modelos predictivos

precisos y confiables. Una preparación cuidadosa de los datos es esencial para maximizar el rendimiento y la precisión de los modelos predictivos, asegurando que sean robustos y efectivos en la práctica.

3.2 Selección de Modelos y Ajuste de Hiperparámetros

Se probaron varios modelos de aprendizaje automático para predecir la conversión de clientes debido a su diversidad en términos de algoritmos y capacidades predictivas. "La evaluación de diversos modelos de aprendizaje automático, incluyendo la regresión logística, árboles de decisión y bosques aleatorios, es fundamental para determinar el mejor enfoque predictivo. Investigaciones han resaltado la precisión de estos modelos en contextos educativos" (McCarthy et al., 2019).

Aquí está una breve explicación de cada uno de los modelos utilizados. La regresión logística se utiliza comúnmente para problemas de clasificación binaria. Estima la probabilidad de que una instancia pertenezca a una clase particular utilizando la función logística. Es un modelo simple y fácil de interpretar, adecuado como punto de partida en problemas de clasificación.

El árbol de decisión divide el conjunto de datos en subconjuntos más pequeños basados en características específicas, formando un árbol de decisiones. Cada nodo interno representa una característica, cada rama representa una regla de decisión y cada hoja representa el resultado. Los árboles de decisión son fáciles de entender e interpretar y pueden capturar relaciones no lineales en los datos.

El Random Forest es una extensión de los árboles de decisión donde se entrena una variedad de árboles de decisión en diferentes subconjuntos de datos y se promedian los resultados para mejorar la precisión y reducir el sobreajuste. Es robusto, preciso y puede manejar grandes conjuntos de datos con alta dimensionalidad.

Gradient Boosting crea un conjunto de modelos de aprendizaje débiles, generalmente árboles de decisión poco profundos, y los entrena secuencialmente para corregir los errores del modelo anterior. A diferencia de Random Forest, se enfoca en minimizar el error de predicción y no en la reducción de la varianza. Es altamente preciso y resistente al sobreajuste.

El Support Vector Machine (SVM) es un algoritmo de aprendizaje supervisado que encuentra el hiperplano de separación óptimo entre clases en un espacio multidimensional. Puede manejar datos lineales y no lineales y es efectivo en espacios de alta dimensionalidad. SVM busca el margen máximo entre las clases y es particularmente útil en conjuntos de datos con pocas características.

El K-Nearest Neighbors (KNN) clasifica una instancia basada en la mayoría de las clases de sus vecinos más cercanos en el espacio de características. No requiere entrenamiento explícito y es simple de entender e implementar. Sin embargo, puede ser computacionalmente costoso para grandes conjuntos de datos y sensible a la elección del número de vecinos (K).

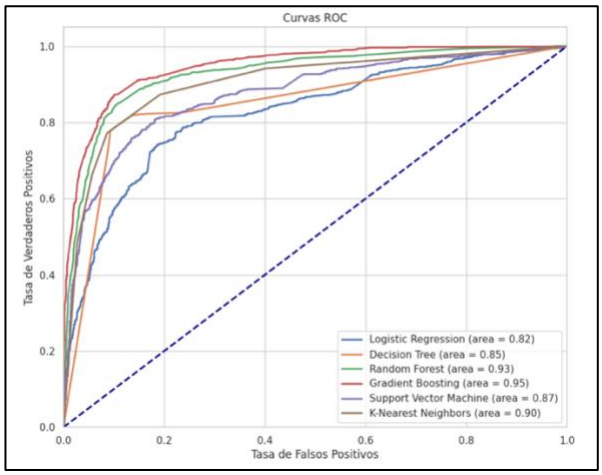
Las Figuras 11 y 12 ilustran las métricas de evaluación de los modelos y la curva ROC, respectivamente. La Figura 11 muestra la precisión, precisión, recall, F1-Score y AUC-ROC de cada modelo evaluado. La Figura 12 muestra la curva ROC, que ayuda a visualizar el rendimiento de los modelos en términos de la tasa de verdaderos positivos frente a la tasa de falsos positivos.

Figura 11: Métricas de evaluación de los modelos

Modelo	**Accuracy**	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
Logistic Regression	0.77	0.77	0.60	0.67	0.82
Decision Tree	0.86	0.84	0.78	0.81	0.85
Random Forest	0.87	0.86	0.80	0.83	0.93
Gradient Boosting	0.88	0.87	0.83	0.85	0.95
Support Vector Machine	0.82	0.82	0.68	0.75	0.87
K-Nearest Neighbors	0.86	0.85	0.77	0.81	0.90

Fuente: Elaboración propia

Figura 12: Curva de ROC



Fuente: Elaboración propia

La elección de Gradient Boosting y Random Forest como los dos mejores modelos se basa en una evaluación exhaustiva de su rendimiento en varias métricas clave. Estas métricas proporcionan una visión completa de cómo cada modelo se desempeña en la tarea de predecir la conversión de leads y ayudan a determinar cuál es más adecuado para este propósito.

El modelo de Gradient Boosting se destaca por su excepcional rendimiento en varias métricas importantes. Con una exactitud del 88%, demuestra una capacidad sólida para realizar predicciones precisas sobre la conversión de leads. Además, su F1-Score del 85% indica un buen equilibrio entre precisión y recall, lo que significa que el modelo es capaz de identificar correctamente la mayoría de las conversiones mientras mantiene un bajo número de falsos positivos. La métrica AUC-ROC, que mide la capacidad del modelo para discriminar entre las clases positivas y negativas, alcanza un impresionante 95%, lo que sugiere una alta capacidad de clasificación. Su recall del 83% muestra que el modelo puede capturar una proporción significativa de conversiones verdaderas, lo que lo hace altamente efectivo en la identificación de leads potenciales.

Aunque ligeramente inferior al Gradient Boosting, Random Forest sigue siendo un modelo sólido con un rendimiento notable. Con una exactitud del 87%, demuestra una capacidad considerable para realizar predicciones precisas. Su F1-Score del 83% indica un buen equilibrio entre precisión y recall, lo que significa que el modelo puede identificar correctamente una gran

proporción de conversiones mientras mantiene un bajo número de falsos positivos. La métrica AUC-ROC alcanza un valor de 93%, lo que indica una capacidad sólida para discriminar entre las clases positivas y negativas.

Ambos modelos ofrecen un rendimiento excelente en la tarea de predecir la conversión de leads. Sin embargo, el Gradient Boosting destaca ligeramente debido a su rendimiento general superior en términos de métricas clave, especialmente en AUC-ROC, lo que indica una mejor capacidad para distinguir entre las clases positivas y negativas. Esto hace que el Gradient Boosting sea la opción preferida cuando se busca el mejor rendimiento global y la máxima capacidad de clasificación.

Durante la fase de ajuste de hiperparámetros mediante la búsqueda en cuadrícula, se identificaron los conjuntos óptimos de hiperparámetros para los modelos Gradient Boosting y Random Forest. Para el modelo Gradient Boosting, se determinaron los siguientes valores de hiperparámetros: n_estimators=300 (número de árboles en el modelo), learning_rate=0.1 (tasa de aprendizaje), max_depth=4 (profundidad máxima de cada árbol) y subsample=0.9 (fracción de muestras utilizadas para entrenar cada árbol).

En cuanto al modelo Random Forest, los hiperparámetros elegidos fueron: n_estimators=300 (número de árboles en el bosque), max_depth=20 (profundidad máxima de cada árbol), min_samples_split=5 (número mínimo de muestras requeridas para dividir un nodo), min_samples_leaf=2 (número mínimo de muestras requeridas para estar en un nodo hoja) y max_features='sqrt' (número de características a considerar para la mejor división).

Figura 13: Métricas de Evaluación

	Metric	Gradient Boosting	Random Forest
0	Accuracy	0.895161	0.884793
1	Precision	0.875758	0.874608
2	Recall	0.852507	0.823009
3	F1-Score	0.863976	0.848024
4	AUC-ROC	0.958459	0.948333

Fuente: Elaboración Propia

Estos conjuntos óptimos de hiperparámetros fueron seleccionados después de una exhaustiva exploración de diversas combinaciones, con el objetivo de maximizar el rendimiento predictivo de los modelos. En la figura 13, se puede observar, los resultados de esta búsqueda meticulosa de hiperparámetros revelaron conjuntos óptimos que condujeron a mejoras significativas en el rendimiento predictivo de ambos modelos.

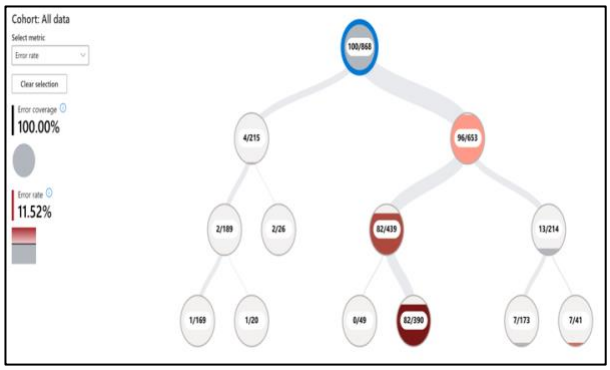
Estos resultados demuestran que la búsqueda de los hiperparámetros óptimos mediante técnicas de ajuste finas como la búsqueda en cuadrícula puede llevar a mejoras significativas en el rendimiento de los modelos de aprendizaje automático, resultando en predicciones más precisas y fiables.

3.3 Matriz de Confusión, Análisis de Importancias y Análisis de Errores sobre Ambos Modelos

El análisis de los errores se centra en identificar patrones en los datos donde los modelos fallan en la clasificación correcta. Las siguientes figuras muestran el análisis de los nodos de error para ambos modelos, destacando áreas específicas donde se concentran los errores. Las Figuras 14 a 23 ilustran cómo se distribuyen los errores en diferentes nodos para los modelos Random Forest y Gradient Boosting.

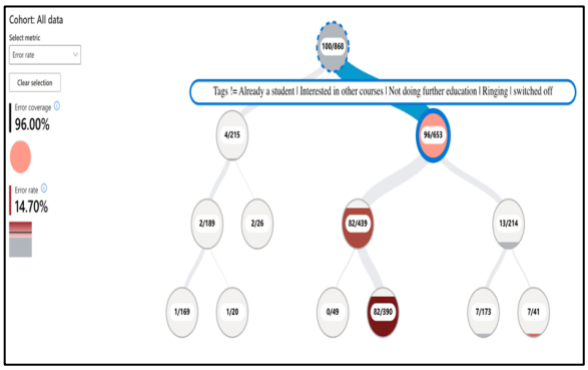
La Figura 14 indica una cobertura de errores del 100% y una tasa de error del 11.52%. El nodo raíz divide los datos en función de las etiquetas (Tags), con una alta concentración de errores en la rama derecha. Esta división inicial sugiere que las etiquetas tienen un impacto significativo en la clasificación incorrecta. La Figura 15 muestra una cobertura de errores del 96% y una tasa de error del 14.70%. Similar a la primera figura, la división principal se basa en las etiquetas (Tags). La alta concentración de errores en la rama derecha, que incluye puntos de datos con etiquetas específicas, sugiere que ciertas etiquetas están fuertemente asociadas con clasificaciones incorrectas.

Figura 14: Análisis de errores de Random Forest



Fuente: Elaboración propia

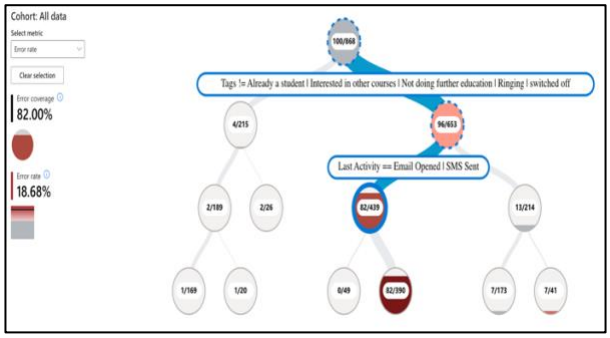
Figura 15: Análisis de errores de Random Forest



Fuente: Elaboración propia

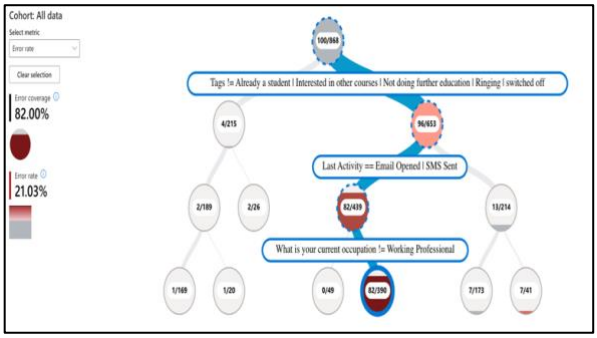
La Figura 16 presenta una cobertura de errores del 82% y una tasa de error del 18.68%. En esta figura, la Última Actividad ("Correo Abierto" o "SMS Enviado") está asociada con una alta tasa de errores. Los errores se concentran en una rama específica, indicando que estas actividades son factores críticos que contribuyen a las clasificaciones incorrectas. La Figura 17 muestra una cobertura de errores del 82% y una tasa de error del 21.03%. La división principal sigue siendo las etiquetas (Tags), pero se observa una alta concentración de errores en la rama derecha cuando la Última Actividad es "Correo Abierto" o "SMS Enviado" y la Ocupación Actual es diferente de "Profesional en Ejercicio". Este patrón resalta la necesidad de ajustar las estrategias basadas en la ocupación.

Figura 16: Análisis de errores de Random Forest



Fuente: Elaboración propia

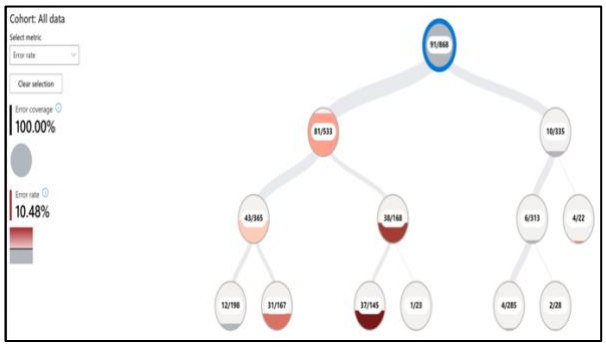
Figura 17: Análisis de errores de Random Forest



Fuente: Elaboración propia

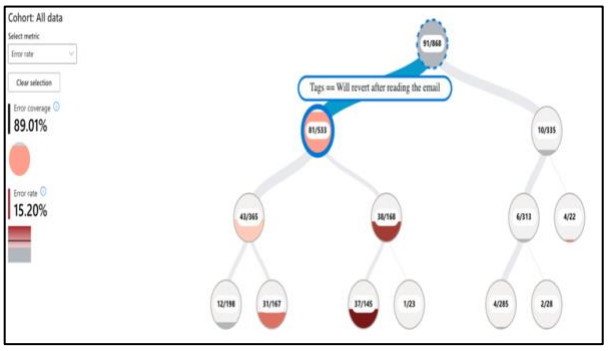
Por otro lado, se analiza los errores del modelo de Gradient Boosting. En la Figura 18, el nodo raíz muestra que todos los puntos de datos están cubiertos por el análisis de errores. La tasa de error en este nodo es del 10.48%. El árbol se divide inicialmente en función de las etiquetas (Tags), con una alta concentración de errores en la rama derecha, sugiriendo que las etiquetas son críticas para la clasificación incorrecta. En la Figura 19, la división principal se basa en las etiquetas (Tags), específicamente "Will revert after reading the email". La rama derecha muestra una alta concentración de errores con una tasa de error del 15.20%. Los errores están concentrados en nodos secundarios, lo que indica que estas etiquetas son factores significativos en las clasificaciones incorrectas.

Figura 18: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

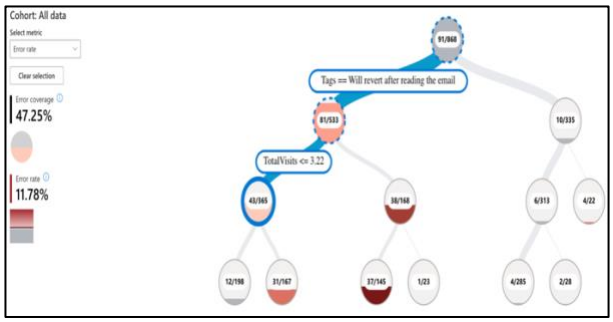
Figura 19: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

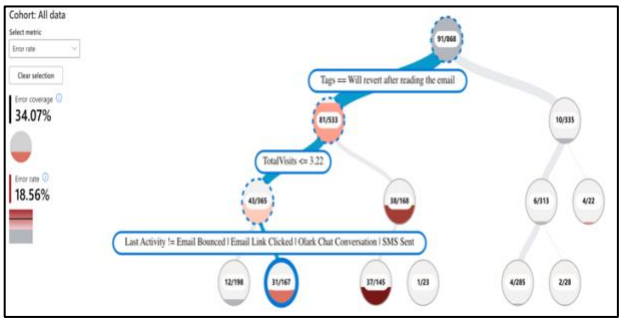
La Figura 20 muestra que la combinación de etiquetas igual a "Will revert after reading the email" y TotalVisits <= 3.22 tiene una tasa de error del 11.78%. La alta concentración de errores en estos nodos sugiere que estas variables son críticas para la clasificación incorrecta y requieren un enfoque detallado. La Figura 21 presenta una cobertura de errores del 34.07% y una tasa de errores del 18.56%. Esta figura destaca la combinación de etiquetas igual a "Will revert after reading the email" y TotalVisits <= 3.22, junto con Last Activity != "Email Bounced | Email Link Clicked | Olark Chat Conversation | SMS Sent". Estos nodos subrayan áreas clave donde se concentran los errores y necesitan atención.

Figura 20: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

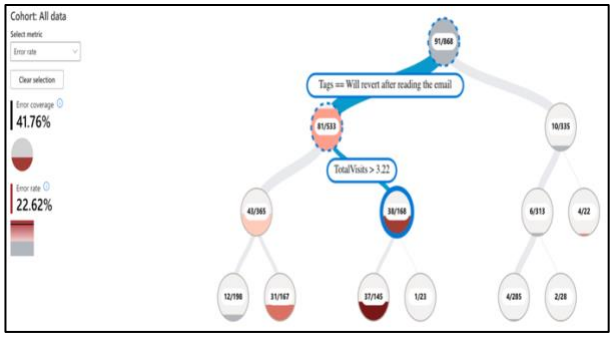
Figura 21: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

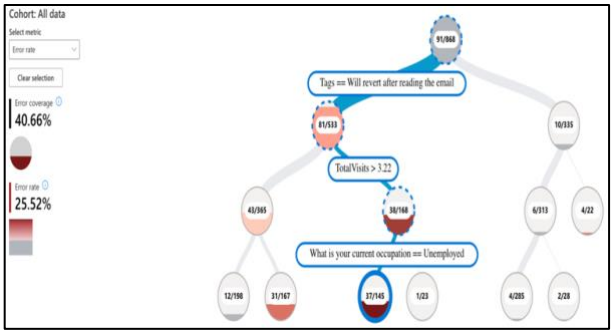
La Figura 22 muestra una cobertura de errores del 41.76% y una tasa de errores del 22.62% cuando Tags es igual a "Will revert after reading the email" y TotalVisits > 3.22. Este nodo es crucial para entender las áreas de mejora en la clasificación. La ocupación también juega un papel importante en este análisis. La Figura 23 presenta una cobertura de errores del 40.66% y una tasa de error del 25.52%. Aquí, Tags igual a "Will revert after reading the email" y TotalVisits > 3.22, junto con "What is your current occupation == Unemployed", muestra una alta tasa de error, destacando la importancia de la ocupación y la cantidad de visitas en la clasificación incorrecta.

Figura 22: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

Figura 23: Análisis de errores de Gradient Boosting



Fuente: Elaboración propia

Para mejorar el rendimiento de los modelos Gradient Boosting y Random Forest en la predicción de la conversión de leads, se identifican varias áreas críticas que requieren atención. La gestión

de etiquetas (Tags) es esencial ya que están asociadas con altas tasas de error en ambos modelos. Es necesario revisar y refinar estas etiquetas para mejorar la precisión de la clasificación. La alta concentración de errores en nodos relacionados con etiquetas específicas sugiere que estas variables son cruciales para la predicción correcta.

La última actividad, como "Correo Abierto" y "SMS Enviado", está vinculada a una alta tasa de errores. Revaluar la relevancia de estas actividades en el modelo y ajustar las estrategias de seguimiento es crucial. La alta concentración de errores en ramas específicas indica que estas actividades son factores críticos en las clasificaciones incorrectas. La ocupación actual muestra que ocupaciones distintas de "Profesional en Ejercicio" tienen altas tasas de error. Es necesario ajustar las estrategias basadas en la ocupación y considerar la introducción de características adicionales para mejorar la precisión del modelo. Los patrones de error relacionados con la ocupación actual resaltan la importancia de esta variable en la clasificación correcta.

Las visitas totales (TotalVisits) muestran que la combinación de etiquetas con el número de visitas al sitio web revela áreas críticas de error. Ajustar la importancia de estas variables en los modelos puede contribuir a mejorar el rendimiento. La alta concentración de errores en nodos con visitas totales específicas sugiere que estas variables son críticas para la clasificación correcta. Al enfocar los esfuerzos en estas áreas y ajustar los modelos en consecuencia, se puede mejorar significativamente la precisión y efectividad de las estrategias de marketing. Esto no solo aumentará la conversión de leads en clientes, sino que también optimizará el rendimiento general de las campañas de marketing en el sector de eLearning.

4.Visualización de resultados

En esta sección se presentan y analizan los resultados obtenidos de la aplicación de los modelos predictivos en el contexto de las empresas de e-learning. Se abordarán cuatro subapartados principales que permiten una comprensión detallada del desempeño de los modelos, la importancia de las variables y las métricas clave utilizadas para evaluar la efectividad de las predicciones.

Primero, se examinan las matrices de confusión de los modelos Gradient Boosting y Random Forest, proporcionando una visión clara de los verdaderos positivos, falsos positivos, falsos negativos y verdaderos negativos, lo cual revela errores en la clasificación que necesitan corrección.

Segundo, se realiza un análisis de la importancia de las características para entender cuáles variables tienen un mayor impacto en las predicciones de los modelos, destacando la relevancia de variables como "Tags" y "Total Time Spent on Website".

Tercero, se presentan y analizan las métricas de rendimiento, incluyendo exactitud, precisión, sensibilidad, F1-Score y AUC-ROC, que muestran que Gradient Boosting supera ligeramente a Random Forest en todas las métricas evaluadas.

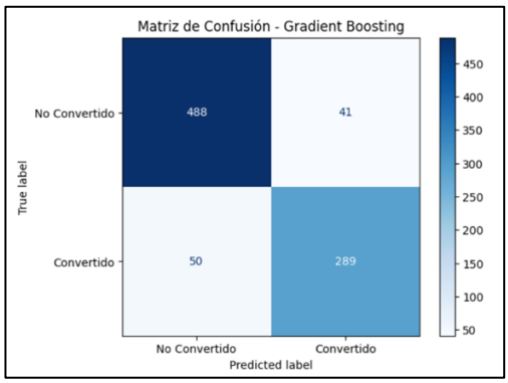
Finalmente, se analiza el impacto de las variables de marketing que más contribuyen a las conversiones, identificando que el tiempo total en el sitio web, las etiquetas de seguimiento positivo y las actividades recientes significativas son claves para optimizar estrategias de marketing y mejorar las tasas de conversión.

4.1 Matrices de Confusión, Análisis de la Importancia de las Características y Métricas de rendimiento

Las matrices de confusión proporcionan una visión clara de los verdaderos positivos (TP), falsos positivos (FP), falsos negativos (FN) y verdaderos negativos (TN) para ambos modelos. A continuación se presentan las matrices de confusión de Gradient Boosting y Random Forest:

En la figura 24 se puede observar que el modelo de Gradient Boosting identificó correctamente 488 leads que se convirtieron en clientes. Clasificó incorrectamente 41 leads como convertidos cuando en realidad no lo eran. Estos son casos donde el modelo predijo que un lead se convertiría en cliente, pero no fue así. Clasificó incorrectamente 50 leads como no convertidos cuando en realidad se convirtieron. Finalmente, identificó correctamente 289 leads que no se convirtieron en clientes.

Figura 24: Matriz de confusión – Gradient Boosting

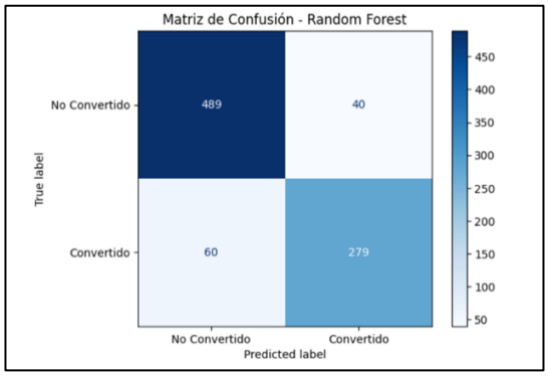


Fuente: Elaboración Propia

El modelo de Gradient Boosting tiene una menor cantidad de falsos positivos en comparación con los falsos negativos. Esto sugiere que es más conservador al predecir conversiones. Sin embargo, hay una cantidad significativa de falsos negativos, lo que indica que algunos leads potenciales no están siendo identificados correctamente como convertidos.

En la figura 25 se puede observar que el modelo de Random Forest identificó correctamente 489 leads que se convirtieron en clientes. Clasificó incorrectamente 40 leads como convertidos cuando en realidad no lo eran. Clasificó incorrectamente 60 leads como no convertidos cuando en realidad se convirtieron. Identificó correctamente 279 leads que no se convirtieron en clientes.

Figura 25: Matriz de confusión – Random Forest



Fuente: Elaboración Propia

El modelo de Random Forest muestra una mayor cantidad de falsos negativos en comparación con Gradient Boosting. Esto indica que Random Forest podría estar perdiendo más oportunidades de conversión al no identificar correctamente algunos leads que se convierten en clientes. De estas matrices se puede interpretar que ambos modelos tienen un buen desempeño en la clasificación de casos positivos (leads convertidos) y negativos (leads no convertidos). Sin embargo, ambos presentan errores que necesitan ser abordados para mejorar la precisión.

El análisis de la importancia de las características es crucial para entender cuáles variables tienen un mayor impacto en las predicciones de los modelos. La tabla a continuación muestra la importancia de las características para ambos modelos. En la tabla, se observa que la característica "Tags" tiene una alta importancia en ambos modelos, pero es particularmente significativa en Gradient Boosting.

Figura 26: Análisis de importancia

Característica	**Importancia Gradient Boosting**	**Importancia Random Forest**
Lead Origin	0,185139	0,089378
Lead Source	0,013247	0,066757
Do Not Email	0,011224	0,01198
TotalVisits	0,015802	0,039923

Total Time Spent on Website	0,282951	0,276483
Page Views Per Visit	0,020303	0,039022
Last Activity	0,024313	0,076927
What is your current occupation	0,03183	0,05758
Tags	0,317084	0,227929
Last Notable Activity	0,098106	0,11402

Fuente: Elaboración propia

De la tabla se puede observar que la característica "Tags" tiene una alta importancia en ambos modelos, pero es particularmente significativa en Gradient Boosting. "Total Time Spent on Website" también es una característica crucial para ambos modelos. Por otro lado, "Lead Source" y "Last Activity" tienen más importancia en Random Forest en comparación con Gradient Boosting.

En esta sección, se presentan y analizan las métricas de rendimiento de dos modelos de clasificación: Random Forest y Gradient Boosting. Estos modelos se han entrenado y evaluado para predecir la conversión de leads en el contexto de una empresa de eLearning. Las métricas de rendimiento incluyen Accuracy (Exactitud), Precision (Precisión), Recall (Sensibilidad), F1-Score y AUC-ROC. A continuación, se muestran los resultados obtenidos, junto con gráficos comparativos y una interpretación detallada de cada métrica. La Figura 27 presenta una tabla comparativa de las métricas de rendimiento de los modelos Random Forest y Gradient Boosting. Estas métricas son fundamentales para evaluar la efectividad de los modelos en predecir la conversión de leads. A continuación se detallan cada una de las métricas y su interpretación:

Figura 27: Métricas de rendimiento

	Metric	Gradient Boosting	Random Forest
0	Accuracy	0.895161	0.884793
1	Precision	0.875758	0.874608
2	Recall	0.852507	0.823009
3	F1-Score	0.863976	0.848024
4	AUC-ROC	0.958459	0.948333

Fuente: Elaboración propia

"Total Time Spent on Website" también es una característica crucial para ambos modelos. Por otro lado, "Lead Source" y "Last Activity" tienen más importancia en Random Forest en comparación con Gradient Boosting. Estas diferencias en la importancia de las características reflejan cómo cada modelo procesa y utiliza la información disponible.

En esta sección, se presentan y analizan las métricas de rendimiento de dos modelos de clasificación: Random Forest y Gradient Boosting. Estos modelos se han entrenado y evaluado para predecir la conversión de leads en el contexto de una empresa de eLearning. Las métricas de rendimiento incluyen Accuracy (Exactitud), Precision (Precisión), Recall (Sensibilidad), F1-Score y AUC-ROC.

La exactitud mide la proporción de predicciones correctas (tanto verdaderos positivos como verdaderos negativos) sobre el total de predicciones. En este caso, el modelo Gradient Boosting tiene una ligera ventaja sobre el modelo Random Forest, indicando que Gradient Boosting realiza predicciones correctas con mayor frecuencia. La precisión es la proporción de verdaderos positivos sobre el total de predicciones positivas.

Una mayor precisión implica menos falsos positivos. Ambos modelos tienen valores muy similares, pero Gradient Boosting muestra una ligera mejora en la precisión. El recall es la proporción de verdaderos positivos sobre el total de positivos reales. Indica la capacidad del modelo para capturar todos los positivos reales.

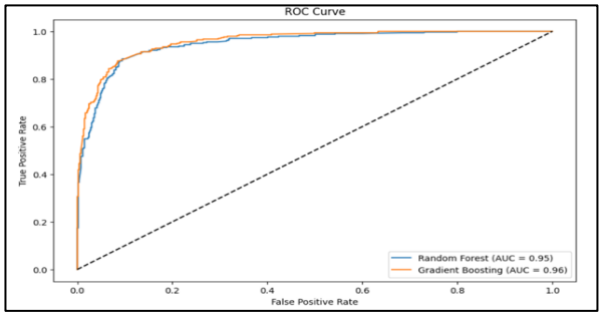
Gradient Boosting supera a Random Forest en esta métrica, lo que sugiere que captura una mayor proporción de los leads que realmente se convierten. El F1-Score es la media armónica

de la precisión y el recall. Proporciona un balance entre estas dos métricas, siendo útil cuando se necesita considerar tanto los falsos positivos como los falsos negativos.

Gradient Boosting tiene un F1-Score más alto, indicando un mejor balance entre precisión y recall. El AUC-ROC (Área Bajo la Curva - Curva ROC) mide la capacidad del modelo para distinguir entre clases positivas y negativas. Un valor más cercano a 1 indica un mejor rendimiento.

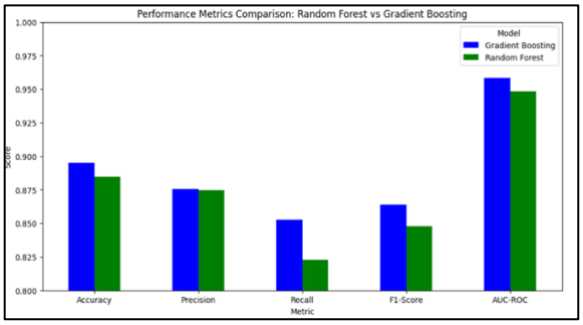
Ambos modelos muestran un excelente rendimiento, pero Gradient Boosting tiene una ligera ventaja, lo que indica una mejor capacidad para discriminar entre clases. La Figura 28 presenta la curva ROC de ambos modelos, mientras que la Figura 29 proporciona una representación gráfica de la comparación de las métricas de rendimiento entre los modelos Random Forest y Gradient Boosting.

Figura 28: Curva ROC



Fuente: Elaboración propia

Figura 29: Comparación de métricas



Fuente: Elaboración propia

La Figura 28 ilustra la curva ROC, mostrando la relación entre la tasa de verdaderos positivos y la tasa de falsos positivos para ambos modelos. Un AUC-ROC más alto sugiere una mejor capacidad del modelo para distinguir entre clases positivas y negativas. La Figura 29, por otro lado, compara las métricas clave de rendimiento de ambos modelos, facilitando la visualización de las diferencias en su rendimiento.

Se visualiza claramente que el modelo de Gradient Boosting tiene un rendimiento superior en todas las métricas evaluadas en comparación con el modelo de Random Forest. Esta representación gráfica ayuda a destacar las áreas donde Gradient Boosting sobresale, facilitando

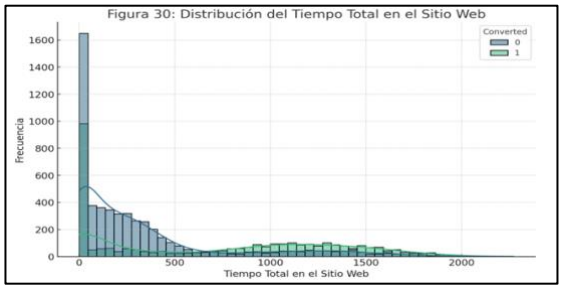
la toma de decisiones informadas para optimizar estrategias de marketing y mejorar las tasas de conversión. Evaluar y entender estas métricas es fundamental para mejorar continuamente los modelos predictivos y asegurar que se están tomando las mejores decisiones posibles basadas en datos precisos y fiables.

4.2 Variables de Marketing que Contribuyen a las Conversiones

A continuación, se presenta el análisis detallado de las variables de marketing que más contribuyen a las conversiones. Este análisis se apoya en gráficos que ilustran la distribución y el impacto de estas variables clave. El tiempo total que un cliente potencial pasa en el sitio web es una de las variables más importantes tanto para Random Forest como para Gradient Boosting. Este tiempo refleja el nivel de interés del usuario en los cursos ofrecidos. La mayoría de los usuarios convertidos pasan un tiempo considerable en el sitio web, lo que indica un mayor compromiso y una mayor probabilidad de conversión (Figura 30).

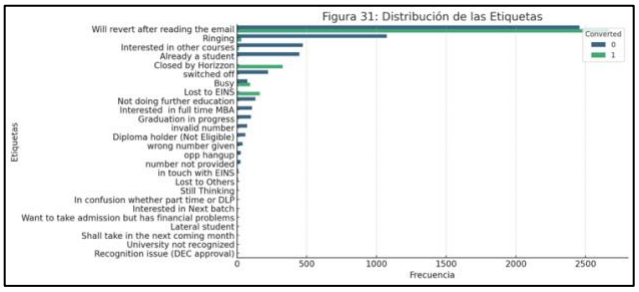
Las etiquetas asignadas a los leads son cruciales para identificar su estado actual y nivel de interés. Las categorías principales incluyen "Will revert after reading the email", "Ringing", "Interested in other courses", "Already a student", y "Closed by Horizzon". Estas etiquetas ayudan a segmentar los leads y a enfocar las estrategias de conversión (Figura 31).

Figura 30: Distribución del Tiempo



Fuente: Elaboración propia

Figura 31: Distribución de las Etiquetas

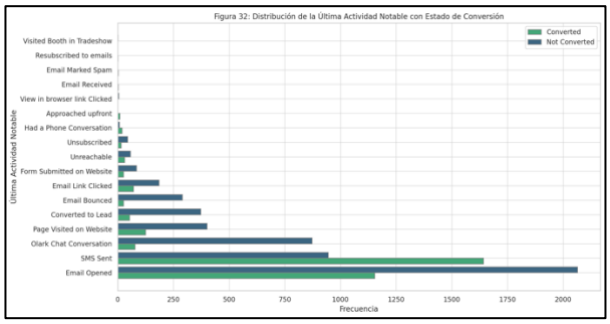


Fuente: Elaboración propia

Actividades recientes como "Email Opened" y "Page Visited on Website" son indicadores importantes de conversión. Los leads con estas actividades tienen una mayor probabilidad de convertirse. Estas interacciones recientes reflejan un alto nivel de compromiso y, por lo tanto, una mayor probabilidad de conversión (Figura 32). Los leads que se originan a través de "API"

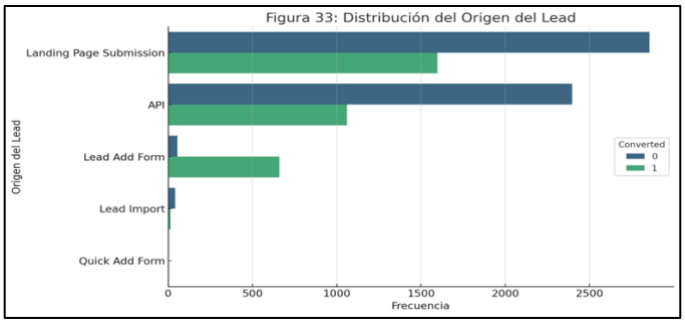
y "Landing Page Submission" tienen tasas de conversión más altas. Estas fuentes de lead generan leads de alta calidad que tienen más probabilidades de convertirse (Figura 33).

Figura 32: Distribución de la Última Actividad Notable



Fuente: Elaboración propia

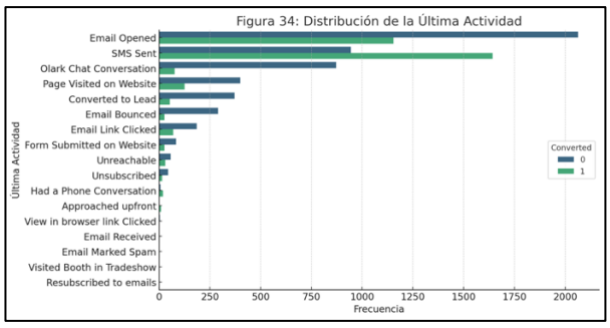
Figura 33: Distribución del Origen del Lead



Fuente: Elaboración propia

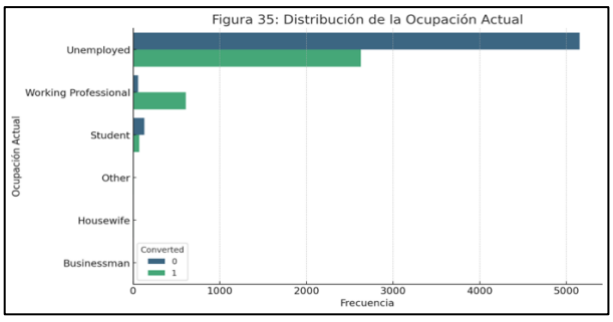
La última actividad realizada por el cliente, como "Email Opened" y "SMS Sent", es un fuerte predictor de conversión. Los leads con estas actividades muestran un mayor compromiso y una mayor probabilidad de conversión (Figura 34). La ocupación actual del cliente potencial influye en las tasas de conversión. Los "Working Professionals" y "Students" tienen mayores tasas de conversión en comparación con otros grupos (Figura 35). Este hallazgo subraya la importancia de adaptar las estrategias de marketing según el perfil ocupacional de los leads.

Figura 34: Distribución de Actividad



Fuente: Elaboración propia

Figura 35: Distribución de la Ocupación

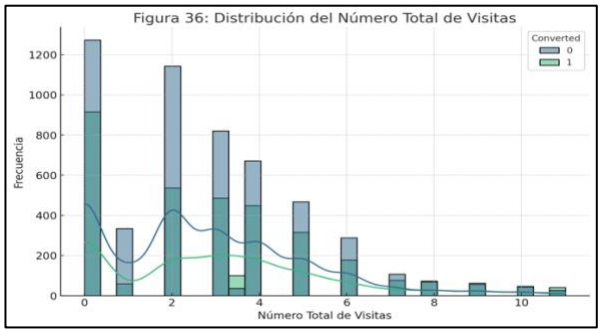


Fuente: Elaboración propia

El número total de visitas al sitio web también es un indicador clave de conversión. Los clientes potenciales con más visitas tienen una mayor probabilidad de conversión (Figura 36). Este patrón sugiere que una mayor exposición al contenido del sitio web está correlacionada con un mayor interés y compromiso.

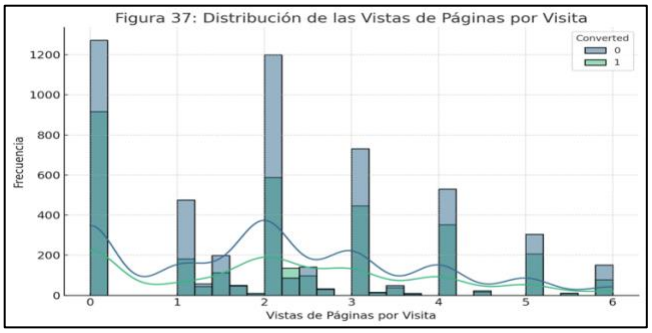
El número promedio de páginas vistas por visita indica el nivel de interés del cliente. Un mayor número de páginas vistas por visita se correlaciona con una mayor probabilidad de conversión (Figura 37). Esto refuerza la importancia de proporcionar contenido atractivo y relevante para retener la atención de los visitantes.

Figura 36: Distribución del Número Total de Visitas



Fuente: Elaboración propia

Figura 37: Distribución de las Vistas de Páginas por Visita



Fuente: Elaboración propia

Las variables clave que contribuyen a las conversiones en el marketing incluyen el tiempo total en el sitio web, las etiquetas de seguimiento positivo, y las actividades recientes significativas. Estas variables permiten a las empresas de eLearning identificar y priorizar los leads con mayor probabilidad de conversión, optimizando así sus estrategias de marketing para mejorar las tasas de conversión.

Conclusión

En el trabajo realizado se abordó el objetivo general de optimizar las estrategias de marketing en empresas de e-learning mediante el uso de modelos predictivos. Para ello, en un primer apartado se exploró el ecosistema de datos en el contexto organizacional de las empresas de e-learning, analizando la interacción de diversos elementos como las plataformas de gestión del aprendizaje y los sistemas de gestión de datos estudiantiles. Se concluyó que la integración de herramientas analíticas y de inteligencia empresarial es crucial para optimizar la oferta educativa y mejorar la toma de decisiones estratégicas.

En el segundo apartado, se describió la metodología utilizada para la limpieza y selección de datos, así como la identificación de variables clave para la predicción de conversiones. Los resultados mostraron que las variables "Total Time Spent on Website" y "Tags" son fundamentales para predecir la conversión de leads. Este hallazgo subraya la importancia de un enfoque meticuloso en la selección y procesamiento de datos para garantizar la precisión de los modelos predictivos.

En el tercer apartado, se presentaron y evaluaron varios modelos de aprendizaje automático, incluyendo la regresión logística, árboles de decisión, bosque aleatorio, Gradient Boosting, SVM y KNN. Los modelos de Gradient Boosting y Random Forest destacaron por su rendimiento superior. Gradient Boosting alcanzó una exactitud del 88%, un F1-Score del 85% y un AUC-ROC del 95%, demostrando una capacidad notable para identificar patrones complejos en los datos.

En el cuarto apartado, se visualizó y analizó los resultados obtenidos de los modelos predictivos. Las matrices de confusión y las curvas ROC demostraron la capacidad de los modelos para discriminar entre clases positivas y negativas. Además, se confirmó que variables de marketing como "Total Time Spent on Website" y "Tags" son cruciales para mejorar las estrategias de conversión, destacando su impacto significativo en la precisión de las predicciones.

La adopción de modelos predictivos presenta múltiples beneficios estratégicos para las empresas de e-learning. En primer lugar, permite una segmentación más precisa de la audiencia, identificando a los prospectos con mayor probabilidad de conversión y permitiendo que los

esfuerzos de marketing se dirijan de manera más eficiente. En segundo lugar, la personalización de las campañas de marketing se mejora significativamente, ya que los modelos pueden identificar patrones de comportamiento y preferencias, permitiendo ajustes en las estrategias de comunicación y ofertas. Finalmente, la utilización de análisis predictivo facilita una mejor gestión de los recursos, optimizando los esfuerzos de ventas y marketing y reduciendo los costos asociados con la adquisición de clientes.

Para futuras investigaciones, se recomienda explorar la integración de técnicas de aprendizaje profundo y modelos de redes neuronales que pueden capturar aún más complejidades en los datos y potencialmente ofrecer mejoras adicionales en la precisión de las predicciones. Además, la incorporación de nuevas fuentes de datos, como el análisis de sentimiento de las interacciones de los usuarios en redes sociales y foros, podría proporcionar información valiosa que complemente los datos existentes. También se sugiere investigar el impacto de diferentes estrategias de segmentación y personalización en la conversión de leads utilizando experimentos controlados y pruebas A/B para identificar las tácticas más efectivas. La continua optimización de los modelos mediante técnicas avanzadas de ajuste de hiperparámetros y validación cruzada será crucial para mantener y mejorar el rendimiento predictivo.

En conclusión, la implementación de modelos predictivos ha tenido un impacto significativo y positivo en la optimización de estrategias de marketing para empresas de e-learning. Los logros alcanzados en términos de precisión y eficacia de los modelos de Gradient Boosting y Random Forest destacan la importancia de utilizar herramientas analíticas avanzadas para mejorar la conversión de prospectos en clientes. Estos modelos mejoran la capacidad de las empresas para tomar decisiones informadas y estratégicas, proporcionando una base sólida para la innovación continua en el campo del marketing educativo. "Los modelos predictivos no solo mejoran la precisión en la predicción de la conversión de leads, sino que también optimizan la personalización de estrategias de marketing, alineándose con las tendencias emergentes y necesidades del mercado educativo" (Namoun & Alshanqiti, 2021; Zhang et al., 2021). La adopción de estos enfoques predictivos asegura que las empresas de e-learning puedan adaptarse rápidamente a las cambiantes demandas del mercado, ofreciendo experiencias educativas personalizadas y optimizando sus estrategias de marketing para maximizar el impacto y la

rentabilidad. La investigación futura y la mejora continua en este ámbito serán esenciales para mantener el liderazgo en el competitivo y dinámico sector del e-learning.

Referencias bibliográficas

Adz Zikri, A. F. N., & Widianto, S. (2023). Applying machine learning to predict online customers behaviour. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4430029

Ahmed, M., & Elaraby, I. (2014). Data mining: A prediction for student's performance using classification method. *World Journal of Computer Application and Technology, 2*(2), 43-47. https://doi.org/10.13189/wjcat.2014.020203

Alsghaier, H., Ming-Syan, C., & Karim, A. (2017). El éxito de las empresas de eLearning radica en su capacidad para adaptarse a las cambiantes demandas educativas y tecnológicas de una audiencia diversa.

Arango, J. D. F. (2021). Predicción de abandono de clientes en telecomunicaciones mediante el aprendizaje automático. *Edu.co*. Recuperado el 25 de noviembre de 2023, de https://expeditiorepositorio.utadeo.edu.co/bitstream/handle/20.500.12010/22247/Tesis%20Churn-Jesus%20David%20Falla.pdf?sequence=1&isAllowed=y

Belén, F., & Donoso, T. (2023). Generación de un modelo de clasificación de Lead Scoring para una empresa SaaS. Memoria para optar al título de Ingeniera Civil Industrial. Universidad de Chile Facultad de Ciencias Físicas y Matemáticas Departamento de Ingeniería Industrial. Recuperado de https://repositorio.uchile.cl/bitstream/handle/2250/194914/Generacion-de-un-modelo-de-clasificacion-de-Lead-Scoring-para-una-empresa-SAAS.pdf?sequence=1&isAllowed=y

Curso de Modelos Analíticos de Marketing, UC Online. (s.f.). *UC Online*. Recuperado de https://www.uc.cl/cursos/modelos-analiticos-de-marketing

Imran, M., Latif, S., Mehmood, D., & Shah, M. S. (2019). Student academic performance prediction using supervised learning techniques. *International Journal of Emerging Technologies in Learning (IJET), 14*(14), 92-104. https://doi.org/10.3991/ijet.v14i14.10310

Jayalekshmi, K. R. (s/f). A comparative analysis of predictive models using machine learning algorithms for customer attrition in the mobile telecom sector. *Tojdel.net*. Recuperado el 25 de noviembre de 2023, de https://tojdel.net/journals/tojdel/articles/v11i01c02/v11i01-01.pdf

Lee, J., Jung, O., Lee, Y., Kim, O., & Park, C. (2021). A comparison and interpretation of machine learning algorithm for the prediction of online purchase conversion. *Journal of Theoretical and Applied Electronic Commerce Research, 16*(5), 1472–1491. https://doi.org/10.3390/jtaer16050083

Lindberg, A.-M. (2018, noviembre 25). Use of predictive analytics in B2B sales lead generation. *Theseus.fi*. Recuperado de https://www.theseus.fi/bitstream/handle/10024/155313/Thesis-Anna- Maria_Lindberg.pdf?sequence=1

Martins, M. P. G., Miguéis, V. L., Fonseca, D. S. B., & Alves, A. (2019). A data mining approach for predicting academic success: A case study. *International Journal of Educational Technology in Higher Education* (pp. 45–56). Cham: Springer.

Mathur, N., Kumar, S., Joshi, T., & Dhuliya, P. (2022). Analyzing Consumer Behavior Predictions: A Review of Machine Learning Techniques. *2022 International Conference on

Advances in Computing, Communication and Materials (ICACCM)*, Dehradun, India, 2022, pp. 1-5. https://doi.org/10.1109/ICACCM56405.2022.10009209

McCarthy, R. V., McCarthy, M. M., Ceccucci, W., & Halawi, L. (2019). Introduction to predictive analytics. In *Applying Predictive Analytics*. Springer International Publishing. https://doi.org/10.1007/978-3-030-14038-0

Mueen, A., Zafar, B., & Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science, 8*(11), 36-42. https://doi.org/10.5815/ijmecs.2016.11.05

Namoun, A., & Alshanqiti, A. (2021). Predicting student performance using data mining and learning analytics techniques: A systematic literature review. *Applied Sciences, 11*(1), 237. https://doi.org/10.3390/app11010237

Navarrete, I., & Sebastian, J. (2021). Desarrollo de una estrategia de marketing digital utilizando análisis predictivo para la comercialización de productos de la Empresa Digimax. Riobamba Universidad Nacional de Chimborazo.

QuestionPro. (s.f.). Modelos de aprendizaje automático en marketing. *QuestionPro*. Recuperado de https://www.questionpro.com/blog/es/modelos-de-aprendizaje-automatico-en-marketing/

Redalyc. (2015). Revisión de modelos predictivos en la educación superior. *Red de Revistas Científicas de América Latina y el Caribe, España y Portugal*. Recuperado de https://www.redalyc.org

Zhang, Y., Yun, Y., An, R., Cui, J., Dai, H., & Shang, X. (2021). Educational data mining techniques for student performance prediction: Method review and comparison analysis. *Frontiers in Psychology, 12*, 698490. https://doi.org/10.3389/fpsyg.2021.698490

Zulaikha, S., Mohamed, H., Kurniawati, M., Rusgianto, S., & Rusmita, S. A. (2020). Customer predictive analytics using artificial intelligence. *The Singapore Economic Review, 1–12*. https://doi.org/10.1142/s0217590820480021

Apéndices

Apéndice I - Variables en el dataset

- Prospect ID: Identificación única del cliente.

- Lead Number: Número asignado a cada lead.

- Lead Origin: Origen del lead, como API o envío de página de destino.

- Lead Source: Fuente específica del lead, como Google o búsqueda orgánica.

- Do Not Email: Preferencia del cliente sobre recibir correos electrónicos.

- Do Not Call: Preferencia del cliente sobre recibir llamadas.

- Converted: Variable objetivo que indica si un lead se convirtió exitosamente.

- Total Visits: Número total de visitas realizadas por el cliente en el sitio web.

- Total Time Spent on Website: Tiempo total que el cliente pasó en el sitio web.

- Page Views Per Visit: Número promedio de páginas vistas por visita.

- Last Activity: Última actividad realizada por el cliente, como correos electrónicos abiertos.

- Country: País del cliente, útil para segmentar el mercado.

- Specialization: Dominio de la industria en el que trabajó el cliente antes.

- How did you hear about X Education: Fuente de la que el cliente supo sobre X Education.

- What is your current occupation: Ocupación actual del cliente.

- What matters most to you in choosing this course: Principal motivo del cliente para realizar el curso.

- Search: Indica si el cliente vio el anuncio en alguno de los ítems listados.

- Magazine: Indica si el cliente se enteró del curso a través de una revista.

- Newspaper Article: Indica si el cliente se enteró del curso a través de un artículo de periódico.

- X Education Forums: Indica si el cliente se enteró del curso a través de los foros de X Education.

- Newspaper: Indica si el cliente se enteró del curso a través de un periódico.

- Digital Advertisement: Indica si el cliente se enteró del curso a través de un anuncio digital.

- Through Recommendations: Indica si el cliente llegó a través de recomendaciones.

- Receive More Updates About Our Courses: Preferencia del cliente sobre recibir más actualizaciones.

- Tags: Etiquetas asignadas a los clientes que indican el estado actual del lead.

- Lead Quality: Calidad del lead basada en datos y la intuición del empleado.

- Update me on Supply Chain Content: Preferencia del cliente sobre recibir actualizaciones de la cadena de suministro.

- Get updates on DM Content: Preferencia del cliente sobre recibir actualizaciones de marketing directo.

- Lead Profile: Nivel de lead asignado a cada cliente.

- City: Ciudad del cliente.

- Asymmetrique Activity Index: Índice de actividad asignado a cada cliente.

- Asymmetrique Profile Index: Índice de perfil asignado al cliente.

- Asymmetrique Activity Score: Puntuación basada en la actividad del cliente.

- Asymmetrique Profile Score: Puntuación basada en el perfil del cliente.

- I agree to pay the amount through cheque: Indica si el cliente ha acordado pagar mediante cheque.

- A free copy of Mastering The Interview: Preferencia del cliente sobre recibir una copia gratuita de "Mastering the Interview".

- Last Notable Activity: Última actividad notable realizada por el cliente.

Apéndice II – Codigo para la limpieza de la base de datos

A continuación, se presenta el código utilizado para la limpieza y preparación de los datos. Este proceso es crucial para garantizar la calidad y la integridad de los datos antes de aplicar modelos predictivos. El código se enfoca en la eliminación de valores faltantes, imputación de datos, eliminación de outliers y preparación de las características.

```
# Mostrar los primeros registros para verificar

print(df.head())


# Información general del DataFrame

print("\nInformación del DataFrame:")

print(df.info())


print("\nEstadísticas descriptivas:")

print(df.describe())


# Identificación de valores faltantes

missing_values = df.isnull().sum()

missing_percentage = (missing_values / len(df)) * 100

missing_data = pd.DataFrame({'Valores Faltantes': missing_values, 'Porcentaje': missing_percentage})

missing_data = missing_data[missing_data['Valores Faltantes'] > 0]

print("\nValores faltantes en el DataFrame:")

print(missing_data)
```

```
# Eliminación de columnas con más del 45% de valores faltantes
cols_to_drop = missing_data[missing_data['Porcentaje'] > 45].index
df.drop(columns=[col for col in cols_to_drop if col in df.columns], inplace=True)

# Imputación de valores faltantes para columnas categóricas con la moda
categorical_cols = ['Tags', 'Lead Profile', 'What matters most to you in choosing a course', 'What is your current occupation']
for col in categorical_cols:
    if col in df.columns:
        df[col].fillna(df[col].mode()[0], inplace=True)

# Imputación de valores faltantes para columnas numéricas con la media
numerical_cols = ['TotalVisits', 'Page Views Per Visit']
for col in numerical_cols:
    if col in df.columns:
        df[col].fillna(df[col].mean(), inplace=True)

# Verificación de la imputación
print("\nValores faltantes después de la imputación:")
print(df.isnull().sum())

# Identificación de outliers
```

```
def find_outliers(df, col):
    Q1 = df[col].quantile(0.25)
    Q3 = df[col].quantile(0.75)
    IQR = Q3 - Q1
    lower_bound = Q1 - 1.5 * IQR
    upper_bound = Q3 + 1.5 * IQR
    return lower_bound, upper_bound


outlier_limits = {}
for col in numerical_cols:
    if col in df.columns:
        lower, upper = find_outliers(df, col)
        outlier_limits[col] = (lower, upper)
        print(f"\nLímites de outliers para {col}:")
        print(f"Límite inferior: {lower}, Límite superior: {upper}")


# Identificación y eliminación de outliers
for col in numerical_cols:
    if col in df.columns:
        lower, upper = outlier_limits[col]
        df = df[(df[col] >= lower) & (df[col] <= upper)]


print("\nDespués de la eliminación de outliers:")
```

```
print(df.info())


# Visualización de outliers con boxplots
plt.figure(figsize=(15, 10))
if 'TotalVisits' in df.columns:
    plt.subplot(3, 1, 1)
    sns.boxplot(x=df['TotalVisits'])
    plt.title('Boxplot de TotalVisits')

if 'Total Time Spent on Website' in df.columns:
    plt.subplot(3, 1, 2)
    sns.boxplot(x=df['Total Time Spent on Website'])
    plt.title('Boxplot de Total Time Spent on Website')

if 'Page Views Per Visit' in df.columns:
    plt.subplot(3, 1, 3)
    sns.boxplot(x=df['Page Views Per Visit'])
    plt.title('Boxplot de Page Views Per Visit')

plt.tight_layout()
plt.show()

# Verificar si hay duplicados en el DataFrame
```

```
duplicate_count = df.duplicated().sum()

print(f'Número de filas duplicadas: {duplicate_count}')


# Calcular la frecuencia de las variables categóricas

categorical_features = df.select_dtypes(include=['object']).columns

categorical_frequencies = {}

for col in categorical_features:

    categorical_frequencies[col] = df[col].value_counts()


# Mostrar las frecuencias

for col, freq in categorical_frequencies.items():

    print(f'Frecuencia de {col}:\n{freq}\n')


# Reemplazar "Select" por NaN en las variables categóricas

df.replace('Select', np.nan, inplace=True)


# Calcular el porcentaje de valores nulos por columna

null_percentage = df.isnull().sum() / len(df) * 100


# Identificar columnas con más del 25% de valores nulos

columns_to_drop = null_percentage[null_percentage > 25].index


# Eliminar estas columnas
```

```
df.drop(columns=[col for col in columns_to_drop if col in df.columns], inplace=True)


# Identificar columnas categóricas con solo una categoría

single_value_columns = [col for col in categorical_features if col in df.columns and df[col].nunique() <= 1]


# Eliminar las columnas indicadas

columns_to_drop = ['What matters most to you in choosing a course', 'Search']

df.drop(columns=[col for col in columns_to_drop if col in df.columns], inplace=True)


# Eliminar estas columnas

df.drop(columns=single_value_columns, inplace=True)


# Convertir 'Prospect ID' a índice

if 'Prospect ID' in df.columns:
```

Apéndice III – Optimización de parámetros de los modelos

Para llevar a cabo la optimización de parámetros en los modelos de Gradient Boosting y Random Forest, se utilizó la técnica de búsqueda en cuadrícula (Grid Search). A continuación, se presenta una tabla con los parámetros optimizados y una explicación del proceso:

Modelo	**Parámetro**	**Valor Optimo**
Gradient Boosting	Número de estimadores	100
	Tasa de aprendizaje	0.09

Modelo	Parámetro	Valor Optimo
	Profundidad máxima	7
	Subsample	0.9
	Estado aleatorio	42
Random Forest	Número de estimadores	300
	Profundidad máxima	20
	Mínimas muestras por división	5
	Mínimas muestras por hoja	2
	Máximo de características	Raíz cuadrada
	Estado aleatorio	42

Apendice IIII - Código para la Implementación y Evaluación de Modelos

```
#Definición de características y objetivo
features = df.drop(columns=['Converted'])
target = df['Converted']

# División de los datos en conjuntos de entrenamiento y prueba
X_train, X_test, y_train, y_test = train_test_split(features, target, test_size=0.1, random_state=42)

# Modelo de Gradient Boosting con los mejores hiperparámetros
best_gb = GradientBoostingClassifier(
    n_estimators=100,
```

```
    learning_rate=0.09,
    max_depth=7,
    subsample=0.9,
    random_state=42
)
best_gb.fit(X_train, y_train)
y_pred_gb = best_gb.predict(X_test)
y_prob_gb = best_gb.predict_proba(X_test)[:, 1]

# Métricas para Gradient Boosting
metrics_gb = {
    'Accuracy': accuracy_score(y_test, y_pred_gb),
    'Precision': precision_score(y_test, y_pred_gb),
    'Recall': recall_score(y_test, y_pred_gb),
    'F1-Score': f1_score(y_test, y_pred_gb),
    'AUC-ROC': roc_auc_score(y_test, y_prob_gb)
}

# Modelo de Random Forest con los mejores hiperparámetros
best_rf = RandomForestClassifier(
    n_estimators=300,
    max_depth=20,
    min_samples_split=5,
```

```
    min_samples_leaf=2,
    max_features='sqrt',
    random_state=42
)
best_rf.fit(X_train, y_train)
y_pred_rf = best_rf.predict(X_test)
y_prob_rf = best_rf.predict_proba(X_test)[:, 1]

# Métricas para Random Forest
metrics_rf = {
    'Accuracy': accuracy_score(y_test, y_pred_rf),
    'Precision': precision_score(y_test, y_pred_rf),
    'Recall': recall_score(y_test, y_pred_rf),
    'F1-Score': f1_score(y_test, y_pred_rf),
    'AUC-ROC': roc_auc_score(y_test, y_prob_rf)
}

# Gráficas de matrices de confusión
conf_matrix_rf = confusion_matrix(y_test, y_pred_rf)
conf_matrix_gb = confusion_matrix(y_test, y_pred_gb)

fig, axes = plt.subplots(1, 2, figsize=(14, 6))
```

```
sns.heatmap(conf_matrix_rf, annot=True, fmt='d', cmap='Blues', ax=axes[0])
axes[0].set_title('Random Forest Confusion Matrix')
axes[0].set_xlabel('Predicted')
axes[0].set_ylabel('Actual')

sns.heatmap(conf_matrix_gb, annot=True, fmt='d', cmap='Blues', ax=axes[1])
axes[1].set_title('Gradient Boosting Confusion Matrix')
axes[1].set_xlabel('Predicted')
axes[1].set_ylabel('Actual')

plt.show()

# Gráficas de curvas ROC
fpr_rf, tpr_rf, _ = roc_curve(y_test, y_prob_rf)
fpr_gb, tpr_gb, _ = roc_curve(y_test, y_prob_gb)

plt.figure(figsize=(10, 6))
plt.plot(fpr_rf, tpr_rf, label=f'Random Forest (AUC = {metrics_rf["AUC-ROC"]:.2f})')
plt.plot(fpr_gb, tpr_gb, label=f'Gradient Boosting (AUC = {metrics_gb["AUC-ROC"]:.2f})')
plt.plot([0, 1], [0, 1], 'k--')
plt.xlabel('False Positive Rate')
plt.ylabel('True Positive Rate')
plt.title('ROC Curve')
```

```
plt.legend()

plt.show()


# Retorno de métricas para ambos modelos

metrics_gb, metrics_rf
```

Anexo – Reporte del Tutor

El trabajo de Sofía titulado *Optimización Estratégica de Marketing en Empresas de eLearning. Procesamiento de Datos en Python y Modelos Predictivos para la Conversión de Clientes*, es excelente en cuanto refiere a un trabajo final de especialización que consiste en un trabajo integrador. En este sentido, el trabajo presentado por Sofía, integra diferentes aspectos revisados en la especialización. De esto se da cuenta en prolija la organización del trabajo en tres apartados que van avanzando progresivamente en la consecución del objetivo general declarado: *analizar cómo el procesamiento de datos en Python y los modelos predictivos pueden mejorar las estrategias de marketing en empresas de e-learning*. Para ello, en un primer apartado expone una problemática organizacional vinculada con el tratamiento de los datos en empresas de e-learning. Seguidamente, en el segundo apartado, realiza un procedimiento operativo en el que expone un método cuantitativo para el procesamiento de los datos y cómo se procedió para la selección de esos datos. El tercer apartado presenta el procedimiento para establecer un criterio que permita realizar la selección y la evaluación de modelos predictivos. En el cuarto apartado, realiza un ejercicio de visualización de resultados. El trabajo finaliza con conclusiones en las que se dejan planteadas unas líneas de indagación a futuro.

El proceder propuesto en el trabajo permite a Sofía ir desarrollando un argumento que, al mismo tiempo que desarrolla el objetivo general, deja en evidencia la articulación del contenido de los módulos propuestos en la especialización. También, sobre todo en el cuarto apartado, se presenta cierto contenido técnico que surge del interés de Sofía. Adicionalmente, las líneas de investigación futuras propuestas en su trabajo pueden ser retomadas por elaboraciones futuras, tanto en trabajos de especialización como en trabajos de maestría. Es importante remarcar que estas líneas de indagación no se limitan únicamente a la Carreras de Posgrado de Métodos Cuantitativos para la Gestión y el Análisis de Datos en Organizaciones, si no que pueden ser abordadas por trabajos más generales que presenten interrogantes vinculados con el tratamiento de los datos. En esta línea, el trabajo de Sofía realiza una contribución directa al proyecto UBACyT que dirijo titulado *Economía de plataformas y común digital. Visualización de las lógicas en disputa en el territorio virtual digital*. Como mentor del trabajo final de la especialización y como director del proyecto mencionado, quisiera finalizar este reporte indicando la facilidad que implicó realizar este trabajo en compañía de Sofía: los intercambios

fueron fluidos y la organización de los tiempos por parte de ella excelente. Espero también haber contribuido con la orientación de su trabajo.