UIDE
Powered by
Arizona State University

Maestría en

SISTEMAS DE INFORMACIÓN

Mención **Inteligencia de Negocios y Analítica de Datos Masivos.**

Tesis previa a la obtención del título de Magíster en Sistemas de Información mención Inteligencia de Negocios y Analítica de Datos Masivos.

AUTORES:

Jaime David Maldonado Núñez

Jimmy David Rivera Cárdenas

Renato Sebastián Jácome Granda

Esteban Sebastián Caza Jácome

TUTOR:

Mgtr. Marco Rengifo

Análisis y desarrollo de un modelo de simulación para evaluar estrategias de marketing de fidelización, enfocado en estrategias de segmentación y posicionamiento en una Institución de Educación Superior del Ecuador.

APROBACIÓN DEL TUTOR

Yo, Marco Rengifo Pozo, certifico que conozco los autores del presente trabajo siendo la responsables exclusivos tanto de su originalidad y autenticidad, como de su contenido.

Marco Rengifo Pozo
DIRECTOR DE TESIS

Certificación de autoría

Nosotros, **Jaime David Maldonado Núñez, Jimmy David Rivera Cárdenas, Renato Sebastián Jácome Granda, Esteban Sebastián Caza Jácome,** declaramos bajo juramento que el trabajo aquí descrito es de nuestra autoría; que no ha sido presentado anteriormente para ningún grado o calificación profesional y que se ha consultado la bibliografía detallada.

Cedemos nuestros derechos de propiedad intelectual a la Universidad Internacional del Ecuador (UIDE), para que sea publicado y divulgado en internet, según lo establecido en la Ley de Propiedad Intelectual, su reglamento y demás disposiciones legales.

Firma del graduando
Jaime David Maldonado Núñez

Firma del graduando
Jimmy David Rivera Cárdenas

Firma del graduando

Renato Sebastián Jácome Granda

Firma del graduando

Esteban Sebastián Caza Jácome

Autorización de Derechos de Propiedad Intelectual

Nosotros, **Jaime David Maldonado Núñez, Jimmy David Rivera Cárdenas, Renato Sebastián Jácome Granda, Esteban Sebastián Caza Jácome**, en calidad de autores del trabajo de investigación titulado **Análisis y desarrollo de un modelo de simulación para evaluar estrategias de marketing de fidelización, enfocado en estrategias de segmentación y posicionamiento en una Institución de Educación Superior del Ecuador**, autorizamos a la Universidad Internacional del Ecuador (UIDE) para hacer uso de todos los contenidos que nos pertenecen o de parte de los que contiene esta obra, con fines estrictamente académicos o de investigación. Los derechos que como autores nos corresponden, lo establecido en los artículos 5, 6, 8, 19 y demás pertinentes de la Ley de Propiedad Intelectual y su Reglamento en Ecuador

D. M. Quito, (septiembre 2024)

Firma del graduando
Jaime David Maldonado Núñez

Firma del graduando
Jimmy David Rivera Cárdenas

Firma del graduando
Renato Sebastián Jácome Granda

Firma del graduando
Esteban Sebastián Caza Jácome

Dedicatoria

En primer lugar, quiero dedicar este trabajo a Dios por brindarme la fuerza en los momentos más difíciles, por la salud y por la vida, ya que sin ello nada sería posible. A mis padres, Renato y Patricia, por siempre brindarme su apoyo y siempre estar a mi lado y nunca dejarme solo. A mis hermanos, Mateo y Karla, que siempre estuvieron ahí para apoyarme y alentarme a seguir adelante. A Alejandra Erazo, mi compañera de aventuras, que con su paciencia y apoyo incondicional me ayuda a seguir adelante y me inspira a ser mejor persona. Por último, a mis abuelitos por estar a mi lado y siempre guiarme por el camino correcto.

Esteban Sebastián Caza Jácome

Con sincera gratitud y amor, dedico este logro a Dios, cuya inquebrantable guía ha iluminado mi camino, a mis queridos padres, cuyo incesante apoyo y amor han sido la piedra angular de mi trayectoria, y a mi hermana, cuyo constante aliento y compañía han sido fuente de alegría y fortaleza. A cada uno de mis amigos, mi más profundo agradecimiento por su ayuda y por ser mis firmes compañeros en cada momento crucial de mi vida.

Jimmy David Rivera Cárdenas

Dedico este fruto de mi trabajo a mi familia quienes son mi inspiración y apoyo para alcanzar mis metas. A mis padres, por su amor incondicional y por transmitirme la importancia de la perseverancia y el esfuerzo. A mi hermana, por motivarme constantemente a superarme y a perseguir mis sueños.

Renato Sebastián Jácome Granda

Dedico este trabajo a Dios por la salud y las oportunidades que me ha brindado en este camino, a mis padres y hermano José, quienes con su amor, sacrificio y apoyo incondicional han sido el pilar fundamental en mi vida. A mi enamorada Michelle, cuyo apoyo y aliento me han motivado en cada paso de este proceso. Este logro es un reflejo de la dedicación y el amor que todos ustedes me han brindado.

Jaime David Maldonado Núñez

Agradecimientos

Quiero agradecer en primer lugar a Dios por darme un día más de vida, ya que con salud se puede lograr cualquier cosa. A mis padres y hermanos, por siempre estar a mi lado y brindarme su apoyo, también quiero agradecerles por siempre tratar de darnos lo mejor a mí y a mis hermanos. A Alejandra Erazo, mi compañera de aventuras, por su apoyo infinito y por sus palabras de aliento. A mis profesores, por siempre por ser una guía y brindarme sus conocimientos para ser un gran profesional y finalmente quiero agradecer a mis amigos, que de igual manera estuvieron a mi lado brindando su apoyo cuando yo lo requería

Esteban Sebastián Caza Jácome

"Agradezco a mis padres por ser mi mayor motivación, por inculcarme sus valiosos principios y por su apoyo incondicional. A mi hermana, que siempre me alienta a seguir adelante y a esforzarme cada día por ser mejor. A mi familia, que, a pesar de la distancia, siempre está pendiente de mí y de mis logros."

Renato Sebastián Jácome Granda

Agradezco a mis padres, hermano, mi abuelita y tíos por su amor y confianza en mí y el respaldo constante que me han brindado durante todo este proceso. Han sido un motor que me impulsó a seguir adelante incluso en circunstancias difíciles.

A mis compañeros, gracias por haber compartido este trayecto conmigo. Juntos hemos enfrentado retos, aprendido valiosas lecciones y forjado amistades que perdurarán más allá de este logro.

Quiero hacer un agradecimiento al Msc. Marco Rengifo, por su apoyo y orientación a lo largo de este trabajo.

Jaime David Maldonado Núñez

Resumen

El objetivo del proyecto es desarrollar un modelo predictivo para identificar posibles estudiantes interesados en ingresar a una institución de educación superior a partir de campañas de marketing realizadas previamente. Para garantizar la privacidad de los datos, se trabajó con datos sintéticos generados a partir de patrones históricos reales y la anonimización de datos personales, lo que permitió simular escenarios sin comprometer la confidencialidad de la información.

El enfoque metodológico se basó en recopilar diversas fuentes de datos, incluyendo interacciones en redes sociales, datos demográficos, visitas, y llamadas telefónicas, que se encontraban en un CRM. A partir de estos datos, se construyó un modelo de predicción que permite segmentar y clasificar a los posibles estudiantes según su probabilidad de inscripción.

La implementación del modelo se realizó utilizando herramientas de machine learning que permitieron evaluar diferentes algoritmos para encontrar el más adecuado en términos de precisión y rendimiento. Entre las técnicas evaluadas se incluyeron KNN, árboles de decisión y redes neuronales, siendo Random Forest el modelo seleccionado por su robustez y facilidad de interpretación en este contexto.

Palabras Claves: Modelo predictivo, Identificación de posibles estudiantes, Campañas de marketing, Datos sintéticos, Machine learning, Privacidad de datos

Abstract

The objective of the project is to develop a predictive model to identify possible students interested in entering a higher education institution based on previously carried out marketing campaigns. To guarantee data privacy, we worked with synthetic data generated from real historical patterns and the anonymization of personal data, which allowed scenarios to be simulated without compromising the confidentiality of the information.

The methodological approach was based on collecting various data sources, including social media interactions, demographic data, visits, and phone calls, which were in a CRM. From these data, a prediction model was built that allows potential students to be segmented and classified according to their probability of enrollment.

The implementation of the model was carried out using machine learning tools that allowed different algorithms to be evaluated to find the most suitable one in terms of precision and performance. The techniques evaluated included KNN, decision trees and neural networks, with Random Forest being the model selected for its robustness and ease of interpretation in this context.

Keywords: Predictive model, Identification of potential students, Marketing campaigns, Synthetic data, Machine learning, Data privacy

Contenido

Índice de figuras

Abreviaturas

ADS	Es el nombre del administrador de anuncios de pagos en instagram.
CRM	Customer Relationship Management
PCA	Análisis de Componentes Principales
MAE	Error Absoluto Medio
MSE	Error Cuadrático Medio
RMSE	Raíz Cuadrada del Error Cuadrático Medio
ETL	Extract, Transform, Load
VMs	Máquinas Virtuales
SQL Server	Es un sistema de gestión de bases de datos relacionales desarrollado por Microsoft.
INEC	Instituto Nacional de Estadística y Censos
COVID-19	Es una enfermedad infecciosa causada por un nuevo coronavirus.
RGPD	Reglamento General de Protección de Datos
LOPDGDD	Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales
DPD	Delegado de Protección de Datos
AEPD	Agencia Española de Protección de
político, económico, sociocultural,	pueden influir en una organización: tecnológico, ecológico y legal.
PIB	Producto Interior Bruto
SaaS	Software as a Service

AWS	Amazon Web Services

CAPÍTULO I

Introducción

Las redes sociales han revolucionado cómo las organizaciones, incluidas las instituciones educativas, realizan campañas de marketing. Según estadísticas recientes, más del 70% de las decisiones de compra de los consumidores son influenciadas por las redes sociales. Este hecho subraya la necesidad de optimizar las estrategias de marketing digital para maximizar su efectividad. (KEMP, 2023)

¿Qué es Marketing?

El marketing es un proceso completo que engloba diversas estrategias y acciones destinadas a convertir a un posible comprador en un cliente satisfecho y fiel. Desde la investigación detallada del mercado hasta la creación de campañas publicitarias efectivas, el marketing busca demostrar el valor y la relevancia de un producto o servicio, fomentando la lealtad hacia la marca y aumentando las ventas totales. (Snyder, 2024)

Sin embargo, lograr este objetivo no es tarea sencilla. Los profesionales del marketing deben invertir tiempo en comprender a fondo a su público objetivo. Este entendimiento profundo les permite diseñar estrategias que realmente conecten con los consumidores, permitiéndoles destacar en un mercado saturado de mensajes y tácticas promocionales.

Además, con el auge de las plataformas digitales y las redes sociales, es esencial que las estrategias de marketing se adapten continuamente. El uso de análisis de datos y tecnologías emergentes, como la inteligencia artificial y la automatización, puede ofrecer

insights valiosos para optimizar campañas y llegar al público correcto en el momento adecuado.

En esencia, el éxito en marketing requiere una combinación de conocimiento profundo del consumidor, innovación estratégica y adaptabilidad en un entorno en constante evolución

¿Marketing en Universidades?

En el competitivo panorama educativo actual, las universidades deben implementar estrategias de marketing efectivas para atraer a estudiantes potenciales y destacar entre la competencia. Desarrollar un plan de marketing integral implica comprender profundamente al público objetivo, establecer una identidad de marca sólida y aprovechar las plataformas digitales para interactuar con futuros estudiantes.

Conocer las características y preferencias de los posibles alumnos es esencial. Identificar factores como el rango de edad, la ubicación geográfica, el estado socioeconómico y los intereses académicos permite adaptar los mensajes para que resuenen de manera más efectiva. Además, analizar el comportamiento en línea y las preferencias de comunicación ayuda a diseñar estrategias que conecten verdaderamente con la audiencia. (Chaves, 2024)

Construir una marca atractiva es clave para posicionar a la institución como la mejor opción. Esto incluye definir una misión y valores claros, crear elementos visuales memorables como logos y eslóganes, y destacar lo que diferencia a la universidad, ya sea la excelencia académica, programas innovadores u oportunidades únicas.

Modernizar los materiales de marketing con imágenes impactantes y contenido interactivo puede aumentar significativamente el compromiso. Utilizar fotografías de alta

calidad, infografías, recorridos virtuales por el campus y aprovechar las redes sociales permite ampliar el alcance y la interacción con una audiencia más amplia. Fomentar la generación de contenido por parte de los estudiantes y exalumnos también puede amplificar la visibilidad de la institución.

El análisis de datos y la toma de decisiones basadas en datos son vitales para la mejora continua. Emplear herramientas analíticas para monitorear el rendimiento en diferentes canales permite identificar qué funciona y dónde se necesitan ajustes. Establecer indicadores clave de rendimiento alineados con los objetivos de marketing garantiza que los esfuerzos sean enfocados y medibles. (Nayak, 2024)

En resumen, el éxito en el marketing universitario requiere una combinación estratégica de comprensión del público, construcción de una marca sólida, adopción de plataformas digitales y un análisis constante de datos para adaptar y mejorar las estrategias en un mercado educativo en constante evolución.

Basándonos en los datos obtenidos de diversos medios de marketing—tanto digitales como no digitales—y de las personas que se matriculan directamente en el área de admisiones, contamos con una amplia base de información que nos permite predecir si un cliente potencial será un futuro candidato a ser estudiante de la institución de educación superior.

Al desarrollar un modelo de aprendizaje supervisado utilizando esta data, podemos analizar patrones y comportamientos específicos de los leads. Por ejemplo, acciones como el colegio del cual se graduó, si trabaja actualmente, la fuente por la cual tuvo el interés de entrar a la institución, pueden ser indicadores significativos del nivel de interés de un cliente potencial. Cada una de estas interacciones refleja un patrón de importancia en la probabilidad de matrícula.

Actualmente, esta información se encuentra dispersa y el puntaje asignado que proporciona el CRM actual es ambiguo, ya que no siempre refleja un interés real. Al centralizar y armonizar estos datos, podemos calcular un puntaje más preciso basado en métricas específicas de los leads, lo que nos permitirá predecir con mayor exactitud su probabilidad de convertirse en estudiantes.

Este enfoque no solo optimiza los procesos de análisis y evaluación de campañas, sino que también mejora la capacidad de la universidad para atraer estudiantes de manera más efectiva y eficiente. Al tener un puntaje predicho más preciso, podemos priorizar y analizar las diferentes estrategias y fidelización. Esto permite enfocar los esfuerzos y recursos de marketing en diferentes insights para que exista mayor probabilidad de matricularse, incrementando así el impacto de las inversiones en marketing.

Además, esta mejora beneficia al área de admisiones, ya que les permite identificar rápidamente a los clientes potenciales que requieren información oportuna y personalizada. Al priorizar a estos leads, se fomenta una interacción más efectiva, aumentando las posibilidades de concretar una futura matrícula.

Desarrollo del modelo: Este modelo servirá como herramienta para identificar y clasificar a los clientes potenciales en función de su probabilidad de matriculación y éxito académico.

Recogida y uso de datos: Por motivos de confidencialidad y privacidad, la investigación utilizará un conjunto de datos sintéticos que reflejen la estructura y complejidad de los datos reales del sistema de gestión de relaciones con los clientes (CRM) de la institución educativa. Este enfoque garantiza la integridad y aplicabilidad del modelo sin comprometer los datos personales.

Enfoque analítico: El estudio se centrará en el desarrollo y las implicaciones del modelo de puntuación en un entorno simulado. Esto permitirá un análisis detallado y el perfeccionamiento iterativo del modelo, garantizando su solidez y aplicabilidad en entornos reales.

Impacto de negocio

Al utilizar un modelo que nos permita identificar los estudiantes con mayor probabilidad de ingresar a la institución de educación superior, la institución puede agilizar el proceso de selección, enfocándose en candidatos con mayor probabilidad de éxito y compromiso. Al identificar y priorizar a los candidatos más prometedores, se pueden reducir los costos asociados con la gestión de solicitudes y entrevistas de candidatos menos aptos.

Al identificar y aplicar estrategias efectivas de fidelización, se puede crear un modelo de simulación que permita prever el impacto de diversas estrategias, ayudando a la institución a invertir recursos en las tácticas más efectivas. Al enfocarse en estrategias que tienen un mayor retorno sobre la inversión, la institución puede reducir costos asociados con la captación y retención de estudiantes.

Necesidades y objetivos: El modelo de simulación se va a componer de aristas como: interacciones previas, cantidad de clics sobre mailings enviados, interacciones con la página web, el medio de donde se hizo contacto con el interesado. Este matiz evocará como resultado un score que tiene como objetivo garantizar que los stakeholders, no solo prioricen a sus clientes potenciales eficientemente, permitiendo a su vez crezca la tasa de participación, tratamiento y conversión. De esta manera se gestionará una metodología dinámica que no solo mejora la satisfacción de los clientes potenciales, sino que también mejora a la asignación de recursos por parte del área encargada.

Recopilar datos: Para el trabajo de titulación, se empleará datos sintéticos con un enfoque similar al productivo, pero la fuente donde se almacena la data está en una base de datos SQL Server, por lo que, para garantizar la coherencia y estructura del entorno productivo, se replicará la misma estructura no productivo, lo que permitirá analizar estrategias sin comprometer información confidencial. (google, 2024)

Tratamiento de datos: En esta fase nos enfocaremos en uno de los aspectos más críticos, que involucra la transformación de datos sin procesarla, es decir trasladar a un formato para su respectivo análisis mediante los siguientes pasos:

Limpieza, transformación e integración de datos: Corresponden a las tareas esenciales para el manejo de valores atípicos (outliers). Identificar y tratar de manera correcta estos valores garantiza un análisis más confiable y preciso, ya sea mediante eliminación o transformación.

La normalización es otro punto para considerar, esto implica escalar datos numéricos en un rango estándar, este proceso es primordial para garantizar que todas las características aporten por igual al modelo.

Plataforma tecnología: Puesto que la base de datos es SQL Server se utilizará la misma para el trabajo de titulación en entorno local, para la parte de PCA y modelado se usará Google Colab y como herramienta de visualización Power BI.

Modelos y algoritmos: Dado que el resultado que vamos a obtener en nuestro modelo es un valor que no depende de un conjunto de datos etiquetados y se va a predecir un score (valor numérico), vamos a emplear modelos de aprendizaje no supervisado como: clustering, PCA, Autoencoders. Para estimar dicho valor en función de las variables disponibles. De esta manera podemos aprovechar las relaciones intrínsecas entre las variables sin depender de etiquetas.

Evaluación del modelo:

Elegir métricas relevantes que midan el desempeño del modelo en relación con los objetivos definidos. Ejemplos comunes incluyen:

R^2 (Coeficiente de Determinación): Mide la proporción de la varianza en la variable dependiente que es predecible a partir de las variables independientes.

MAE (Error Absoluto Medio): La media de los errores absolutos entre las predicciones y los valores reales.

MSE (Error Cuadrático Medio): La media de los cuadrados de los errores entre las predicciones y los valores reales.

RMSE (Raíz del Error Cuadrático Medio): La raíz cuadrada del MSE, que penaliza más los errores grandes.

Se realizará una evaluación de cada modelo para poder determinar cuál de ellos nos da los resultados óptimos para el proyecto

Presentación de resultados:

Los resultados obtenidos se los presentará en un Dashboard en Power Bi, el cual se comunicará claramente los hallazgos del modelo. Aquí se va a elaborar un listado de los indicadores clave de desempeño, tanto para las fases de desarrollo como de puesta en valor del proyecto. Algunos ejemplos de KPI son:

- Oferta académica más demandada.
- Tipo de género determinante en cada carrera.
- Plataforma de redes sociales a través de la cual los interesados se contactan con la Institución de Educación Superior por carreras.

- Precisión del puntaje en base a la interacción con el CRM y las variables más determinantes.

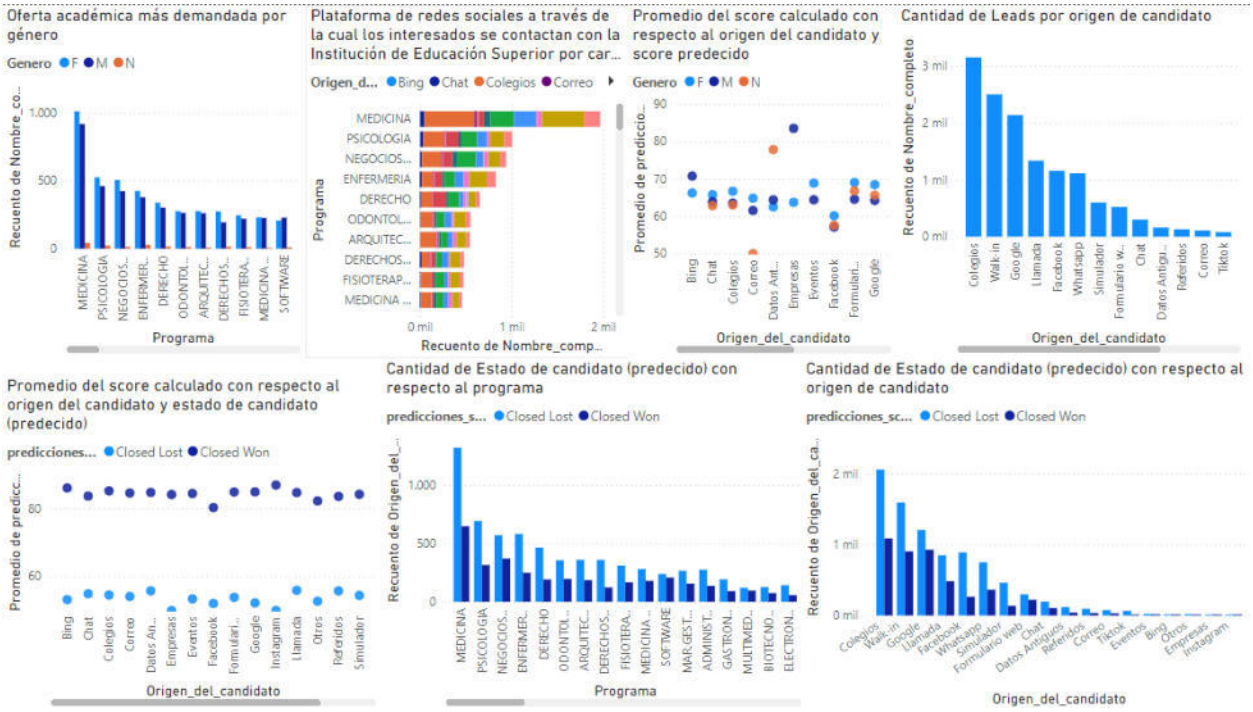

Figura 1 Indicadores a utilizar para la visualización de resultados

Despliegue:

Desarrollo - Integración

Iniciamos preparando el entorno acorde al objetivo del proyecto, asegurándonos de documentar y replicar todo el software y librerías utilizadas en el desarrollo. Evaluamos si la infraestructura actual es adecuada o si se requieren recursos

adicionales. Posteriormente, configuramos un entorno de adecuado para realizar pruebas exhaustivas sin impactar el sistema operativo, asegurando la integración fluida del modelo con los sistemas existentes.

Pruebas

Se realizan pruebas funcionales para validar que el modelo de simulación cumpla con los objetivos de evaluar estrategias de marketing específicas de fidelización, enfocadas en segmentación y posicionamiento. Además, ejecutamos pruebas de carga para evaluar su rendimiento bajo diversas condiciones. Identificamos y resolvemos cualquier conflicto entre librerías y versiones de software, garantizando el funcionamiento adecuado.

Producción

Implementamos planes de contingencia que incluyen la creación de copias de seguridad y estrategias de restauración. Posteriormente, migramos el modelo al entorno de producción, monitoreamos continuamente su rendimiento y estabilidad, realizando ajustes según sea necesario.

Puesta en valor

Estos modelos permiten a la institución evaluar y ajustar eficientemente sus tácticas de marketing mediante el análisis de datos históricos y simulaciones futuras. La toma de decisiones basada en datos reales y precisos reduce la incertidumbre, permitiendo adaptarse ágilmente a las dinámicas del mercado educativo y las preferencias de los estudiantes. Además, la capacidad de segmentar y personalizar las estrategias mejora la satisfacción estudiantil al ofrecer experiencias más relevantes y atractivas,

fortaleciendo así la relación institución-estudiante y promoviendo la fidelización a largo plazo.

Seguimiento

Se establece un sistema de monitoreo continuo para evaluar el rendimiento del modelo y detectar problemas a tiempo. Manteniendo la documentación actualizada y accesible para todos los involucrados en la operación. Este sistema de monitoreo continuo permitirá evaluar el rendimiento del modelo y hacer ajustes en tiempo real. Asegurando que las estrategias de marketing se mantengan efectivas y adaptables a cambios en el comportamiento y preferencias de los estudiantes.

Objetivo General

Desarrollar un modelo de simulación que evalúe la eficacia de las estrategias de marketing de fidelización mediante técnicas de segmentación y posicionamiento en una institución educativa de Ecuador.

Objetivos Específicos

- Crear un modelo de simulación completo que incorpore variables relevantes para los comportamientos y preferencias de los estudiantes. Este modelo permitirá evaluar diferentes estrategias de segmentación y posicionamiento en el contexto del marketing de fidelización.
- Determinar las características demográficas, psicográficas y de comportamiento más influyentes de los estudiantes que repercuten en la fidelización y la retención.
- Evaluar diversas estrategias de marketing para determinar su eficacia a la hora de aumentar la fidelidad de los estudiantes.

Alcance

El proyecto se enfocará en el desarrollo de un modelo de simulación que pueda asignar un puntaje para el cual se realizará un modelo utilizando el lenguaje de programación Python que permita a las instituciones superiores evaluar la efectividad de diferentes estrategias de marketing de fidelización, considerando la segmentación y el posicionamiento de clientes en base a datos de campañas recientes y anteriores, esto va a permitir ponderar a sus mejores clientes potenciales, para facilidad del entendimiento de los resultados, se realizará un dashboard para que pueda visualizar de una manera más clara los resultados y puedan tener mejor toma de decisiones las áreas de admisiones y marketing. El proyecto se centra en la evaluación de las estrategias de marketing de fidelización mediante simulación, pero no incluye la implementación real de estas estrategias en las instituciones superiores.

Análisis PESTEL

Político

En las instituciones públicas, tuvieron un acceso limitado por políticas de admisión y exámenes como el "Ser Bachiller". Sin embargo, con la eliminación de este examen en 2023, se espera un aumento en la matrícula. Históricamente, aproximadamente entre 60-70% de los estudiantes que ingresan a la educación superior lo hacen a través de instituciones públicas. Las universidades privadas tradicionalmente han absorbido a estudiantes que no han podido ingresar a las universidades públicas. Representa aproximadamente entre un 30-40% de la matrícula en educación superior. Con las recientes políticas y cambios, incluyendo la eliminación del "Ser Bachiller", es posible

que este porcentaje cambie, con un potencial aumento en la matrícula privada debido a la mayor demanda. (Estadísticas de ingreso a la educación superior en Ecuador, 2023)

Económico

- Perdida en relación con el poder adquisitivo de los clientes potenciales para las distintas carreras, considerando el puntaje obtenido y su aceptación.
- Los costos asociados con el desarrollo e implementación de modelos de simulación y estrategias de marketing deben considerarse.

Para el año 2024, la economía ecuatoriana enfrentará desafíos significativos, entre ellos la suspensión de las actividades de extracción de crudo, lo que, junto a un limitado crecimiento económico, repercutirá profundamente en la educación superior. Se espera un crecimiento del PIB de apenas el 1.0%, lo que implicará una asignación reducida de fondos gubernamentales para la educación. Esto podría afectar negativamente tanto la calidad como el acceso a la educación superior. La inflación y la disminución en la capacidad económica de las familias también podrían limitar aún más el acceso a este nivel educativo. Se anticipa que hasta el 2027, estos retos seguirán presentes, lo que exigirá que el sector educativo se adapte para mantener su calidad y sostenibilidad ante un entorno de recursos escasos. (Banco Central del Ecuador, 2024).

Social

- Mayor valorización de la educación superior, aumento de la diversidad estudiantil. Cambios demográficos que reduzcan la población en edad escolar, disminución del interés por ciertas carreras.

- Tendencias demográficas, como el número de jóvenes mayores de asistir a la universidad y la graduación de secundaria.

El Instituto Nacional de Estadística y Censos (INEC) ha presentado proyecciones que indican un cambio significativo en la demografía de Ecuador. Se prevé un decrecimiento de la población infantil y adolescente y un aumento en la población de adultos mayores entre 2030 y 2050. El país, que alcanzará los 17.9 millones de habitantes en 2024, verá una disminución en los nacimientos y una creciente esperanza de vida, con mujeres alcanzando los 85.5 años y hombres los 79.6 años para 2050. La población mayor de 60 años se triplicará, mientras que la población joven y en edad escolar disminuirá. Estos cambios reflejan la influencia de factores sociales y culturales en las tendencias demográficas. (INEC, 2024)

Tecnológico

- Las nuevas tecnologías pueden mejorar la captación de estudiantes y la predicción de su éxito.
- Considerar la infraestructura y software para el análisis de datos y la predicción de scores.

La pandemia de COVID-19 aceleró la adopción de la educación virtual en Ecuador, impulsando una rápida transición desde la enseñanza presencial hacia plataformas digitales como Zoom y Google Classroom. Esta transición, aunque crucial para mantener la continuidad educativa, reveló desafíos significativos, incluyendo la desigualdad en el acceso a tecnología y la necesidad de capacitación para docentes y estudiantes (INEC, 2024).

En el periodo post-COVID, Ecuador ha adoptado un modelo híbrido que combina educación presencial y virtual, mejorando la flexibilidad y adaptándose a las nuevas realidades del aprendizaje. Sin embargo, la transición hacia la educación digital también ha resaltado desafíos persistentes, como la brecha en el acceso a tecnología entre diferentes regiones y grupos socioeconómicos. A pesar de los avances en infraestructura digital y metodologías innovadoras, la equidad en el acceso a recursos tecnológicos sigue siendo un obstáculo importante para garantizar una educación inclusiva y de calidad (INEC, 2024).

Ecológico

El factor ecológico no se ha analizado en profundidad por limitaciones en la disponibilidad de datos y a priorizar otros aspectos en las instituciones educativas. Estudios futuros podrían explorar con mayor detalle el impacto ambiental del sector educativo y proponer estrategias para mejorar su sostenibilidad.

Legal

Al presentar un marco legal en desarrollo en cuanto a la protección de datos en Ecuador se tomó en cuenta el modelo de ley de protección de datos de España.

La protección de datos personales en las universidades españolas la regula el Reglamento General de Protección de Datos (RGPD) y la Ley Orgánica de Protección de Datos Personales y garantía de los derechos digitales (LOPDGDD). (Agencia Española de Protección de Datos, 2018). Estas normativas establecen un marco legal robusto para garantizar el tratamiento lícito, leal, transparente y seguro de los datos

personales de los estudiantes, profesores, personal administrativo y demás miembros de la comunidad universitaria.

Principales Aspectos Para Considerar

- **Designación del delegado de Protección de Datos (DPD):** Todas las universidades, tanto públicas como privadas, están obligadas a designar un DPD, quien será el responsable de velar por el cumplimiento de la normativa de protección de datos dentro de la institución.
- **Consentimiento Informado:** Las universidades deben obtener el consentimiento libre, informado, específico e inequívoco de los interesados antes de tratar sus datos personales. Este consentimiento debe ser documentado.
- **Principios Rectores:** El tratamiento de los datos personales debe basarse en los principios de licitud, lealtad, transparencia, limitación de la finalidad, minimización de datos, exactitud, integridad y confidencialidad.
- **Seguridad de los Datos:** Las universidades deben implementar medidas técnicas y organizativas adecuadas para garantizar la seguridad de los datos personales y protegerlos contra el tratamiento no autorizado o ilícito, la pérdida, la destrucción o el daño accidental.
- **Derechos de los Interesados:** Los titulares de los datos tienen derecho a acceder, rectificar, suprimir, oponerse al tratamiento, a la portabilidad de sus datos y a la limitación del tratamiento.
- **Notificación de Brechas de Seguridad:** En caso de producirse una brecha de seguridad que pueda poner en riesgo los derechos y libertades de los interesados, la universidad deberá notificarla a la Agencia Española de Protección de Datos (AEPD) y, en su caso, a los afectados.

Stakeholders y áreas del negocio involucradas

Al centralizar los datos de diferentes campañas en un solo conjunto de datos, facilitamos el análisis y la toma de decisiones. Se desarrollarán modelos de regresión utilizando datos históricos para segmentar de manera eficaz y optimizar el gasto en marketing y fidelización de clientes. Esto involucra a varios actores clave, incluidos los departamentos de marketing, admisiones y análisis de datos, además de considerar el comportamiento de los potenciales estudiantes. Las áreas directamente involucradas en este proceso son marketing, admisiones y análisis de datos.

Definir fuentes de información

Para el proyecto presentado se recopilará todas las fuentes involucradas en el proyecto, tanto internas como externas, y la relación existente entre ellas si la hubiese. Se deberá clasificar los tipos de datos con los que se trabajará. En el caso de que se necesitase hacer un tratamiento de la información, describir en qué consistirá este tratamiento.

Para el trabajo de titulación, se emplearán datos sintéticos según un enfoque similar al productivo, pero la fuente donde se almacena la data está en una base de datos SQL Server, por lo que, para garantizar la coherencia y estructura del entorno productivo, se replicará la misma estructura no productivo, esto nos permitirá analizar estrategias sin comprometer información confidencial.

Se usarán los datos históricos de campañas anteriores realizadas en una Institución de educación superior, con base en estos datos se seleccionarán las mejores variables para construir el modelo de predicción.

CAPÍTULO II

Regulación y protección de datos

Para proteger la privacidad de los datos reales y cumplir con las regulaciones de protección de datos se usarán datos sintéticos, lo que permitirá generar datos sintéticos que representen las características del CRM con la información de las campañas realizadas en una institución de educación superior. Dentro del esquema se entrenará el modelo de score sin revelar información confidencial sobre las personas reales y las características de estas. Se generará un dataset sintético que simule las características de los datos reales. Esto incluye variables relevantes como el comportamiento en el sitio web, la participación en eventos y la interacción con campañas que se realizaron en la institución de educación superior. Posteriormente con estos datos sintéticos se entrenará los modelos que asignaran score para identificar a los potenciales clientes. (Patki, Wedge, & Veeramachaneni, 2016).

Arquitectura del modelo

Aquí se definirá o se planteará un esquema de la arquitectura que se va a montar para este proyecto.

Un ejemplo de arquitectura es el siguiente:

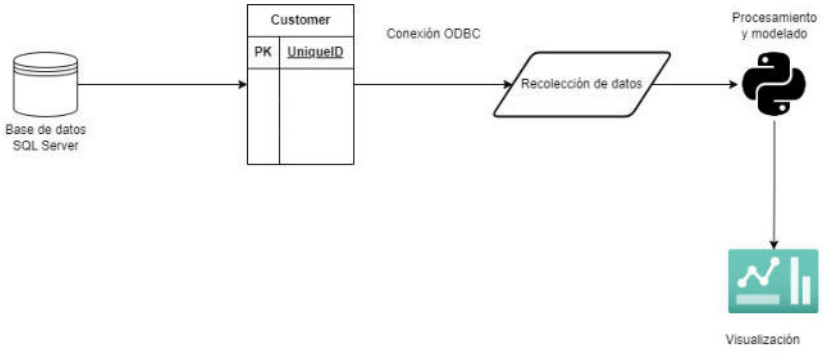

Figura 2 Arquitectura del modelo

Como el alcance de este trabajo de titulación no se contempla la instalación en el entorno productivo, la arquitectura anterior se basa en un entorno local, pero para definirlo, se debe considerar la infraestructura y presupuestos de la Institución de Educación Superior.

Infraestructura

En el caso de estudio actual, se dispone de servidores físicos que operan con el sistema operativo Windows, en un entorno on-premises. Para optimizar el procesamiento de datos, se propone la implementación de un proceso ETL (Extracción, Transformación y Carga). Este proceso permitirá extraer los datos de las bases donde se almacenan los insumos, procesarlos a través del modelo correspondiente y actualizar de manera continua el score de los clientes potenciales.

El uso de ETL es beneficioso en este contexto, ya que permite automatizar el flujo de datos. Por ejemplo, se puede configurar un job que procese automáticamente cualquier nuevo archivo creado, asegurando que el flujo de datos se mantenga actualizado y que los scores se reflejen en tiempo real.

Oportunidades de Negocio

A pesar de tener una infraestructura física, una buena alternativa es migrar a un modelo SAAS, ya que tanto en costo, escalabilidad, mantenimiento, respaldo son mayores. A continuación, se detalla un poco sobre los costos de ambos servicios:

Costos anuales infraestructura actual:

- **Servidores físicos:** Incluye la compra, instalación y mantenimiento de los servidores. Costo anual aprox ($30,000).
- **Licencias de software:** Windows Server, bases de datos, y cualquier otro software necesario. Costo anual aprox ($10,000).
- **Costos de energía y refrigeración:** Electricidad para mantener los servidores en funcionamiento. Costo anual aprox ($10,000).
- **Personal técnico:** Salarios para el personal encargado de mantenimiento y soporte. Costo anual aprox ($100,000).
- **Espacio físico:** Costos de alquiler o mantenimiento del espacio donde se alojan los servidores. Costo anual aprox ($15,000).
- **Depreciación de hardware:** El costo anual por la depreciación del hardware. Costo anual aprox ($10,000).
- **Respaldo y recuperación de datos:** Sistemas y procedimientos para asegurar la continuidad del negocio. Costo anual aprox ($10,000).
  Total: $185,000 aprox.

Aspectos para considerar para la migración en la nube:

- **Servicios de computación en la nube:** Dependiendo del proveedor (AWS, Azure, Google Cloud), y de los recursos necesarios (VMs, almacenamiento, bases de datos, etc.).

- **Licencias de software:** Algunas licencias pueden estar incluidas en los servicios en la nube, otras no.
- **Costos de transferencia de datos:** Gastos relacionados con la migración inicial y el tráfico continuo.
- **Mantenimiento y soporte:** Personal para la gestión y monitoreo de los servicios en la nube.
- **Respaldo y recuperación de datos:** Servicios en la nube para respaldo y recuperación.

Costos anuales infraestructura en la nube:

- **Servicios en la nube:** Microsoft Azure SQL Database, Azure Data Factory, Power BI Service: $60,000 (dependiendo del uso y la escalabilidad).
- **Transferencia de datos:** $5000 (costo inicial de migración y tráfico mensual).
- **Mantenimiento y soporte:** $50,000 costo aprox.
- **Respaldo y recuperación de datos:** Azure Backup $6000 (dependiendo del uso y la escalabilidad).
  Total: $121,000 aprox.

Migrar a la nube puede ofrecer una mayor flexibilidad, reducción de costos de mantenimiento y escalabilidad, especialmente si la carga de trabajo es variable o si se planea un crecimiento futuro. Este enfoque permite ajustar la capacidad y los servicios según sea necesario, lo cual es ideal para casos de estudio en crecimiento y con necesidades variables. Además, un modelo basado en SaaS proporciona la ventaja de que el proveedor se encarga del mantenimiento y las actualizaciones, lo que reduce los costos operativos asociados al personal. También es una solución flexible, ya que solo se paga por los recursos utilizados. Es importante considerar que esta migración impactará en otras áreas que podrían necesitarse trasladarse a la nube, por lo que es

fundamental analizar el panorama completo para tomar decisiones informada y evitar efectos negativos en otras áreas involucradas.

Planteamiento Agile

Para llevar a cabo un proyecto de BI para el sector educativo se tendrá la interacción del usuario con la página web de la institución, se utilizará la metodología Kanban. Se empezará por establecer un sistema de gestión visual claro, conocido comúnmente como tablero Kanban.

Kanban, que se traduce como "cartelera" en japonés, se originó como un sistema de control de inventario y cadena de suministro para la fabricación de Toyota en la década de 1940, con el objetivo de reducir el trabajo en curso y alinear el suministro de piezas de automóviles con la demanda.

En esencia, Kanban se basa en dos principios clave:

- Reducción del trabajo en curso
- Visualización del flujo de trabajo

Kanban comienza con una lista de características o tareas potenciales, similar al concepto de backlog en Scrum, que se colocan en la columna inicial "Por hacer" de un tablero Kanban. Este tablero sirve como una representación visual del flujo de trabajo. En un tablero Kanban simple de tres columnas, una vez que un equipo comienza a trabajar en una tarea, la tarjeta Kanban correspondiente se mueve de la columna "Por hacer" a la columna "En proceso". Cuando se completa la tarea, se mueve a la columna "Terminado". Los tableros Kanban pueden tener columnas adicionales; por ejemplo, los equipos de software pueden dividir la columna "En proceso" en "En desarrollo" y "En prueba". Algunos equipos dividen las "Pruebas" en "Verificación" y "Validación".

Aunque comenzó en la fabricación, Kanban ha sido adoptado por varias industrias, incluido el desarrollo de software. Recientemente, los equipos de ciencia de datos también lo han adoptado. Según una encuesta de 2020, Kanban fue el tercer proceso más popular utilizado en proyectos de ciencia de datos. Su popularidad se debe en parte a su simplicidad y alineación con las prácticas ágiles. (Anderson, 2010).

Kanban vs. Scrum

En comparación con Scrum, Kanban es menos prescriptivo. No define roles, reuniones ni bloques de tiempo. Muchos equipos de Scrum utilizan Kanban como un método secundario para gestionar el flujo de trabajo durante los sprints, aunque a menudo no implementan límites de trabajo en progreso.

Kanban puede ser eficaz para proyectos de ciencia de datos debido a sus procesos flexibles y menos rigurosos, que permiten a los científicos de datos trabajar sin la presión de plazos constantes. De manera similar a otros métodos ágiles, Kanban divide el trabajo en pequeños incrementos, lo que permite iteraciones rápidas y entrega continua. Para los equipos de ciencia de datos que carecen de estructura, Kanban puede proporcionar un marco que ayude a organizar sus procesos. (Anderson, 2010).

Se precisa sus beneficios:

- Altamente visual: la naturaleza visual de Kanban lo hace eficaz para comunicar rápidamente el trabajo en progreso a los miembros del equipo y las partes interesadas.
- Flexible: los equipos pueden incorporar tareas una por una, evitando el enfoque de ciclo por lotes de Scrum y permitiendo una mayor flexibilidad para cambiar las prioridades.
- Ligero y adaptable: sin límites de tiempo, roles ni reuniones, Kanban tiene menos gastos generales que los métodos en cascada y Scrum.

- Evita el choque cultural: Kanban no redefine los roles del equipo, lo que genera menos resistencia en la adopción en comparación con Scrum.
- Mejor coordinación: su simplicidad y flexibilidad pueden fomentar un mejor trabajo en equipo.
- Minimiza el trabajo en progreso: los límites de trabajo en progreso pueden aumentar la eficiencia general y reducir la cantidad de trabajo sin terminar.

Desafíos

- Falta de interacción con el cliente puesto que kanban no incluye inherentemente procesos para la retroalimentación regular del cliente, a diferencia de Scrum.
- Inconstancia de fechas limite, las tareas pueden tardar más de lo necesario en completarse, lo que requiere disciplina del equipo.
- Definición de la columna Kanban ya que al configurar un tablero Kanban adaptado a los procesos de ciencia de datos puede ser un desafío debido a la variabilidad de los pasos involucrados.
- Sin valor agregado para algunos: para ciertos equipos, Kanban puede parecer una carga administrativa innecesaria sin beneficios significativos.

Recomendaciones

Si bien Kanban puede ser eficaz para algunos equipos, es posible que deba complementarse con otros procesos que promuevan la interacción con el cliente. Es ideal para equipos que están en transición a Agile, aquellos que tienen dificultades con Scrum o equipos que necesitan flexibilidad y colaboración sin plazos estrictos. Kanban es ideal para organizaciones que prefieren mejoras graduales sin prácticas estrictamente prescriptivas.

Fase 1: Inicialización y puesta en marcha

Configuración del tablero Kanban: El proyecto arranca con la configuración del tablero Kanban, que se divide en varias columnas como “Por hacer”, “En curso”, “Test” y “Hecho”. Cada tarea o característica del proyecto BI está representada por una tarjeta Kanban que se mueve de izquierda a derecha a medida que avanza el trabajo.

Backlog: Se crea una lista más fluida para permitir adiciones y cambios continuos llamadas incidencias.

Fase 2: Flujo de trabajo continuo

Límites de trabajo en curso: Para garantizar la eficacia y gestionar la carga de trabajo, se establecen límites de trabajo en curso para cada columna del tablero. Estos límites evitan que el equipo se comprometa en exceso y ayudan a identificar los cuellos de botella en las primeras fases del proceso.

Registro de impedimentos: Los problemas o bloqueos identificados durante las reuniones eventuales o el trabajo regular se registran y destacan en el tablero Kanban o en un registro de impedimentos dedicado.

Fase 3: Seguimiento y Optimización

Entrega y Revisión Continuas: Kanban fomenta la entrega continua y la retroalimentación. A medida que las tareas se completan y se mueven a la columna “Hecho”, se revisan y se entregan de forma incremental, lo que permite ajustes continuos basados en la retroalimentación del equipo y el rendimiento del sistema.

Figura 3 Tablero de tareas en KANBAN

Se muestran 3 columnas en el tablero KANBAN las cuales representan una lista de cosas por hacer, en curso y listas (hechas). La limpieza de datos es el primer paso, y puede dar lugar a diversas subtareas asignadas a diferentes miembros del equipo. Por ejemplo, un desarrollador podría encargarse de la visualización de los datos, otro de crear el modelo y un tercero de realizar pruebas exhaustivas. Utilizando etiquetas específicas en cada tarjeta, podemos rastrear con precisión el progreso de cada subtarea y quién es el responsable. Una vez completadas todas las subtareas, la tarea principal se marca como 'Done'."

CAPÍTULO III

Fuentes de información

Se ha llevado a cabo un exhaustivo análisis de los leads generados a lo largo de siete periodos anteriores, con el objetivo de identificar patrones que influyen en la decisión de los estudiantes de inscribirse en la institución de educación superior. Como fuente principal para este análisis, se ha utilizado una robusta base de datos relacional en SQL Server, la cual ha permitido almacenar y organizar de manera sistemática los datos de los leads correspondientes a cada uno de los periodos evaluados.

Gracias a esta infraestructura, se ha logrado rastrear y analizar minuciosamente cada interacción que los potenciales estudiantes han tenido con la institución. Esto incluye la identificación de las diferentes fuentes a través de las cuales estos leads fueron generados, tales como canales digitales, correos electrónicos y formularios completados. Al desglosar estos datos, se ha obtenido una visión clara de cuáles de estos canales resultaron ser más efectivos en captar el interés de los estudiantes y, en última instancia, en influir en su decisión final de inscribirse. Este enfoque ha proporcionado información valiosa sobre las preferencias y comportamientos de los estudiantes durante el proceso de toma de decisiones.

Anonimización de información

En el desarrollo de un trabajo de titulación, la protección de la confidencialidad de los datos personales se ha identificado como un aspecto crucial y de vital importancia. Debido a la naturaleza altamente sensible de la información manejada, se ha determinado que no es factible divulgar los datos originales de los individuos involucrados en el estudio, ya que esto podría poner en riesgo su privacidad. Para cumplir con los rigurosos estándares éticos y normativos vigentes, se ha implementado un exhaustivo proceso de anonimización parcial, cuyo principal objetivo es evitar cualquier posibilidad de identificación directa o indirecta de los sujetos de estudio. Esto es esencial para garantizar que los participantes no puedan ser vinculados con los datos utilizados en la investigación.

Con el fin de abordar esta necesidad de manera efectiva y garantizar el cumplimiento de los requisitos éticos, se ha optado por la generación de datos sintéticos. Estos datos replican fielmente las características más relevantes de los datos reales, pero sin comprometer en ningún momento la privacidad de las personas involucradas. Para llevar a cabo esta tarea de manera eficiente, se ha empleado la librería Faker de Python, una herramienta ampliamente reconocida y valorada por su capacidad para crear datos ficticios de forma realista y coherente con la estructura de la información original.

La librería Faker permite la generación de una amplia variedad de elementos, tales como nombres, direcciones, números de identificación, y otros datos relevantes, manteniendo siempre la integridad del análisis. De esta forma, se garantiza la protección de la información sensible mientras se asegura que la base de datos utilizada, aunque ficticia, conserva todas las propiedades estadísticas necesarias para obtener resultados válidos y útiles para el análisis final (Meqlad, 2023). Esta técnica ha sido fundamental para que

el proyecto continúe avanzando de manera segura, sin comprometer la privacidad de los clientes potenciales y asegurando que todas las medidas de protección de datos sean respetadas, permitiendo que el trabajo avance conforme a los más altos estándares de seguridad y confidencialidad.

Análisis exploratorio de datos

El análisis exploratorio de datos es un método estadístico que utiliza técnicas visuales y numéricas para investigar a fondo un conjunto de datos, con el objetivo de revelar patrones subyacentes, identificar valores atípicos y explorar relaciones entre variables.

Al realizar el EDA en este proyecto nos ayudará a comprender de mejor manera la distribución de las variables, explorar posibles relaciones entre variables y poder identificar si los datos que poseemos se encuentran óptimos para la elaboración del modelo. (Tukey, 1977)

Inicialmente, se realizó un proceso de limpieza para identificar y corregir problemas en los datos. Se detectaron variables con valores ausentes, para mitigar este problema, se aplicaron diferentes estrategias, como la imputación de valores faltantes con valores que se encontraban más comunes en cada una de las variables, al ser estas variables categóricas se realizó un análisis de cada una de ellas y se logró determinar que valores eran los más comunes y con ellos hacer la imputación de los valores faltantes.

Se hizo uso de gráficos de barras para la identificación de las variables óptimas para el modelo. Las variables categóricas fueron codificadas utilizando One-Hot Encoding para garantizar que el modelo pudiera interpretarlas correctamente. Esto creó nuevas columnas binarias para cada categoría única. De igual manera, la variable objetivo, que

indica si un cliente es potencial o no, fue transformada en una variable binaria con valores 1 (Closed_won) y 0 (Closed_lost), facilitando su uso en la predicción con el modelo de clasificación.

Modelos de Machine Learning

En este apartado se describen los modelos de predicción utilizados para identificar posibles clientes en campañas de marketing universitarias. Se han seleccionado tres algoritmos de machine learning: K-Nearest Neighbors (KNN), Redes Neuronales y Random Forest, dada su eficacia para la clasificación de datos y su capacidad para manejar conjuntos de datos complejos. A continuación, se detallan los fundamentos y la implementación de cada uno de estos modelos.

Redes Neuronales

Las Redes Neuronales son modelos de machine learning inspirados en el cerebro humano, que se componen de capas de neuronas interconectadas. Estas capas procesan la información a través de funciones matemáticas y ajustan sus parámetros durante el entrenamiento para mejorar la precisión de las predicciones. (Goodfellow, Bengio, & Courville, 2016).

Funcionamiento Básico:

Una red neuronal consta de una capa de entrada, una o más capas ocultas, y una capa de salida. Las neuronas de cada capa transforman la información y la pasan a la siguiente capa para generar una predicción final. (Goodfellow, Bengio, & Courville, 2016).

Aplicación en el Proyecto:

En el contexto de este proyecto, se utilizó una red neuronal con varias capas ocultas para mejorar la precisión de la predicción. Se ajustaron hiperparámetros como el número

de capas, neuronas por capa y la tasa de aprendizaje. El modelo fue entrenado utilizando los datos de campañas de marketing, y evaluado en términos de precisión y sensibilidad.

K-Nearest Neighbors (KNN)

El modelo K-Nearest Neighbors (KNN) es un algoritmo de clasificación que asigna una clase a una nueva instancia basándose en las clases de sus "K" vecinos más cercanos. La cercanía se mide a través de la distancia euclidiana entre los puntos de datos. (Mitchell, 1997)

Funcionamiento Básico:

KNN compara la nueva instancia con las "K" observaciones más cercanas en el conjunto de datos de entrenamiento. La clase más frecuente entre estos vecinos es asignada a la nueva instancia. (Mitchell, 1997).

Aplicación en el Proyecto:

En este proyecto, KNN fue implementado para clasificar las respuestas de los estudiantes a las campañas de marketing. Se evaluaron diferentes valores de "K" para encontrar el que produjera la mejor precisión y balance entre clases.

Random Forest

Random Forest es un algoritmo de machine learning basado en el uso de múltiples árboles de decisión para mejorar la precisión de las predicciones. Este modelo combina las predicciones de varios árboles, lo que reduce el riesgo de sobreajuste y aumenta la robustez del modelo. (Breiman, 2001).

Funcionamiento Básico:

Random Forest crea varios árboles de decisión a partir de muestras aleatorias del conjunto de datos. La predicción final se obtiene mediante la votación mayoritaria entre los resultados de estos árboles, lo que mejora la estabilidad y precisión del modelo. (Breiman, 2001).

Aplicación en el Proyecto:

En este proyecto, Random Forest fue utilizado para predecir el interés de los estudiantes en base a las respuestas recibidas en las campañas de marketing. Se ajustaron parámetros como el número de árboles y la profundidad máxima de cada árbol para mejorar el rendimiento del modelo.

CAPITULO IV

Extracción de datos

La información necesaria para nuestro modelo está almacenada en una base de datos SQL Server. Para acceder a estos datos, se estableció una conexión desde un entorno Jupyter Notebook, permitiendo ejecutar consultas directamente desde el código y obtener la data requerida para el análisis.

```
server = 'DAVID\\SQLEXPRESS'  # Nombre del servidor
database = 'BDD_PBL'  # Nombre de la base de datos
connection_string = (
    'DRIVER={SQL Server};'
    'SERVER=' + server + ';'
    'DATABASE=' + database + ';'
    'Trusted_Connection=yes;'
)
```

Se define el servidor y la base de datos a la cual se desea conectar. Estas variables almacenan el nombre del servidor y la base de datos de SQL Server.

Se crea una cadena de conexión (connection_string) que indica el tipo de controlador (DRIVER={SQL Server}), el servidor al que se accede, la base de datos, y especifica que se usará una conexión de confianza (con autenticación basada en las credenciales de Windows.

```
conn = pyodbc.connect(connection_string)
```
0.0s

```
cursor = conn.cursor()
```
0.0s

```
query = "SELECT * FROM data_unificada_final"
cursor.execute(query)
columns = [column[0] for column in cursor.description]
rows = cursor.fetchall()
cursor.close()
conn.close()
```

Mediante el cursor, se ejecuta la consulta SQL SELECT * FROM data_unificada_final, que selecciona todas las columnas y filas de la tabla data_unificada_final y se recuperan todas las filas resultantes de la consulta utilizando rows = cursor.fetchall(), lo que almacena los datos obtenidos en una lista de tuplas.

Finalmente, se cierra el cursor y la conexión a la base de datos para liberar los recursos.

Desarrollo del EDA

En primer lugar se observo las dimensiones del dataset, en donde se obtuvieron las primeras filas y los valores nulos por cada columna.

```python
#DataFrame
print("Dimensiones del DataFrame:", datos.shape)
print("Datos nulos por columna:")
print(datos.head)
print(datos.isnull().sum())
```

Se pudo observar la información general sobre las columnas y los tipos de datos y la suma de valores nulos para verificar que no exista ninguno.

```python
print(datos.info())
print(datos.isnull().sum())
```

```
 11  Edad_grupos                  49309 non-null  object
 12  TieneDiscapacidad            49309 non-null  object
 13  TrabajaActualmente           49309 non-null  object
 14  TipoAdmision                 49309 non-null  object
 15  EstadoCivil_1                49309 non-null  object
 16  CitaAdmisiones               49309 non-null  object
 17  pi_score_c_grupos            49309 non-null  object
 18  Ciudad                       49309 non-null  object
 19  LeadContacto                 49309 non-null  object
 20  NoLlamar                     49309 non-null  object
 21  ListaNegra                   49309 non-null  object
 22  tour_c                       49309 non-null  object
 23  Titulo_Registrado_Senescyt_c 49309 non-null  object
 24  Identificación               49309 non-null  object
dtypes: object(25)
memory usage: 9.4+ MB
None
Nombre_completo                0
Programa                       0
Estado_de_candidato            0
Afluente                       0
Periodo                        0
Origen_del_candidato           0
CodPrograma                    0
Genero                         0
Provincia                      0
Campaña                        0
TipoColegio                    0
Edad_grupos                    0
TieneDiscapacidad              0
```

Para tener una visualizacion más clara sobre los datos del dataset se hizo uso de graficos en barra para poder tener una mejor comprensión de lo datos. Se obtuvieron las gráficas de las 10 categorías más frecuentes en variables con más de 10 categorías.

```python
# Función para graficar solo las 10 categorías más frecuentes en variables con más de 10 categorías
def plot_top_10_categories(df, column, target, palette='Set2'):
    plt.figure(figsize=(14, 8))

    if df[column].nunique() > 10:
        top_10 = df[column].value_counts().nlargest(10).index
        sns.countplot(data=df[df[column].isin(top_10)], x=column, hue=target, palette=palette)
    else:
        sns.countplot(data=df, x=column, hue=target, palette=palette)

    plt.title(f'Distribución de {column} con respecto a {target}')
    plt.xticks(rotation=45)
    plt.tight_layout()
    plt.show()

# Graficar la variable objetivo 'Estado_de_candidato'
plt.figure(figsize=(8, 6))
sns.countplot(data=datos, x='Estado_de_candidato', palette='Set2')
plt.title('Distribución de la Variable Objetivo: Estado_de_candidato')
plt.show()

# Graficar para todas las columnas categóricas
categorical_columns = ['Genero', 'Provincia', 'Edad_grupos', 'Programa', 'Periodo', 'Origen_del_candidato',
                       'CodPrograma', 'EstadoCivil_1', 'pi_score_c_grupos', 'Ciudad']

for column in categorical_columns:
    plot_top_10_categories(datos, column, 'Estado_de_candidato')
```

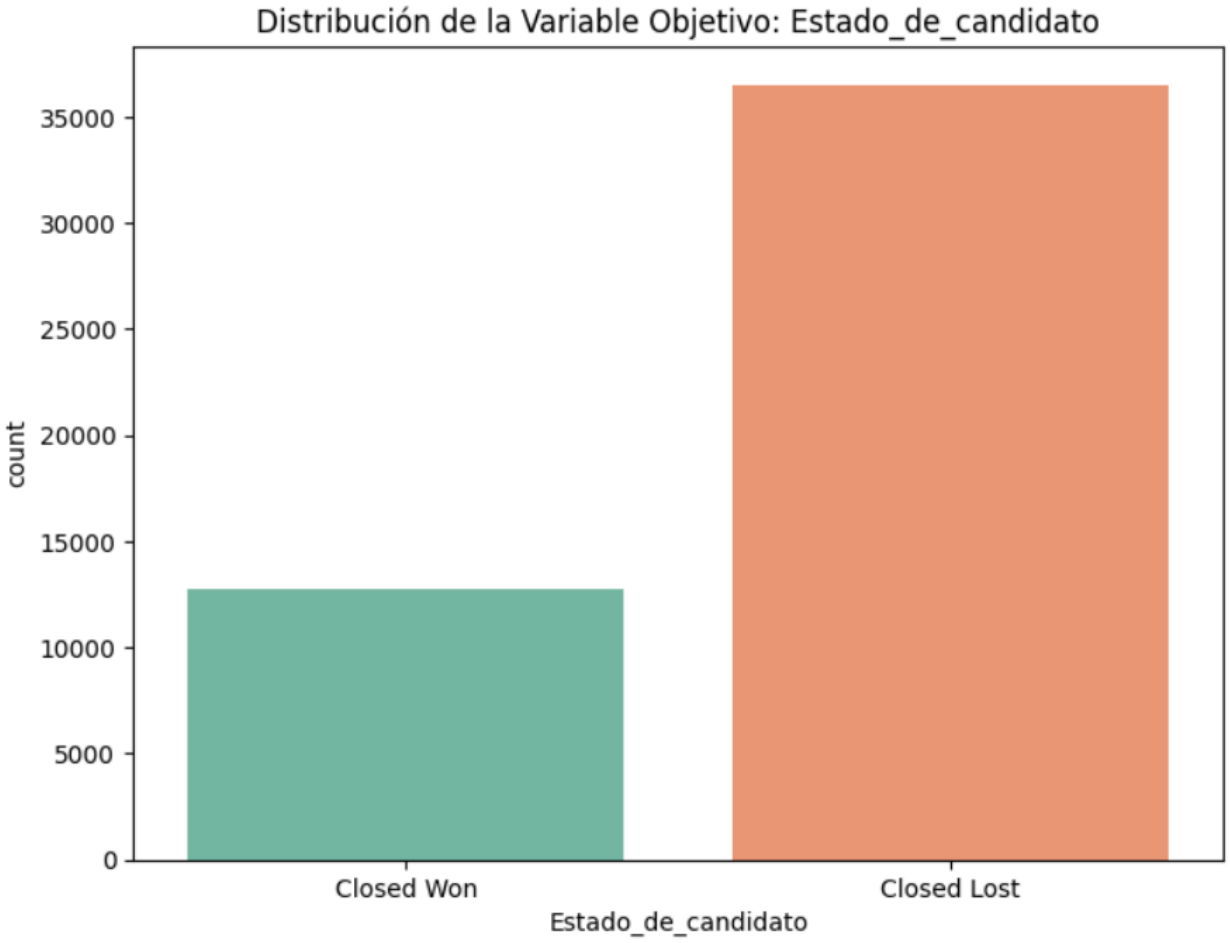

Figura 4 Distribución de Estado_de_candidato

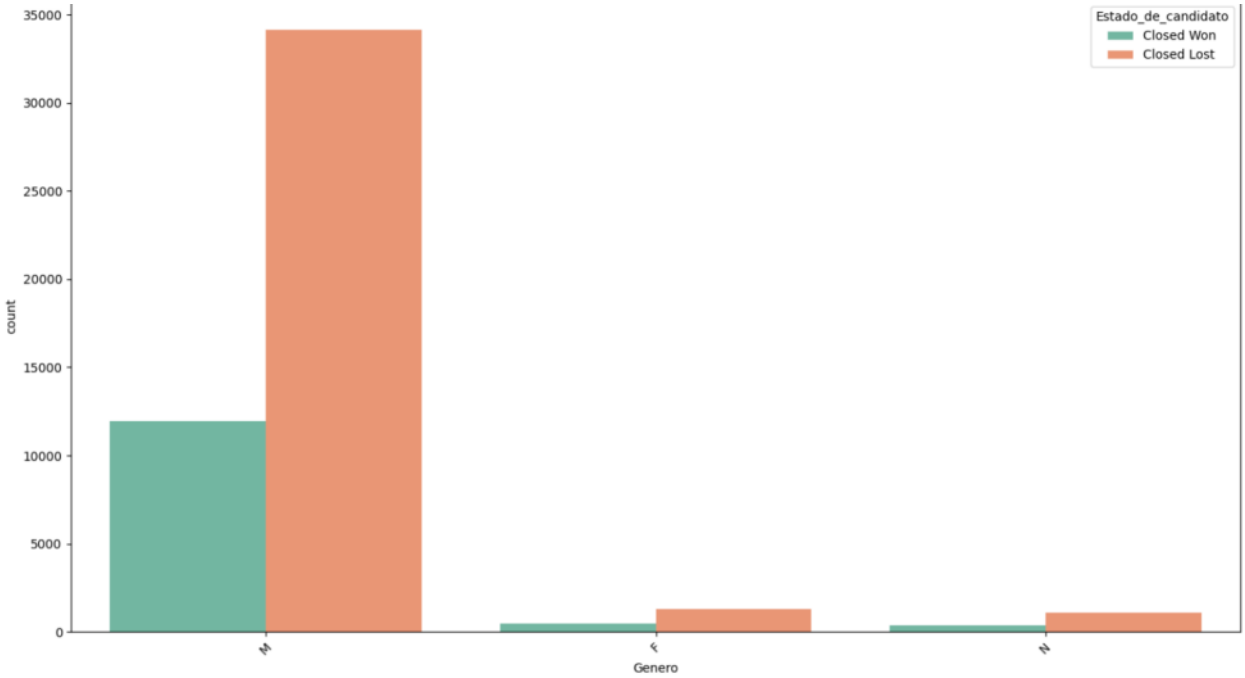

Figura 5 Distribución de estado de candidato por género

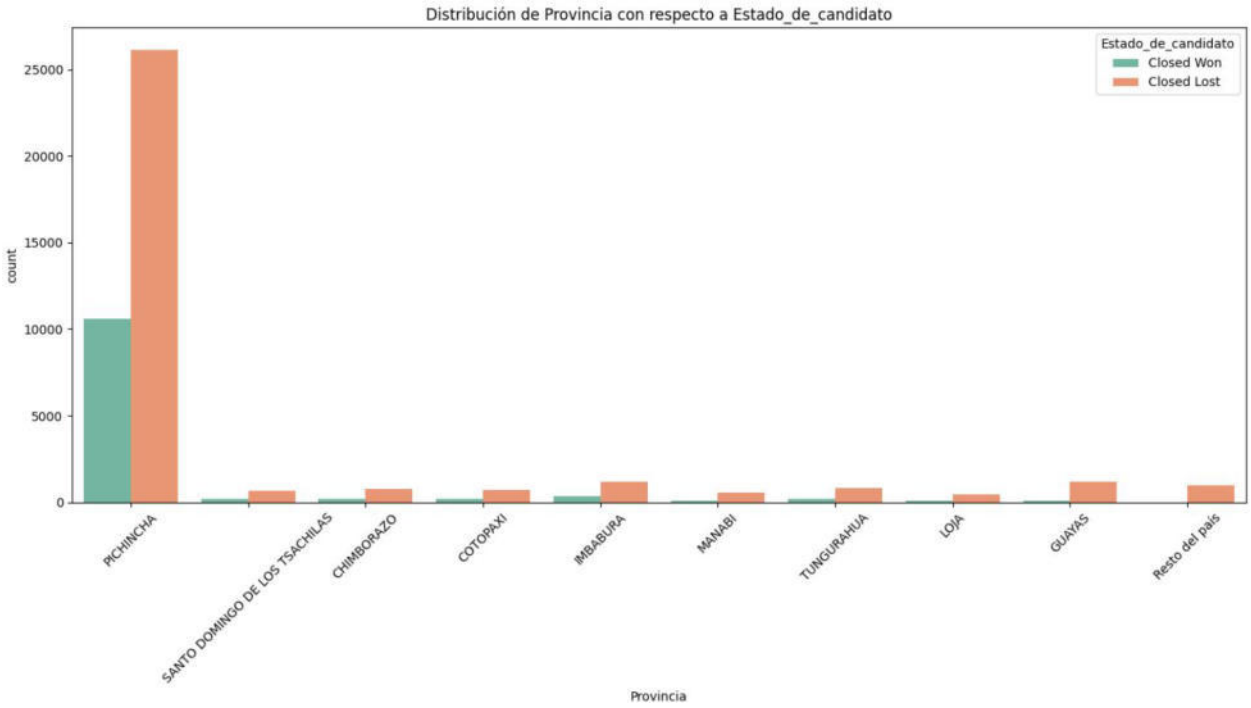

Figura 6 Distribución de estado de candidato por provincia

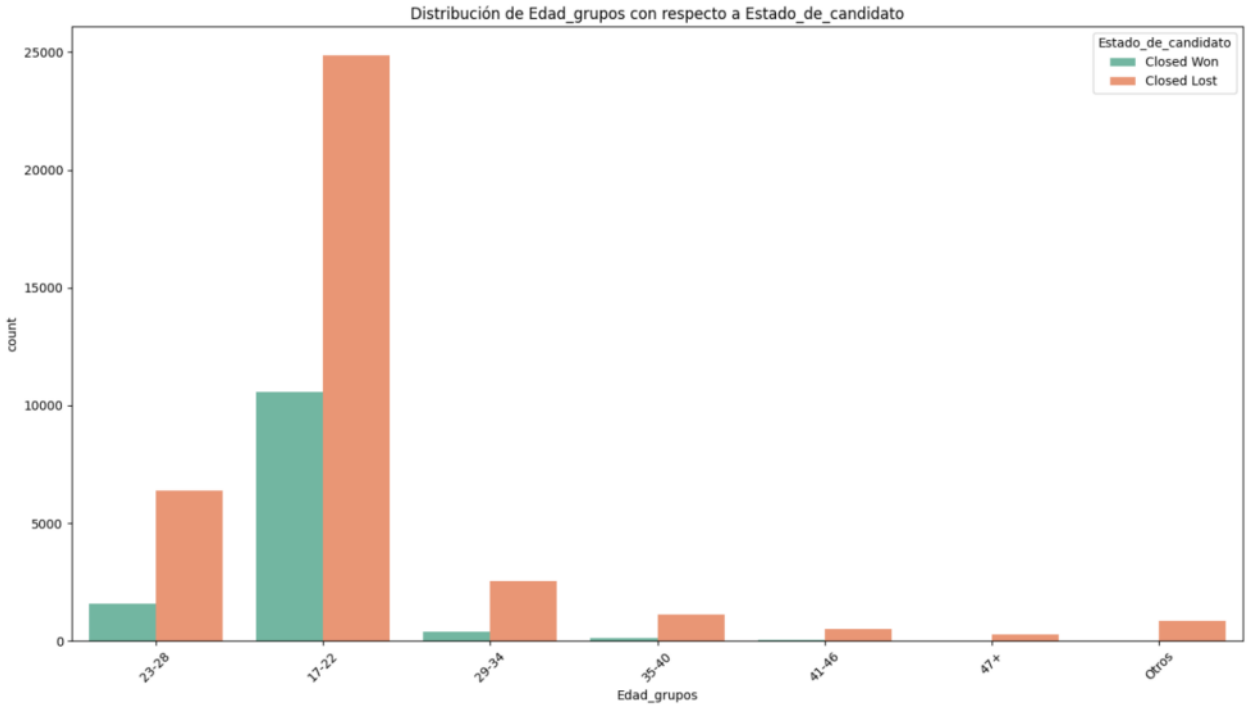

Figura 7 Distribución de estado de candidato por edad

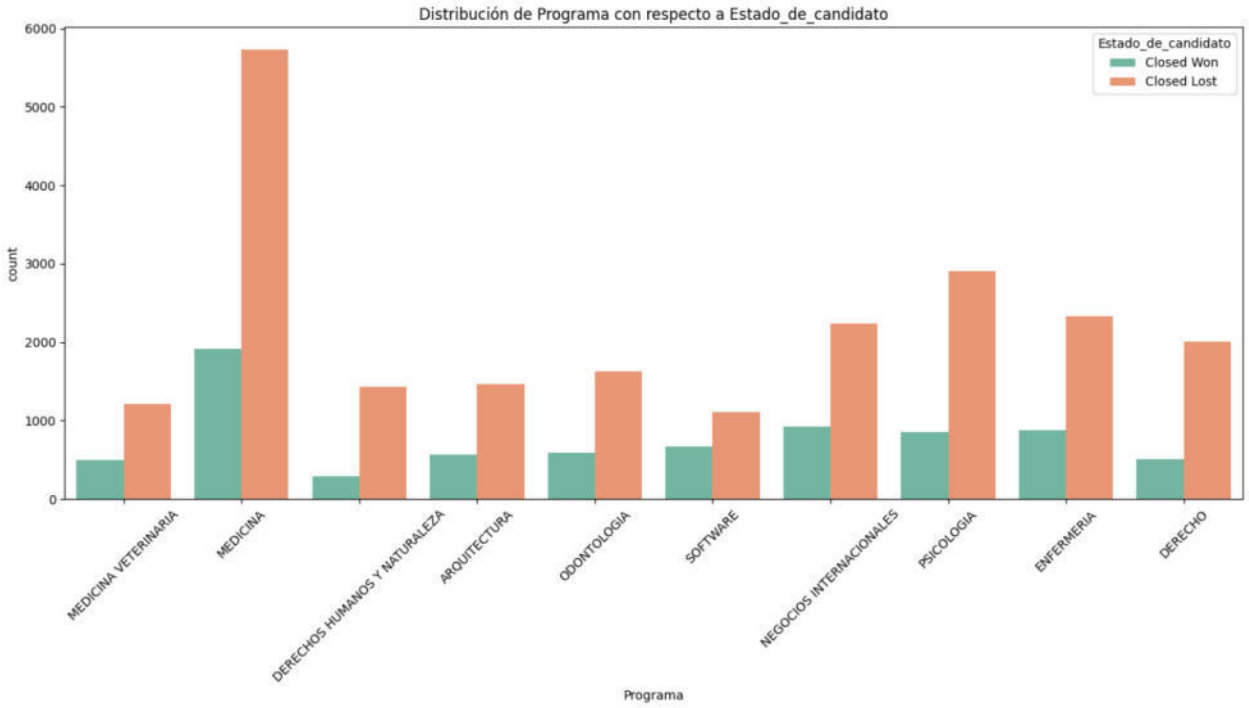

Figura 8 Distribución de estado de candidato por edad

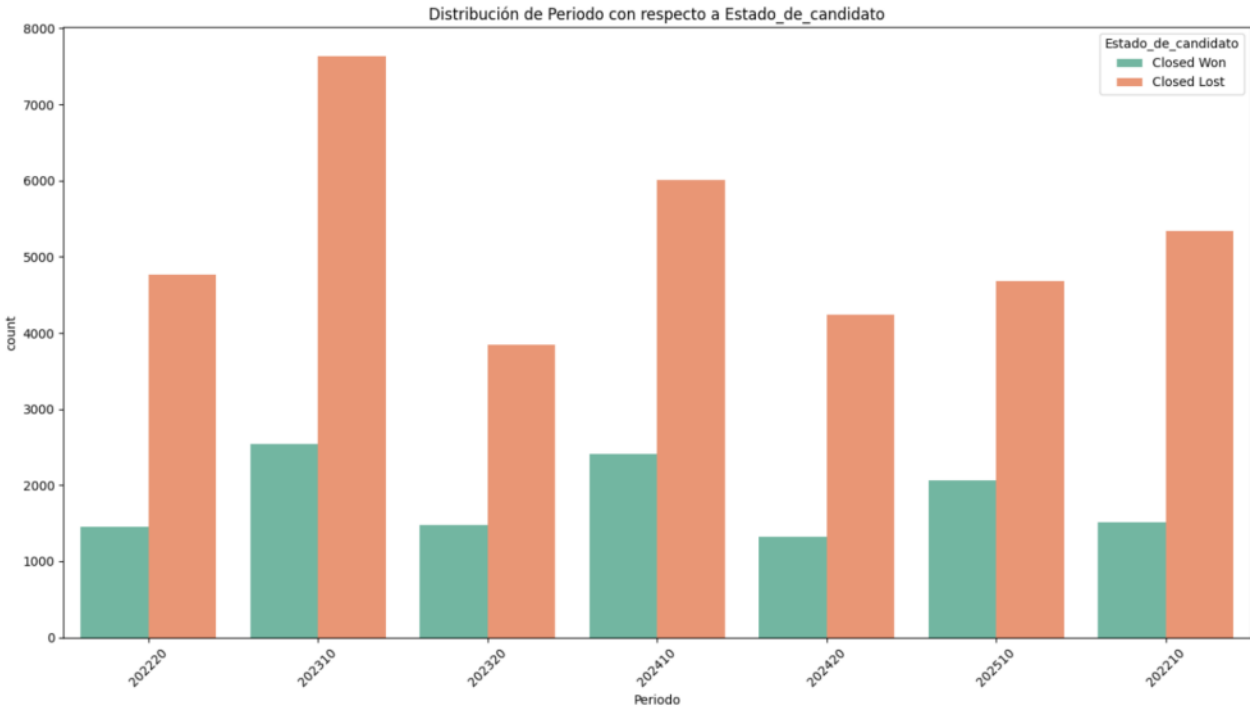

Figura 9 Distribución de estado de candidato por periodo

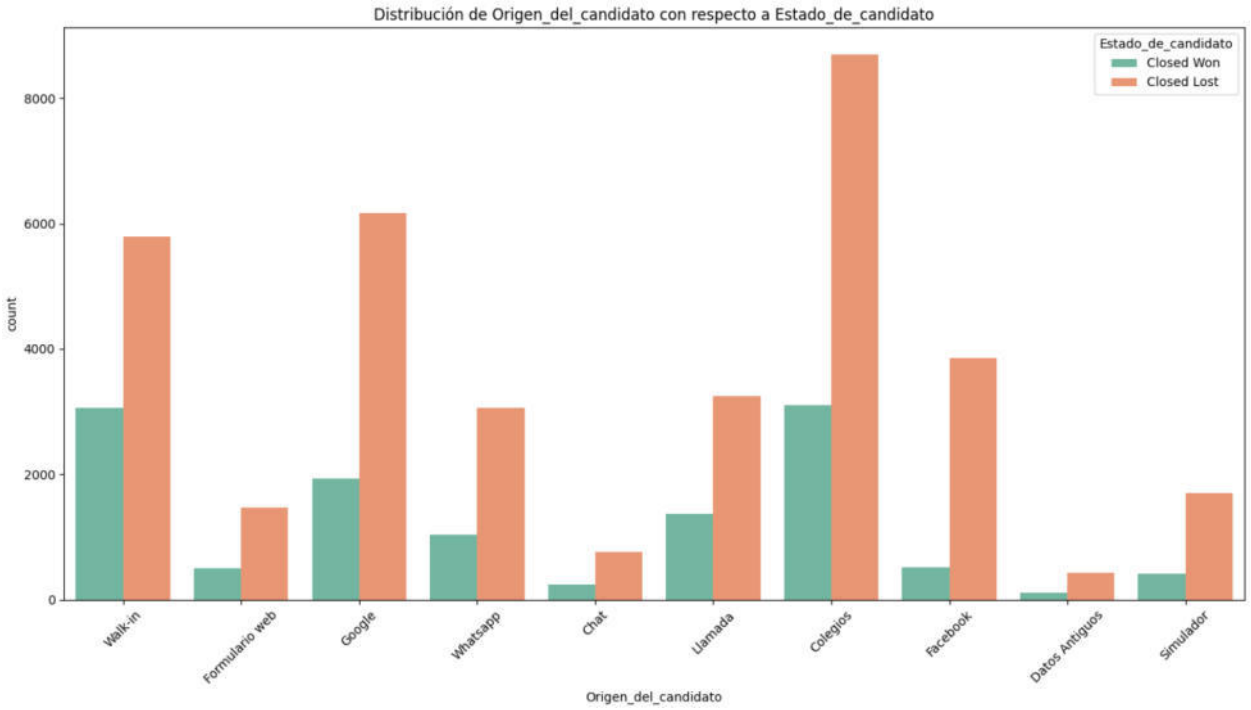

Figura 100 Distribución de estado de candidato por origen

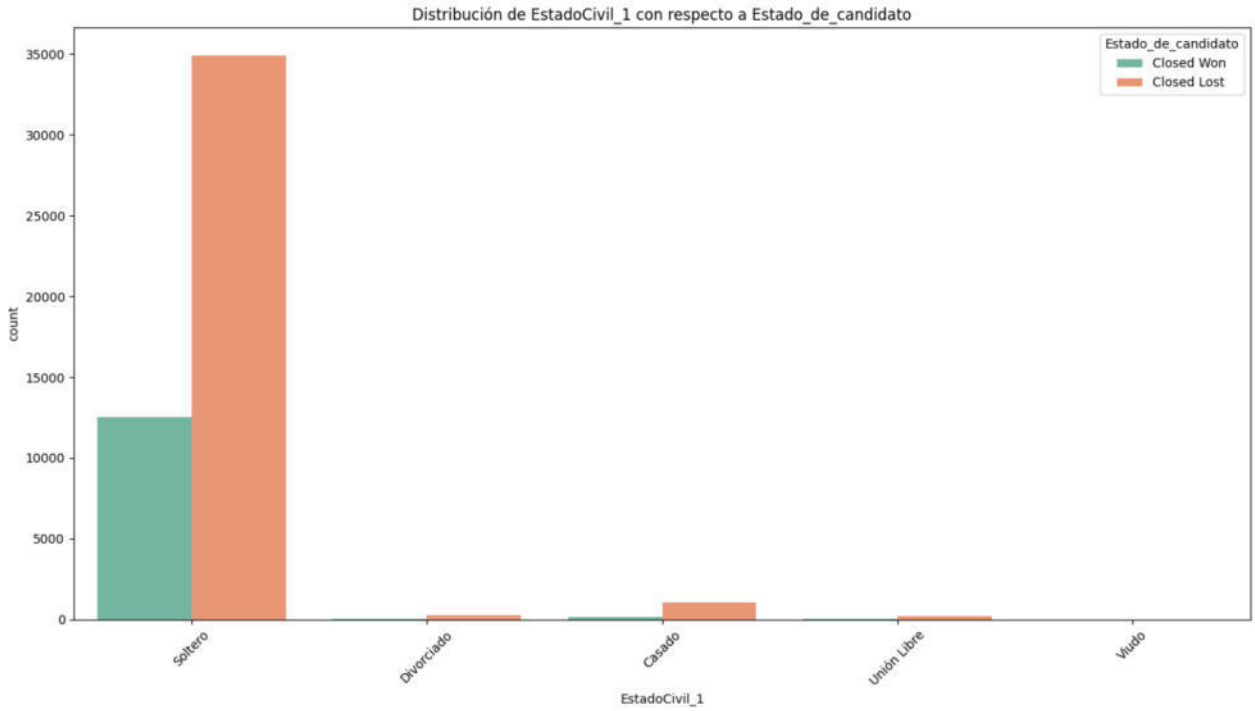

Figura 111 Distribución de estado de candidato por estado civil

Se verificaron los valores duplicados y se procedió a borrarlos.

```
print(datos['Titulo_Registrado_Senescyt_c'].drop_duplicates())
```

```
0     0
36    1
Name: Titulo_Registrado_Senescyt_c, dtype: object
```

Figura 122 Verificación de datos duplicados

Para la conversión de variables categóricas a numéricas se utilizó el método de One-Hot Encoding y Label Encoding.

```
# Paso 1: Aplicar LabelEncoder para variables categóricas binarias (con dos valores)
label_columns = ['Estado_de_candidato', 'Afluente', 'TipoColegio', 'TieneDiscapacidad',
                 'TrabajaActualmente', 'TipoAdmision', 'CitaAdmisiones', 'NoLlamar',
                 'ListaNegra', 'tour_c','Titulo_Registrado_Senescyt_c']

le_dict = {}  # Diccionario para guardar los codificadores y revertir luego

for column in label_columns:
    le = LabelEncoder()
    datos[column] = le.fit_transform(datos[column])
    le_dict[column] = le  # Guardar el codificador para revertir luego

# Paso 2: Aplicar OneHotEncoder para variables con más de dos categorías
```

```
one_hot_columns = ['Programa', 'Origen_del_candidato',
                   'Genero', 'Provincia', 'Campaña', 'Edad_grupos',
'EstadoCivil_1', 'pi_score_c_grupos', ]


# Guardar las columnas originales que serán OneHotEncoded
one_hot_originals = datos[one_hot_columns].copy()


# Aplicar OneHotEncoding y agregar las nuevas columnas al dataset
datos = pd.get_dummies(datos, columns=one_hot_columns, drop_first=True)


# Paso 3: Convertir las columnas que pueden ser numéricas a tipo numérico
for column in datos.columns:
    try:
        datos[column] = pd.to_numeric(datos[column])
    except ValueError:
        continue


# Mostrar las primeras filas del dataset transformado
print(datos.head())
```

Resultados para la conversión de variables categóricas a numéricas

```
                Nombre_completo  Estado_de_candidato  Afluente  Periodo  \
0      María Ángeles del Laguna                    0         0   202210
1         Valeria del Francisco                    0         0   202210
2  Fabiola Salud Márquez Iborra                    0         0   202210
3       Olalla Aura Paz Estevez                    0         0   202210
4       Valentín Nicolau Morera                    0         0   202210

  CodPrograma  TipoColegio  TieneDiscapacidad  TrabajaActualmente  \
0       2P567            0                  0                   0
1       1P634            0                  0                   0
2       1P524            0                  0                   0
3       1P791            0                  0                   0
4       2P257            0                  0                   0

   TipoAdmision  CitaAdmisiones  ... pi_score_c_grupos_101-200  \
0             0               0  ...                         0
1             0               1  ...                         1
2             1               1  ...                         0
3             1               0  ...                         0
4             1               1  ...                         0

  pi_score_c_grupos_201-300  pi_score_c_grupos_301-400  \
0                         0                          0
1                         0                          0
2                         0                          0
...
3                           0
4                           0
```

Figura 133 Resultados conversión de variables categóricas a numéricas

Desarrollo de los modelos

```
# Dividir el dataset en entrenamiento y prueba (80% entrenamiento, 20% prueba)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)


# Estandarizar los datos (necesario para KNN)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

Con las variables seleccionadas, se procedió a implementar y evaluar tres modelos distintos: K-Nearest Neighbors (KNN), Random Forest y Redes Neuronales. Para cada modelo, se configuraron hiperparámetros específicos para garantizar un ajuste óptimo:

Ahora cada modelo fue entrenado usando técnicas de validación cruzada, lo cual fortalece la fiabilidad y la generalización de los resultados al evaluar el rendimiento del modelo en diferentes subconjuntos del dataset.

Finalmente, se presentan los resultados de cada modelo, comparando meticulosamente su precisión, recall y F1-score. Este análisis comparativo ayuda a discernir cuál modelo ofrece la mejor capacidad predictiva para determinar el estado de los candidatos basado en las características seleccionadas.

Modelo KNN

```
# Entrenar el modelo

knn.fit(X_train_scaled, y_train)



# Predecir los resultados

y_pred_knn = knn.predict(X_test_scaled)



# Precisión del modelo

accuracy_knn = accuracy_score(y_test, y_pred_knn)

print(f"\nPrecisión del modelo KNN: {accuracy_knn:.4f}")



# Matriz de confusión

cm_knn = confusion_matrix(y_test, y_pred_knn)



# Reporte de clasificación

report_knn = classification_report(y_test, y_pred_knn, digits=4)

print("\n=== Informe de clasificación para KNN ===\n")

print(report_knn)
```

```
Precisión del modelo KNN: 0.8189

=== Informe de clasificación para KNN ===

              precision    recall  f1-score   support

           0     0.8576    0.9037    0.8800      7249
           1     0.6860    0.5836    0.6307      2613

    accuracy                         0.8189      9862
   macro avg     0.7718    0.7437    0.7554      9862
weighted avg     0.8121    0.8189    0.8140      9862

                Nombre_completo CodPrograma        LeadContacto  \
11127  Eufemia Monreal Carranza       1H564  0038W00001VUHztQAH
3564   Aureliano España Gimenez       1P564  0038W00001XBcyyQAD
48830          Josep Maza Milla       1P704  0031U00001mVF4SQAW
9738      Caridad Vilar Iglesia       1P704  0038W00001pzc2hQAA
147       Tatiana Barros Franco       1P724  003U1000003gFZ6IAM

       Identificación  Periodo     Ciudad  Prediccion_KNN
11127      1350221423   202220   ORELLANA               0
3564       5218840779   202220      QUITO               1
48830      1054930461   202210      QUITO               0
9738       1603355994   202320    OTAVALO               1
147        6441185721   202420  GUAYAQUIL               0
```

Figura 144 Informe de clasificación para KNN

El modelo KNN, ajustado con los mejores hiperparámetros, mostró una precisión general del 81.89%. Al desglosar las métricas por clases, se observa que el modelo alcanzó una precisión del 85.76% para la clase 0 y del 68.60% para la clase 1. Sin embargo, el recall

de la clase 0 (Closed Lost) fue más alto (90.37%) en comparación con el de la clase 1 (Closed Win) (58.36%), lo que refleja que el modelo tuvo un mejor desempeño al identificar correctamente los elementos de la clase 0. El f1-score, que equilibra precisión y recall, fue de 88.00% para la clase 0 y de 63.07% para la clase 1. Estos resultados indican que el modelo es más efectivo para la clase mayoritaria, mientras que la clase 1 presenta mayor dificultad para ser correctamente clasificada.

En cuanto a la predicción de los datos, el DataFrame generado incluye las características seleccionadas, como el nombre completo de los individuos, el programa de estudio y la ciudad, junto con las predicciones hechas por el modelo.

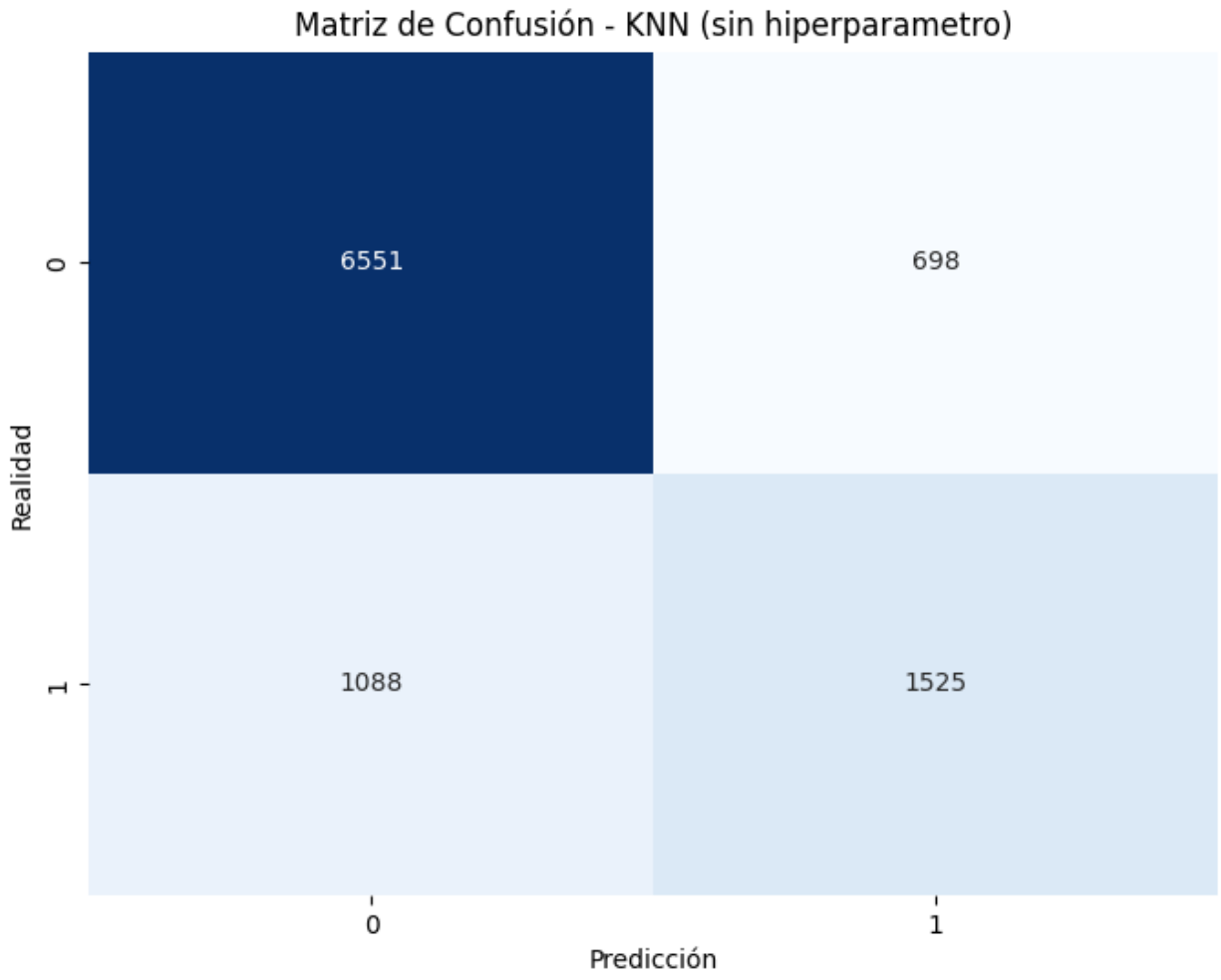

Figura 155 Matriz de confusión - KNN (Sin hiperparámetro)

```
# Inicializar el modelo base KNN
knn = KNeighborsClassifier()


# Definir los hiperparámetros que queremos ajustar
param_dist = {
    'n_neighbors': [3, 5, 7, 9, 11, 13, 15],  # Número de vecinos a probar
    'weights': ['uniform', 'distance'],  # Peso basado en distancia o uniforme
    'metric': ['euclidean', 'manhattan', 'minkowski'],  # Diferentes métricas de distancia
    'p': [1, 2]  # p=1 es manhattan, p=2 es euclidiana
}


# Inicializar RandomizedSearchCV
random_search = RandomizedSearchCV(estimator=knn, param_distributions=param_dist, n_iter=20, cv=3, verbose=2, n_jobs=-1, random_state=42)


# Ajustar el modelo usando RandomizedSearchCV
random_search.fit(X_train_scaled, y_train)


# Mostrar los mejores hiperparámetros
```

```
print(f"Mejores hiperparámetros: {random_search.best_params_}")


# Entrenar el modelo con los mejores hiperparámetros
best_knn_model = random_search.best_estimator_


# Predecir los resultados
y_pred_best_knn = best_knn_model.predict(X_test_scaled)


# Precisión del modelo
accuracy_best_knn = accuracy_score(y_test, y_pred_best_knn)
print(f"\nPrecisión del modelo KNN (mejores hiperparámetros):
{accuracy_best_knn:.4f}")


# Matriz de confusión
cm_best_knn = confusion_matrix(y_test, y_pred_best_knn)


# Reporte de clasificación
report_best_knn = classification_report(y_test, y_pred_best_knn, digits=4)
print("\n=== Informe de clasificación para KNN (mejores hiperparámetros) ===\n")
print(report_best_knn)
```

```
Fitting 3 folds for each of 20 candidates, totalling 60 fits
Mejores hiperparámetros: {'weights': 'distance', 'p': 1, 'n_neighbors': 15, 'metric': 'minkowski'}

Precisión del modelo KNN (mejores hiperparámetros): 0.8431

=== Informe de clasificación para KNN (mejores hiperparámetros) ===

              precision    recall  f1-score   support

           0     0.8696    0.9254    0.8966      7249
           1     0.7481    0.6150    0.6751      2613

    accuracy                         0.8431      9862
   macro avg     0.8089    0.7702    0.7858      9862
weighted avg     0.8374    0.8431    0.8379      9862

               Nombre_completo CodPrograma        LeadContacto  \
11127  Eufemia Monreal Carranza       1H564  0038W00001VUHztQAH
3564   Aureliano España Gimenez       1P564  0038W00001XBcyyQAD
48830          Josep Maza Milla       1P704  0031U00001mVF4SQAW
9738      Caridad Vilar Iglesia       1P704  0038W00001pzc2hQAA
147       Tatiana Barros Franco       1P724  003U1000003gFZ6IAM

       Identificación  Periodo     Ciudad  Prediccion_KNN
11127      1350221423   202220   ORELLANA              0
3564       5218840779   202220      QUITO              0
48830      1054930461   202210      QUITO              0
9738       1603355994   202320    OTAVALO              1
147        6441185721   202420  GUAYAQUIL              0
```

Figura 166 Informe de clasificación para KNN (mejores hiperparámetros)

El algoritmo KNN, tras ser optimizado con los mejores parámetros mediante RandomizedSearchCV, logró una precisión general del 84.31%. Los parámetros seleccionados incluyen pesos basados en la distancia, la métrica Manhattan (p=1) y el uso de 15 vecinos con la distancia Minkowski.

Al examinar los resultados por clase, la clase 0 mostró un desempeño sólido con una precisión del 86.96%, un recall del 92.54% y un f1-score de 89.66%, lo que indica una alta capacidad del modelo para reconocer correctamente esta clase. Por otro lado, para

la clase 1, la precisión fue de 74.81%, el recall de 61.50% y el f1-score de 67.51%, lo que revela mayores dificultades para predecir correctamente los casos de esta clase minoritaria.

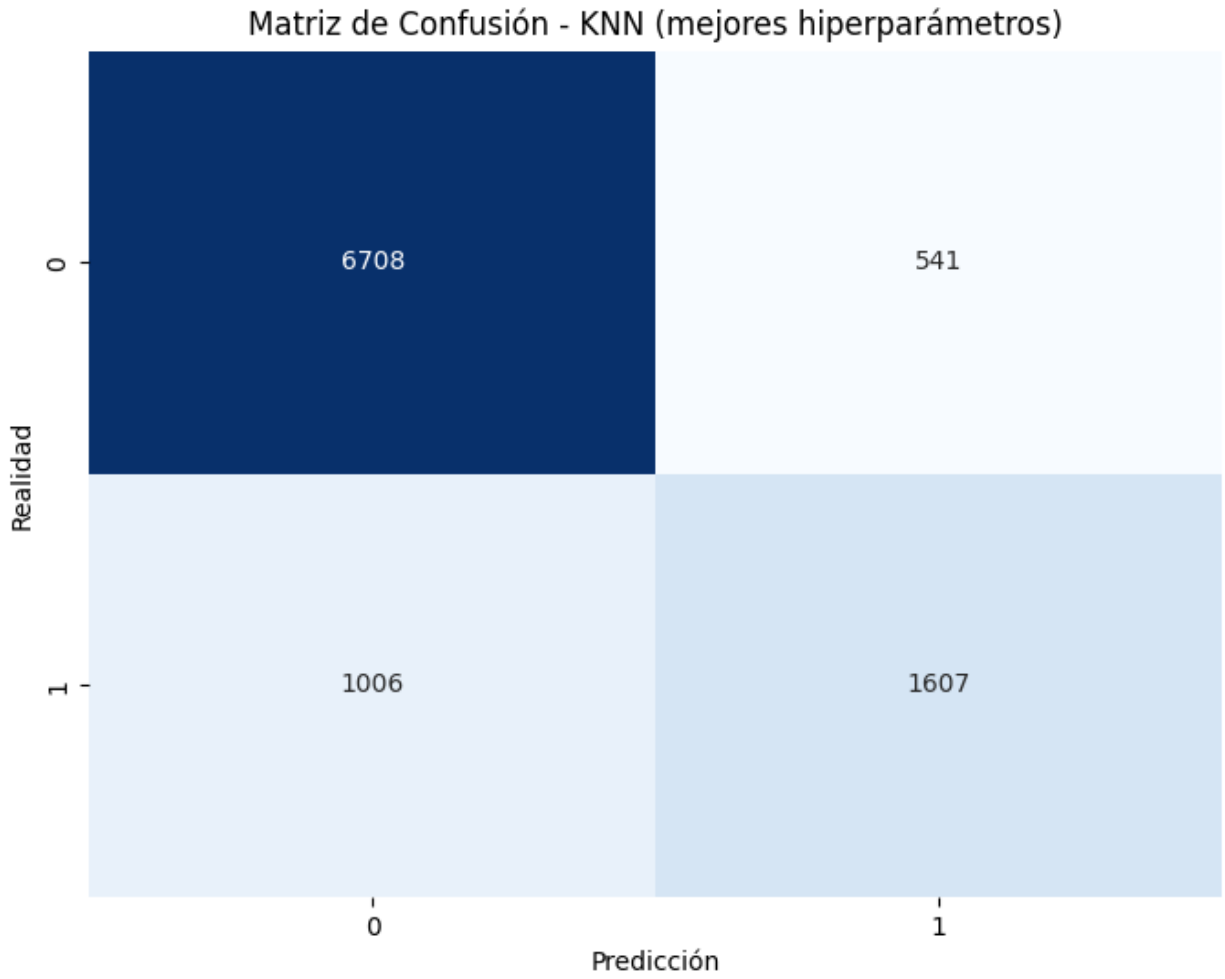

Figura 177 Matriz de confusión - KNN (Mejores hiperparámetros)

Random Forest

En el desarrollo del modelo de RandomForest, se implementaron dos enfoques con el objetivo de optimizar su rendimiento y realizar un análisis comparativo entre ambos:

1. Enfoque sin Selección de Características (Sin SelectKBest)

En el primer enfoque, se entrenó el modelo de RandomForest utilizando todas las características disponibles sin aplicar un proceso previo de selección de las mismas. Este enfoque considerara toda la información proporcionada por las variables de entrada, permitiendo al modelo explorar la totalidad de los datos.

Para este enfoque se usaron los siguientes criterios para el desarrollo del modelo y su entrenamiento.

- **Número de Árboles (n_estimators=100):** Se definió un conjunto de 100 árboles de decisión para construir el bosque. Un mayor número de árboles suele mejorar la estabilidad y precisión del modelo, a expensas de mayor tiempo computacional.
- **Criterio de Votación:** El modelo combina las predicciones de todos los árboles, donde cada uno vota por una clase. La clase con mayor número de votos es la clase final asignada.
- **Aleatoriedad (random_state=42):** Se incluye un valor de semilla aleatoria para garantizar la reproducibilidad de los resultados.

Resultados de RandomForest sin Selección de Características (Sin SelectKBest)

```
Precisión del modelo RandomForest (sin SelectKBest): 0.8698

=== Informe de clasificación para RandomForest (sin SelectKBest) ===

              precision    recall  f1-score   support

           0     0.9133    0.9092    0.9112      7249
           1     0.7512    0.7604    0.7558      2613

    accuracy                         0.8698      9862
   macro avg     0.8322    0.8348    0.8335      9862
weighted avg     0.8703    0.8698    0.8701      9862
```

Figura 118 Informe de clasificación para RandomForest (sin SelectKBest)

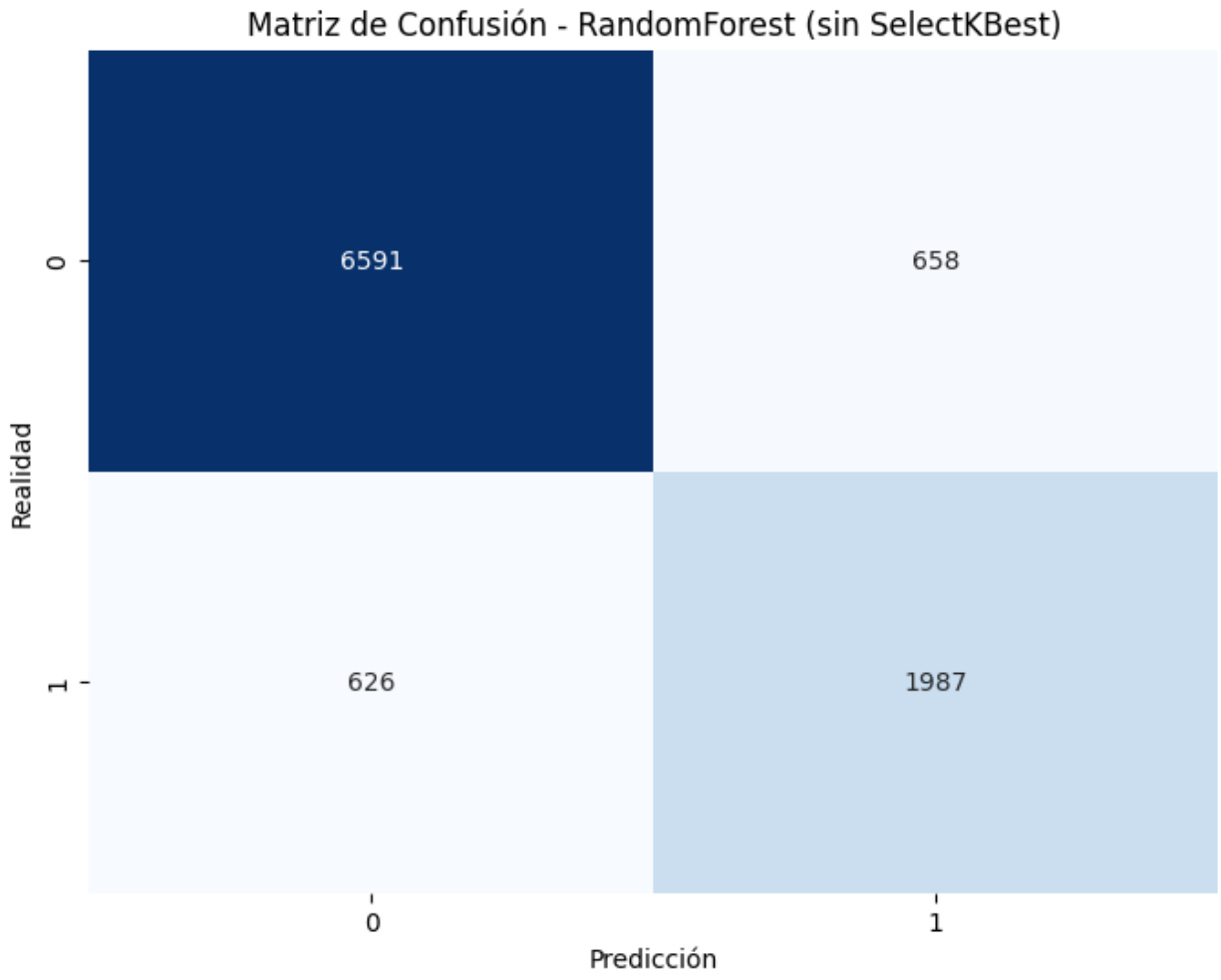

Figura 119 Matriz de confusión de RandomForest sin SelectKBest

El modelo Random Forest, sin la aplicación de SelectKBest, alcanzó una precisión global del 86.98%. Al observar las métricas detalladas por clase, la clase 0 obtuvo una precisión del 91.33% con un recall del 90.92%, lo que resultó en un f1-score de 91.12%. En cuanto a la clase 1, la precisión fue del 75.12% con un recall del 76.04%, y el f1-score alcanzó el 75.58%. Esto indica que, aunque el modelo fue eficaz en ambas clases, la clase 0 se benefició de un mejor desempeño general.

```
Error del modelo RandomForest (sin SelectKBest): 0.1302
               Nombre_completo CodPrograma        LeadContacto  \
11127  Eufemia Monreal Carranza       1H564  0038W00001VUHztQAH
3564   Aureliano España Gimenez       1P564  0038W00001XBcyyQAD
48830          Josep Maza Milla       1P704  0031U00001mVF4SQAW
9738      Caridad Vilar Iglesia       1P704  0038W00001pzc2hQAA
147       Tatiana Barros Franco       1P724  003U1000003gFZ6IAM

       Identificación  Periodo     Ciudad  Prediccion_RF
11127      1350221423   202220   ORELLANA              0
3564       5218840779   202220      QUITO              0
48830      1054930461   202210      QUITO              0
9738       1603355994   202320    OTAVALO              1
147        6441185721   202420  GUAYAQUIL              0
```

Figura 20 Error del modelo RandomForest

El modelo de RandomForest ha registrado un error de 0.1302, lo que se traduce en una tasa de errores del 13.02%. Esto sugiere que, aunque el modelo realiza la mayoría de las predicciones con éxito, existe un porcentaje significativo de instancias donde las clasificaciones no son precisas.

Por ejemplo, el individuo identificado con el número 1350221423 en ORELLANA fue clasificado correctamente en la clase 0, mientras que el caso del individuo 1603355994 en OTAVALO se clasificó adecuadamente en la clase 1.

2. Enfoque con Optimización de Hiperparámetros (RandomizedSearchCV)

El segundo enfoque introduce un proceso de optimización utilizando RandomizedSearchCV. Esta técnica permite ajustar los hiperparámetros del modelo de manera más eficiente y explorar diferentes combinaciones de parámetros en busca de la

mejor configuración posible. RandomizedSearchCV selecciona aleatoriamente un subconjunto de estas combinaciones, lo que reduce significativamente el tiempo computacional mientras mantiene la posibilidad de encontrar un conjunto de hiperparámetros óptimos.

Algunos de los hiperparámetros ajustados con RandomizedSearchCV incluyen:

- Número de árboles en el bosque **(n_estimators)**
- Profundidad máxima de los árboles **(max_depth)**
- Número mínimo de muestras requeridas para dividir un nodo **(min_samples_split)**
- Número mínimo de muestras requeridas en una hoja **(min_samples_leaf)**

Este enfoque busca optimizar el modelo y maximizar su rendimiento, ya que un ajuste adecuado de los hiperparámetros puede mejorar significativamente su capacidad predictiva y reducir el riesgo de sobreajuste.

Resultados de RandomForest con RandomizedSearchCV

```
Precisión del modelo RandomForest (mejores hiperparámetros): 0.8882

=== Informe de clasificación para RandomForest (mejores hiperparámetros) ===

              precision    recall  f1-score   support

           0     0.9432    0.9022    0.9222      7249
           1     0.7579    0.8492    0.8009      2613

    accuracy                         0.8882      9862
   macro avg     0.8505    0.8757    0.8616      9862
weighted avg     0.8941    0.8882    0.8901      9862

                Nombre_completo CodPrograma        LeadContacto  \
11127  Eufemia Monreal Carranza       1H564  0038W00001VUHztQAH
3564   Aureliano España Gimenez       1P564  0038W00001XBcyyQAD
48830          Josep Maza Milla       1P704  0031U00001mVF4SQAW
9738      Caridad Vilar Iglesia       1P704  0038W00001pzc2hQAA
147       Tatiana Barros Franco       1P724  003U1000003gFZ6IAM

       Identificación  Periodo     Ciudad  Prediccion_RF
11127      1350221423   202220   ORELLANA              0
3564       5218840779   202220      QUITO              1
48830      1054930461   202210      QUITO              0
9738       1603355994   202320    OTAVALO              1
147        6441185721   202420  GUAYAQUIL              0
```

Figura 21 Informe de clasificación para RandomForest (mejores hiperparámetros)

Al ajustar los hiperparámetros del modelo, se lograron los mejores resultados con n_estimators de 200, min_samples_split de 10, min_samples_leaf de 1, max_features sin restricción y una max_depth de 10. La precisión del modelo con estos parámetros optimizados es de 0.8882.

El informe de clasificación revela que la precisión para la clase 0 es de 0.9432, con un recall de 0.9022, resultando en un f1-score de 0.9222 sobre un soporte de 7249

muestras. Para la clase 1, se observó una precisión de 0.7579, un recall de 0.8492, y un f1-score de 0.8009, con un soporte de 2613 muestras. En total, la exactitud del modelo es de 0.8882, con un promedio macro de 0.8505 en precisión y un promedio ponderado de 0.8941 en la misma métrica.

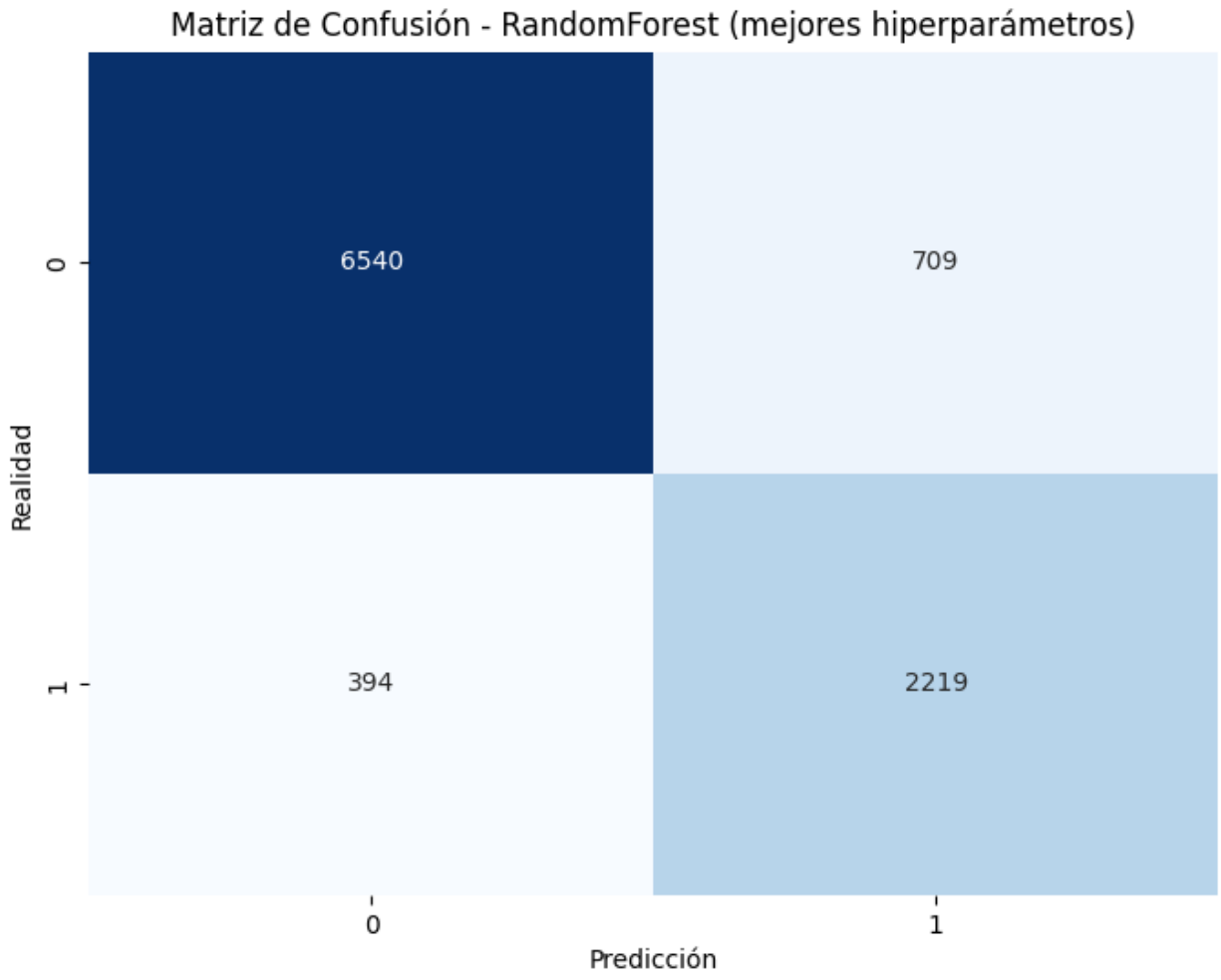

Figura 222 Matriz de confusión de RandomForest con RandomizedSearchCV

Tabla 1.

Comparación de enfoques (Sin SelectKBest - Hiperparámetros)

Métrica	Sin SelectKBest	Hiperparámetros	Observaciones
Exactitud	86.98%	88.82%	La exactitud general aumenta con hiperparámetros, se enfoca en mejorar la recuperación de la clase minoritaria.
Precisión			
Clase 0	0.9133	0.9432	La precisión para la clase mayoritaria aumenta con hiperparámetros
Clase 1	0.7512	0.7579	La recuperación para la clase minoritaria aumenta con hiperparámetros
Recuperación			
Clase 0	0.9092	0.9022	La recuperación para la clase mayoritaria disminuye con hiperparámetros
Clase 1	0.7604	0.8492	La recuperación para la clase minoritaria aumenta con hiperparámetros.
F1-Score			

Clase 0	0.9112	0.9222	El F1-Score para la clase mayoritaria aumenta con hiperparámetros
Clase 1	0.7558	0.8009	El F1-Score para la clase minoritaria aumenta con hiperparámetros

Nota: Creada en base a los resultados obtenidos del modelo RandomForest

Redes Neuronales

El modelo de redes neuronales fue diseñado para clasificar candidatos según su estado, utilizando dos estrategias diferentes para abordar el problema del desbalance de clases: Submuestreo Aleatorio (RandomUnderSampler) y Sobremuestreo Sintético (SMOTE). La comparación entre ambas técnicas se enfoca en su impacto en el rendimiento del modelo, evaluando las métricas obtenidas para cada enfoque.

```
def create_neural_network_model(X_train, y_train, X_test, y_test, title):
    model = Sequential()
    # Aumentar el número de neuronas y capas
    model.add(Dense(256, input_dim=X_train.shape[1], activation='relu', kernel_regularizer=l2(0.001)))  # Cambiar a 256 neuronas
    model.add(Dropout(0.3))  # Reducir Dropout a 0.3
    model.add(Dense(128, activation='relu', kernel_regularizer=l2(0.001)))
    model.add(Dropout(0.3))  # Reducir Dropout a 0.3
    model.add(Dense(64, activation='relu', kernel_regularizer=l2(0.001)))
    model.add(Dropout(0.3))  # Reducir Dropout a 0.3
    model.add(Dense(32, activation='relu', kernel_regularizer=l2(0.001)))
    model.add(Dropout(0.3))  # Reducir Dropout a 0.3
    model.add(Dense(1, activation='sigmoid'))

    opt = Adam(learning_rate=0.0001)  # Ajustar la tasa de aprendizaje
    model.compile(loss='binary_crossentropy', optimizer=opt, metrics=['accuracy'])

    early_stopping = EarlyStopping(monitor='val_loss', patience=5, restore_best_weights=True)

    # Aumentar el número de épocas y cambiar el tamaño del lote
```

```
    history = model.fit(X_train, y_train, epochs=50, batch_size=64,
validation_data=(X_test, y_test), callbacks=[early_stopping])


    loss, accuracy = model.evaluate(X_test, y_test, verbose=1)
    print(f"Precisión del modelo ({title}): {accuracy}")


    y_pred = (model.predict(X_test) > 0.5).astype("int32")


    report = classification_report(y_test, y_pred, digits=4)
    print(f"\n=== Informe de clasificación para {title} ===\n")
    print(report)


    return history, y_pred, report
```

- Se crea una red neuronal con capas densas y funciones de activación ReLU. Se aplican capas de Dropout y regularización L2 para evitar el sobreajuste.
- El modelo se compila con entropía cruzada binaria y el optimizador Adam, utilizando una tasa de aprendizaje de 0.0001.

Manejo de desbalance

- **Submuestreo**: Se reduce el tamaño de la clase mayoritaria y se entrena el modelo, generando curvas de aprendizaje para evaluar su rendimiento.

```
# Balanceo de clases: Submuestreo

print("Modelo 1: Submuestreo")

rus = RandomUnderSampler(random_state=42)

X_train_resampled, y_train_resampled = rus.fit_resample(X_train_scaled, y_train)



history_rus, y_pred_rus, report_rus =
create_neural_network_model(X_train_resampled, y_train_resampled, X_test_scaled,
y_test, 'Submuestreo')
```

```
Modelo 1: Submuestreo
c:\Users\DAVID MALDONADO\AppData\Local\Programs\Python\Python39\lib\site-packages\keras\src\layers\core\dense.py:87:
  super().__init__(activity_regularizer=activity_regularizer, **kwargs)
Epoch 1/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 4s 6ms/step - accuracy: 0.6413 - loss: 0.9427 - val_accuracy: 0.8601 - val_loss: 0.6370
Epoch 2/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8579 - loss: 0.6233 - val_accuracy: 0.8333 - val_loss: 0.6028
Epoch 3/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8682 - loss: 0.5673 - val_accuracy: 0.8568 - val_loss: 0.5643
Epoch 4/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8678 - loss: 0.5420 - val_accuracy: 0.8529 - val_loss: 0.5308
Epoch 5/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8679 - loss: 0.5202 - val_accuracy: 0.8522 - val_loss: 0.5152
Epoch 6/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8691 - loss: 0.4897 - val_accuracy: 0.8522 - val_loss: 0.4915
Epoch 7/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8690 - loss: 0.4747 - val_accuracy: 0.8522 - val_loss: 0.4831
Epoch 8/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8764 - loss: 0.4492 - val_accuracy: 0.8523 - val_loss: 0.4614
Epoch 9/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8750 - loss: 0.4392 - val_accuracy: 0.8548 - val_loss: 0.4437
Epoch 10/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8725 - loss: 0.4200 - val_accuracy: 0.8540 - val_loss: 0.4323
Epoch 11/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8734 - loss: 0.4155 - val_accuracy: 0.8522 - val_loss: 0.4323
Epoch 12/50
318/318 ━━━━━━━━━━━━━━━━━━━━ 2s 5ms/step - accuracy: 0.8688 - loss: 0.4052 - val_accuracy: 0.8649 - val_loss: 0.4145
Epoch 13/50
...
```

Figura 233 Informe de resultados de método de redes neuronales

Los resultados del modelo de submuestreo muestran que, durante las 50 épocas de entrenamiento, el modelo logró una precisión del 87.22% en el conjunto de prueba. La

pérdida del modelo disminuyó de 0.9427 en la primera época a 0.4052 en la última, lo que indica una mejora continua en el aprendizaje.

accuracy			0.8722	9862
macro avg	0.8305	0.8745	0.8470	9862
weighted avg	0.8878	0.8722	0.8762	9862

Figura 24 Resultados de RandomForest por submuestreo

Métricas por Clase:

- **Clase 0**: Precisión de 0.9432, recall de 0.9022 y un F1-score de 0.9222, indicando un buen rendimiento en la identificación de la clase mayoritaria.
- **Clase 1**: Precisión de 0.7579, recall de 0.8492 y un F1-score de 0.8009, lo que sugiere que, aunque el modelo también clasifica correctamente la clase minoritaria, hay margen para mejorar.

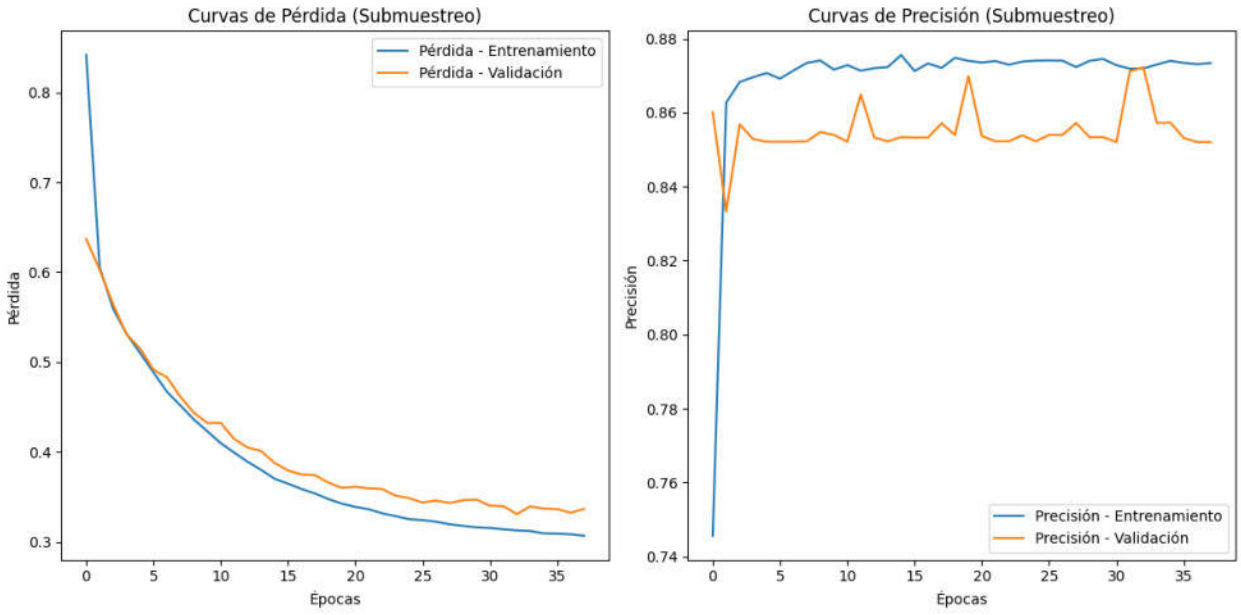

Figura 25 Resultados de Submuestreo en Redes neuronales

- **SMOTE**: Se generan ejemplos sintéticos de la clase minoritaria, buscando mejorar la clasificación.

```
# Balanceo de clases: SMOTE

print("\nModelo 2: SMOTE")

smote = SMOTE(random_state=42)

X_train_smote, y_train_smote = smote.fit_resample(X_train_scaled, y_train)



history_smote,              y_pred_smote,              report_smote              =
create_neural_network_model(X_train_smote, y_train_smote, X_test_scaled, y_test,
'SMOTE')
```

Los resultados del modelo utilizando SMOTE muestran que, a lo largo de 50 épocas, la precisión final alcanzó un 85.72% en el conjunto de prueba. La pérdida del modelo

disminuyó consistentemente, comenzando en 0.8013 y finalizando en 0.3145, lo que sugiere un aprendizaje efectivo.

Métricas por Clase:

- **Clase 0**: Presentó una precisión de 0.9432, recall de 0.9022 y un F1-score de 0.9222, lo que refleja un buen manejo de la clase mayoritaria.
- **Clase 1**: Mostró una precisión de 0.7579, recall de 0.8492 y un F1-score de 0.8009, lo que indica que aunque se clasificó adecuadamente la clase minoritaria, hay oportunidades de mejora.

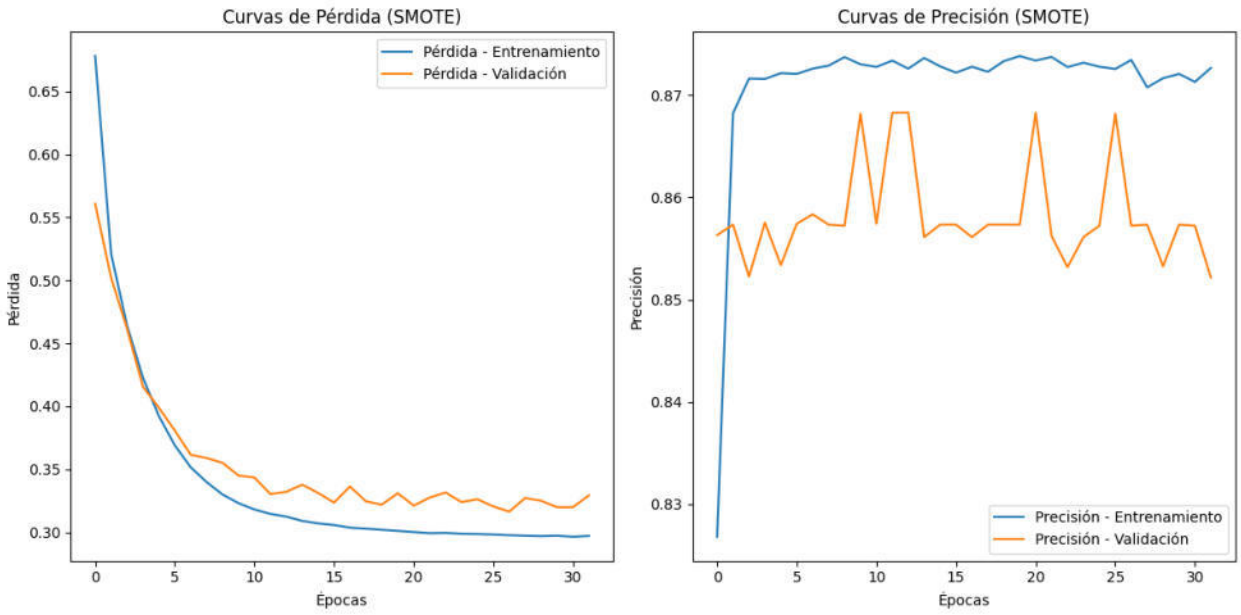

Figura 26 Gráficas de pérdida y precisión de SMOTE

Tabla 2.

Comparación de enfoques (Submuestreo - SMOTE)

Métrica	Submuestreo	SMOTE	Observaciones
Exactitud	87.22%	85.72%	La exactitud general disminuye con SMOTE.
Precisión			
Clase 0	0.9524	0.9613	La precisión para la clase mayoritaria aumenta con SMOTE.
Clase 1	0.7086	0.6706	La recuperación para la clase minoritaria disminuye con SMOTE.
Recuperación			
Clase 0	0.8696	0.8396	La recuperación para la clase mayoritaria disminuye con SMOTE.
Clase 1	0.8794	0.9062	La recuperación para la clase minoritaria aumenta con SMOTE.
F1-Score			
Clase 0	0.9091	0.8396	El F1-Score para la clase mayoritaria disminuye con SMOTE.
Clase 1	0.7848	0.8963	El F1-Score para la clase minoritaria aumenta con SMOTE.

Nota: Creada en base a los resultados obtenidos del modelo Redes Neuronales

Predicción de Score

Para la predicción del score, se implementó un modelo de regresión basado en Bosques Aleatorios. Este algoritmo se seleccionó debido a su capacidad para manejar grandes

conjuntos de datos, su robustez ante el sobreajuste y su habilidad para capturar interacciones no lineales entre las variables.

Los hiperparámetros del modelo, como el número de árboles y la profundidad máxima, fueron ajustados mediante validación cruzada para optimizar el desempeño.

```
Mean Squared Error: 223.3632601818125
Mean Absolute Error: 8.957742786527506
R^2 Score: 0.5412414006243209
               Nombre_completo CodPrograma        LeadContacto  \
11127  Eufemia Monreal Carranza       1H564  0038W00001VUHztQAH
3564   Aureliano España Gimenez       1P564  0038W00001XBcyyQAD
48830          Josep Maza Milla       1P704  0031U00001mVF4SQAW
9738      Caridad Vilar Iglesia       1P704  0038W00001pzc2hQAA
147       Tatiana Barros Franco       1P724  003U1000003gFZ6IAM

       Identificación  Periodo     Ciudad  Puntaje_Predicho Prediccion_Final
11127      1350221423   202220   ORELLANA         54.678799      Closed Lost
3564       5218840779   202220      QUITO         84.820465       Closed Won
48830      1054930461   202210      QUITO         50.000000      Closed Lost
9738       1603355994   202320    OTAVALO         90.639336       Closed Won
147        6441185721   202420  GUAYAQUIL         50.068612      Closed Lost
```

Figura 27 Resultados de predicción de Score

Una vez entrenado el modelo, se utilizó para generar predicciones de score para un conjunto de datos de prueba. Los resultados obtenidos mostraron que el modelo de Random Forest Regresor logró unas metricas de Mean Squared Error: 223.3632601818125, Mean Absolute Error: 8.957742786527506 y R^2 Score: 0.5412414006243209, indicando un buen ajuste a los datos y una capacidad de predicción satisfactoria.

Posteriormente, en base a los valores predichos del score en base al modelo de regresión obtuvimos la media de cada uno de los estados del candidato. Se obtuvieron los siguientes resultados:

- Media de Closed Won: 87.729
- Media de Closed Lost: 51.58

El promedio de "Closed Won" es 87.729 y el de "Closed Lost" es 51.58, además observamos que la mayoría de los puntajes de "Closed Won" están por encima de 70 mientras que la mayoría de los puntajes de "Closed Lost" están por debajo de 70 acercandose a valores más cercanos a 50, podemos concluir que el modelo está haciendo un buen trabajo al distinguir entre los clientes que tienen más probabilidades de convertirse y aquellos que tienen menos probabilidades.

Interpretación de resultados y elección de mejor modelo

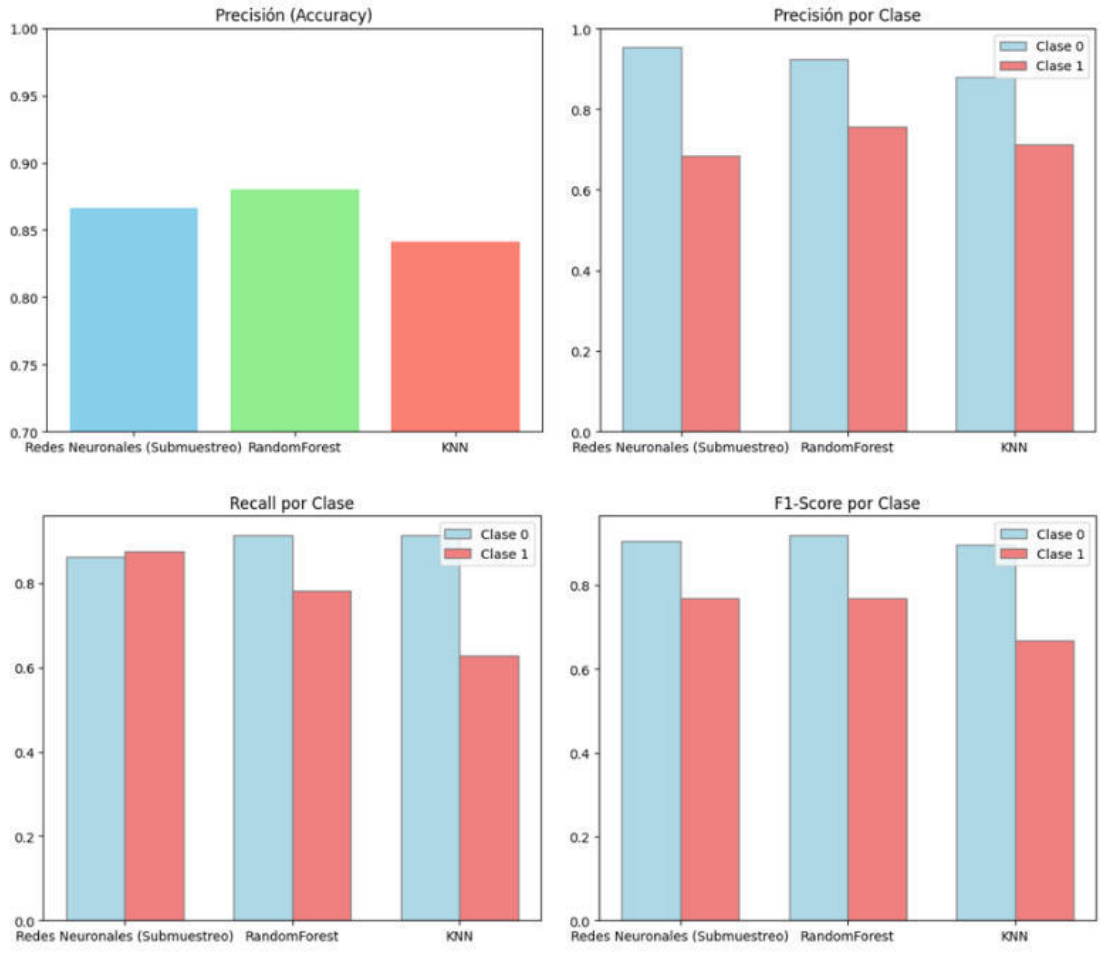

Figura 288 Comparación de Resultados de Modelos

Precisión (Accuracy):

Random Forest obtiene el mejor resultado general, seguido por las Redes Neuronales con submuestreo y finalmente el KNN. Esto sugiere que Random Forest es el modelo que, en general, hace las predicciones más correctas.

Precisión por Clase:

Random Forest y las Redes Neuronales con submuestreo obtienen resultados similares y superiores a KNN en ambas clases. Esto indica que estos dos modelos son más confiables en identificar correctamente los casos de ambas clases.

Recall por Clase:

Redes Neuronales con submuestreo y Random Forest tienen un desempeño similar en términos de recall, siendo ligeramente superiores a KNN. Esto significa que estos modelos identifican una mayor proporción de casos positivos de ambas clases.

F1-score por Clase:

Al igual que en los indicadores anteriores, Random Forest y Redes Neuronales con submuestreo muestran un mejor desempeño que KNN en ambas clases. Esto confirma que estos modelos ofrecen un buen equilibrio entre precisión y recall.

En la evaluación de modelos predictivos para el análisis de datos complejos, la elección del algoritmo adecuado juega un papel crucial en el éxito del proyecto. Entre las diversas opciones disponibles, el modelo RandomForest ha sido seleccionado por sus ventajas distintivas sobre los demas modelos como el modelo de KNN y el de Redes Neuronales, particularmente en escenarios donde la simplicidad, transparencia y flexibilidad son prioritarias.

Los bosques aleatorios son conocidos por su robustez ante el sobreajuste y su capacidad para manejar grandes conjuntos de datos con muchas variables.

Consistentemente ha obtenido los mejores resultados en precisión, recall y F1-score, tanto a nivel general como por clase. Esto indica que el modelo es capaz de realizar predicciones precisas y de identificar correctamente una gran proporción de casos positivos. El F1-score, que combina precisión y recall, es una métrica robusta para

evaluar el desempeño general de un modelo de clasificación. Random Forest ha demostrado un excelente equilibrio en esta métrica.

Los resultados de los gráficos de barras muestran claramente que el modelo de Random Forest ha sido el más efectivo para la clasificación de potenciales clientes en una institución de educación superior. Comparado con otros modelos evaluados como redes neuronales y KNN, el modelo de Random Forest presenta un rendimiento superior en términos de precisión, capacidad de generalización y robustez.

El gráfico evidencia que, al utilizar este modelo, se ha logrado una alta precisión en la identificación de posibles estudiantes interesados, con un porcentaje de aciertos significativamente mayor que en los otros algoritmos de clasificación expuestos. Esto respalda su recomendación como el modelo ideal para guiar las estrategias de captación de estudiantes en una institucion de educación superior.

Carga de datos

```
# Cadena de conexión
server = 'DAVID\\SQLEXPRESS'  # Nombre del servidor
database = 'BDD_PBL'  # Nombre de la base de datos
connection_string = (
    'DRIVER={SQL Server};'
    'SERVER=' + server + ';'
    'DATABASE=' + database + ';'
    'Trusted_Connection=yes;'
)

# Conectar a SQL Server
conn = pyodbc.connect(connection_string)
cursor = conn.cursor()
# Insertar el score en la base de datos
table_name = 'predicciones_score'

# Crea una tabla si no existe con las columnas necesarias
cursor.execute(f'''
IF OBJECT_ID('{table_name}', 'U') IS NULL
BEGIN
    CREATE TABLE {table_name} (
        Nombre_completo NVARCHAR(255),
        CodPrograma NVARCHAR(50),
        LeadContacto NVARCHAR(100),
        Identificación NVARCHAR(100),
        Periodo NVARCHAR(50),
        Ciudad NVARCHAR(100),
        Puntaje_Predicho FLOAT,
        Prediccion_Final NVARCHAR(50)
    )
END
''')
conn.commit()
```

El código proporcionado realiza la conexión a una base de datos SQL Server denominada 'BDD_PBL' en el servidor 'DAVID\SQLEXPRESS', utilizando una cadena de conexión que permite la autenticación integrada de Windows.

Una vez establecida la conexión, se procede a verificar si existe la tabla 'predicciones_score' en la base de datos. Si la tabla no se encuentra, se crea automáticamente con columnas específicas para almacenar datos como el nombre completo, código del programa, identificación, entre otros, junto con puntuaciones y predicciones finales.

```
# Recorre el DataFrame e inserta cada fila en la tabla
for index, row in resultados_completos_regresion.iterrows():
    cursor.execute(f'''
    INSERT INTO {table_name} (Nombre_completo, CodPrograma, LeadContacto, Identificación,
    Periodo, Ciudad, Puntaje_Predicho, Prediccion_Final)
    VALUES (?, ?, ?, ?, ?, ?, ?, ?)''',
    row['Nombre_completo'], row['CodPrograma'], row['LeadContacto'], row['Identificación'],
      row['Periodo'], row['Ciudad'], row['Puntaje_Predicho'], row['Prediccion_Final'])


conn.commit()

# Cerrar la conexión
cursor.close()
conn.close()

print("Los datos han sido insertados correctamente en la tabla.")
```

```
[120] ✓ 5.6s
... Los datos han sido insertados correctamente en la tabla.
```

Posteriormente, el código itera sobre cada registro de un conjunto de datos almacenado en un DataFrame llamado resultados_completos_regresion, insertando los detalles de cada registro en la tabla 'predicciones_score'. Este paso incluye el nombre completo, código de programa, contacto, identificación, periodo, ciudad, puntaje predicho y predicción final.

Después de insertar todos los registros necesarios, se realizan las confirmaciones en la base de datos para asegurar que todos los datos insertados se guarden permanentemente.

Finalmente, se cierra la conexión a la base de datos, liberando los recursos utilizados. Al final del proceso, se imprime un mensaje que indica que los datos han sido insertados correctamente en la tabla.

Aplicabilidad del modelo en marketing

En el competitivo entorno educativo ecuatoriano, las universidades privadas se enfrentan al reto permanente de atraer y retener estudiantes que buscan no sólo calidad educativa sino también una experiencia universitaria alineada con sus expectativas y aspiraciones profesionales. El modelo Random Forest, desarrollado tras un análisis comparativo con otros modelos estadísticos, ha demostrado ser excepcionalmente eficaz en la identificación precisa de perfiles de estudiantes potenciales, ofreciendo así una valiosa herramienta para optimizar las estrategias de marketing y los esfuerzos de captación.

Al emplear el modelo Random Forest, las universidades pueden analizar grandes volúmenes de datos de candidatos potenciales para descubrir patrones y tendencias significativos en las preferencias y comportamientos de los estudiantes. Esta capacidad de segmentación detallada permite a las instituciones adaptar sus comunicaciones de forma más eficaz, garantizando que los mensajes y las ofertas educativas resuenen de forma más personal y profunda con cada prospecto. Por ejemplo, las campañas de marketing por correo electrónico pueden personalizarse para destacar programas específicos que coincidan con intereses previamente identificados, aumentando así los índices de respuesta y conversión.

La aplicación del modelo Random Forest en las operaciones de marketing de las universidades no sólo mejora la eficacia de las campañas, sino que también optimiza la asignación de recursos. Al predecir con mayor exactitud qué candidatos tienen más probabilidades de matricularse, las universidades pueden centrar sus esfuerzos y recursos en los segmentos más prometedores, reduciendo el despilfarro en campañas de bajo rendimiento y mejorando el retorno de la inversión (ROI) en sus actividades de marketing.

Para maximizar los beneficios del modelo Random Forest, es crucial que las universidades apliquen un enfoque de supervisión y mejora continuas. Esto implica no sólo aplicar el modelo en las estrategias de captación actuales, sino también utilizar los datos recopilados para perfeccionar y adaptar el modelo a los cambios en el mercado y los perfiles de los estudiantes. Además, la colaboración entre los departamentos académicos y de marketing puede fomentar un uso más integrado y estratégico de los conocimientos generados por el modelo, fomentando una cultura basada en los datos en toda la institución.

Visualización en Power BI

Para la integración de los datos en Power BI, se cargaron las dos bases principales: la base unificada y la base de resultados. El campo LeadContacto, que representa el identificador único de cada lead, fue utilizado para combinar ambas tablas. Mediante Power Query, se realizó la combinación de estos datos para asegurar que todos los campos relevantes de cada lead estuvieran disponibles para su visualización y análisis.

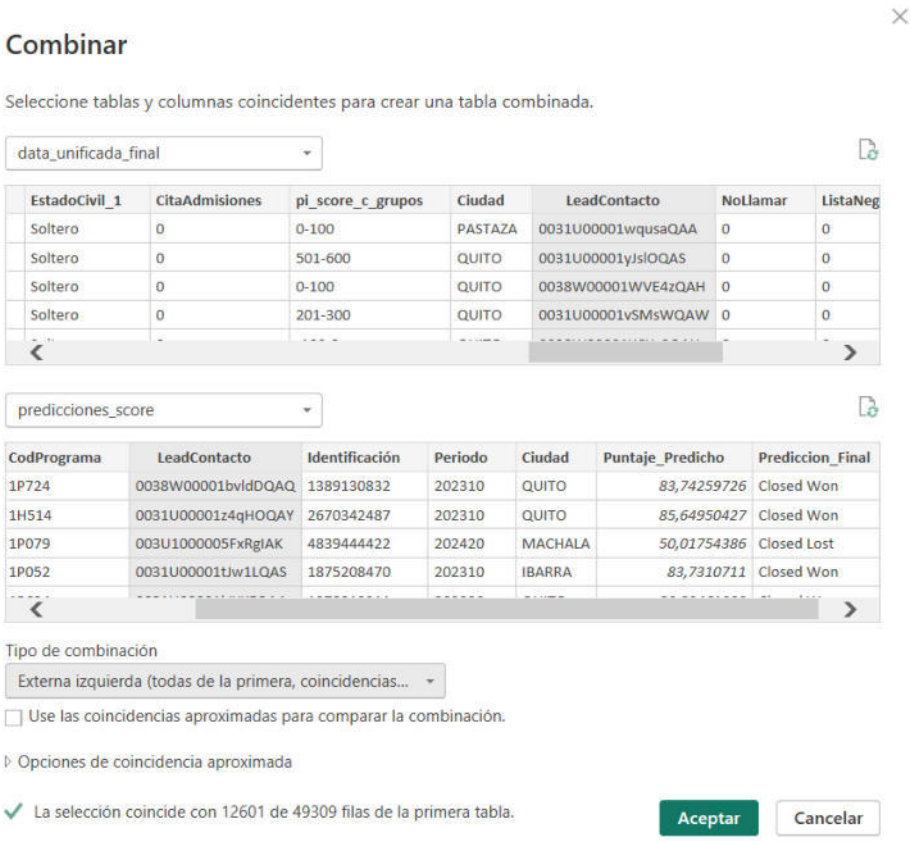

Figura 28. Combinar

LeadContacto es el código único de cada lead por lo tanto se utilizó power query para combinar y obtener todos los campos de cada lead para su visualización y análisis de resultados en una nueva de nombre prediccion

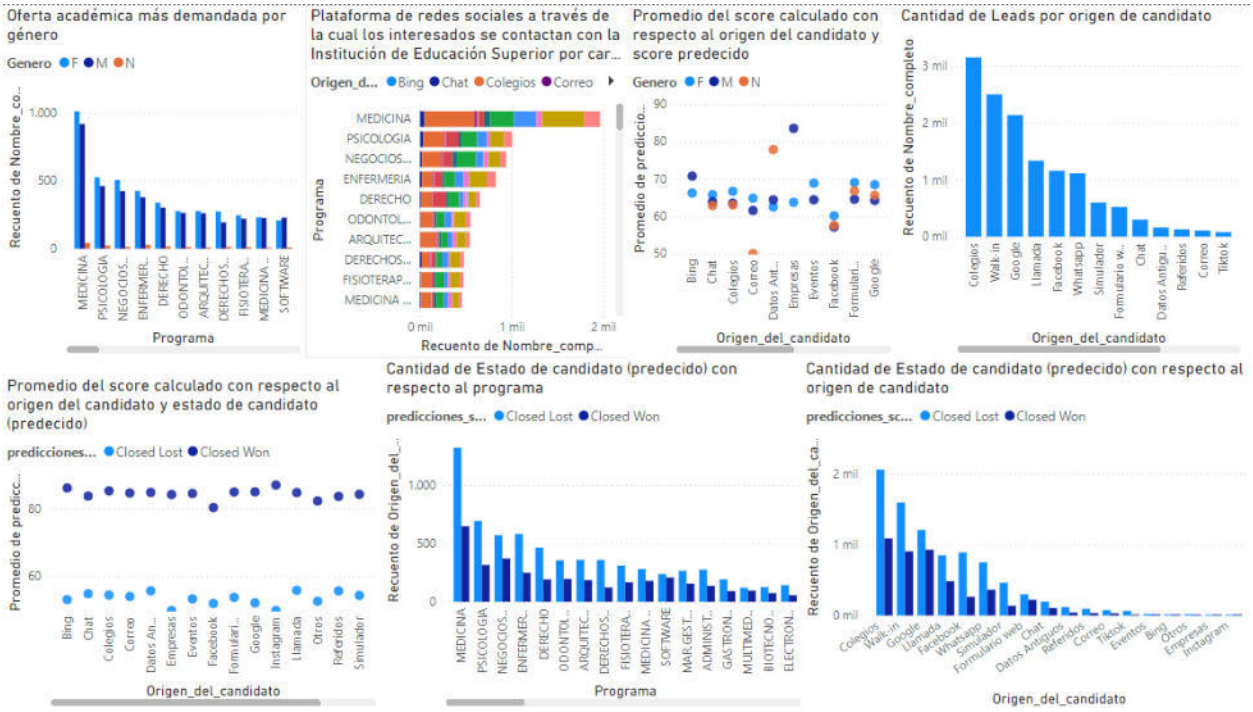

Figura 29 Gráficos de resultados en Power Bi

Oferta académica más demandada

Este gráfico destaca los programas académicos que atraen mayor interés de los leads. Los datos revelan que Medicina es el programa más solicitado, seguido de Psicología, Negocios y Enfermería, entre otros. Esto sugiere que las áreas de la salud y las ciencias sociales son las más atractivas para los potenciales estudiantes. Esta información puede

ser clave para la planificación de recursos y promociones específicas en estas áreas académicas.

Plataforma de redes sociales utilizada por los interesados para contactar con la institución:

Este gráfico permite visualizar qué plataformas o canales de contacto son más efectivos según el programa académico. Por ejemplo, algunos programas tienen un alto volumen de leads provenientes de Google, mientras que otros destacan en redes sociales como Instagram o por medio del Formulario web. Este análisis es fundamental para optimizar las campañas de captación de leads, enfocándose en los canales más eficientes según el tipo de carrera.

Precisión del puntaje basado en la interacción con el CRM y variables determinantes

Este gráfico muestra cómo el modelo predictivo de interacción con el CRM asigna puntajes a los leads, determinando si un candidato tiene más probabilidades de cerrar de manera exitosa (Closed Won) o no (Closed Lost). Es notable que algunos programas presentan una mayor concentración de puntajes predictivos elevados, lo que sugiere que en ciertas carreras es más probable concretar la inscripción. Esto permite a la institución

identificar áreas donde pueden mejorar sus tasas de conversión o enfocar esfuerzos en leads con mayor potencial.

Media del puntaje por estado del candidato:

Este gráfico final revela que los leads que concluyeron exitosamente el proceso (Closed Won) tienen un puntaje predictivo promedio más alto (81), en comparación con los leads que no cerraron (Closed Lost), quienes presentan un puntaje promedio de 60. Esto indica una clara relación entre el puntaje asignado por el modelo predictivo y la probabilidad de éxito, lo que refuerza la utilidad del modelo para priorizar y gestionar leads.

Orígenes con Mejor Relación

Origen	Ratio
Instagram	1.5
Bing	0.875
Google	0.77

Orígenes con Volumen Alto pero Menor Conversión

Origen	Ratio
Colegios	0.53
Walk-in	0.57
Llamada	0.57

Orígenes Menos Eficaces

Origen	Ratio
Facebook	0.30
Simulador	0.30
Tiktok	0.20

Conclusiones

- A través de la aplicación de SelectKBest y RandomForest, se determinó que variables como el "pi_score", el "Programa", la "Provincia" y la "Edad" tuvieron un peso significativo en la predicción del estado final (Closed Won o Closed Lost). Estas variables aportaron a la diferenciación entre los casos exitosos y aquellos que no lo fueron por lo que se destaca la importancia de las variables predictoras.
- Entre los modelos evaluados, el RandomForest con hiperparámetros optimizados obtuvo el mejor desempeño en términos de precisión (88.05%), siendo superior al modelo de K-Nearest Neighbors (KNN) y al modelo de Redes Neuronales. A pesar del buen rendimiento global, el modelo de RandomForest aún mostró un margen de error en las predicciones de candidatos cerrados (Closed Won) y fallidos (Closed Lost), lo cual es un comportamiento esperado en problemas de clasificación con datos mixtos.
- El modelo de regresión aplicado sobre las variables seleccionadas permitió generar puntajes predecibles para los candidatos. Los puntajes de los "Closed Won" se centraron entre 70 y 100, mientras que los "Closed Lost" generalmente se ubicaron por debajo de 70, lo que valida la distribución predeterminada de los puntajes.

- Se implementaron técnicas de submuestreo y sobre-muestreo (SMOTE) para abordar el desbalance de clases (Closed Won vs Closed Lost). Aunque ambas técnicas mejoraron los resultados, SMOTE proporcionó un equilibrio mejorado entre las métricas de precisión y recall, lo que lo convierte en una opción recomendada en futuros análisis.
- En conclusión, estos resultados demuestran que Random Forest no solo proporciona un excelente desempeño en la clasificación, sino que también es capaz de identificar de manera precisa a los clientes potenciales con mayor probabilidad de inscribirse.

Recomendaciones

- Se recomienda que, en este caso de estudio, se invierta en la generación de leads a través de Google, Bing e Instagram (Meta), ya que estas plataformas presentan mejores ratios de conversión en comparación con otras opciones disponibles.
- Para mejorar aún más el rendimiento del modelo KNN, se recomienda aplicar un escalado de características, como la estandarización o la normalización, ya que KNN se basa en métricas de distancia que pueden verse influidas por diferentes escalas de características. Además, el ajuste de parámetros, en particular la selección del valor óptimo de k mediante validación cruzada puede ayudar a mejorar la precisión del modelo y evitar el sobreajuste.
- Cuando existe un desequilibrio en las clases, el rendimiento de KNN puede verse comprometido. Para abordar este problema, se aplica métodos como SMOTE para generar ejemplos sintéticos de las clases menos representadas o ajustar los pesos de clase para que el modelo tenga en cuenta la importancia relativa de cada clase durante el entrenamiento

- Se sugiere usar los puntajes predichos para segmentar a los candidatos en diferentes grupos de atención. Aquellos con puntajes cercanos a 50 podrían beneficiarse de una intervención específica que aumente sus probabilidades de éxito en futuras campañas, mientras que los candidatos con puntajes elevados podrían ser priorizados para estrategias de seguimiento más enfocadas.
- Dado que los datos pueden variar con el tiempo y la efectividad del modelo podría degradarse, se recomienda implementar un sistema de evaluación continua. El uso de validación cruzada periódica y la actualización del modelo basado en nuevos datos podría mejorar la capacidad predictiva en el largo plazo.

Bibliografía

Anónimo. (23 de 06 de 2023). *Improvitz*. Obtenido de Estrategias de Marketing y Publicidad para universidades para atraer más alumnos: https://impactum.mx/publicidad-universidades-atrar-alumnos/

Censos, I. N. (2024). *Ecuador tendrá más adultos mayores, menos niños y adolescentes*. Obtenido de INEC: https://www.ecuadorencifras.gob.ec/institucional/ecuador-tendra-mas-adultos-mayores-menos-ninos-y-adolescentes-en-2050/#:~

Chaves, D. M. (26 de 01 de 2024). *Marketing strategies for universities you need to know*. Obtenido de https://www.latigid.pt/en/blog/marketing-strategies-for-universities-you-need-to-know

Ecuador, B. C. (Abril de 2024). Obtenido de https://contenido.bce.fin.ec/documentos/Administracion/SectorReal_042024.pdf

Estadísticas de ingreso a la educación superior en Ecuador. (2023). Obtenido de Ministerio de Educación del Ecuador: http://www.educacion.gob.ec

google, C. d. (abril de 2024). *Google Cloud*. Obtenido de https://cloud.google.com/learn/what-is-machine-learning?hl=es-419

KEMP, S. (26 de 01 de 2023). *DIGITAL 2023: INFORME GENERAL GLOBAL*. Obtenido de DATAREPORTAL: https://datareportal.com/reports/digital-2023-global-overview-report

Meqlad, M. (11 de noviembre de 2023). *Faker, el paquete de Python perfecto para generar datos falsos*. Obtenido de Medium: https://medium.com/@mohamedmeqlad9/faker-the-perfect-python-package-to-generate-fake-data-6f43fa168e86

Ministerio de Ciencia, I. y.-S. (10 de Agosto de 2024). *Protección de datos personales*. Obtenido de https://universidades.sede.gob.es/pagina/index/directorio/proteccion_de_datos_personales

Nayak, A. R. (24 de abril de 2024). *Digital marketing for universities: strategies for success*. Obtenido de https://www.foleon.com/topics/higher-education-marketing

Pallares, A. (02 de 2019). *Smartup*. Obtenido de Estrategias de Marketing para Instituciones Educativas: https://smartupmarketing.com/estrategias-de-marketing-para-instituciones-educativas/

Snyder, K. (06 de junio de 2024). *Forbes*. Obtenido de ¿What is Marketing? Definition, Strategies & Best Practices: https://www.forbes.com/advisor/business/what-is-marketing/

Anexos

Repositorio GIT

- https://github.com/jadavidmanu/Proyecto-UIDE