MÁSTER EN BUSINESS ANALYTICS

Estrategia predictiva y Data-Driven para aumentar la Captación y Conversión de Leads en la Industria Farmacéutica en la escuela Pharma Business Education

Presentado por:

ESCUDERO NIETO, SERGIO JOEL

GAINZA JIMÉNEZ, ADRIANA VANESSA

MARTÍNEZ BRAGADO, INÉS MARIA

MÉNDEZ FIGUERA, NELSON SIMÓN

Dirigido por:

CASTILLO FAULI, VICENTE

YESTE MORENO, VICTOR MANUEL

CURSO ACADÉMICO 2024-2025

Resumen

La transformación digital y la incorporación de la inteligencia artificial en procesos organizativos han supuesto una revolución en los modelos operativos de las instituciones de educación. Concretamente, las escuelas de negocio enfrentan el desafío de integrar tecnologías predictivas para mejorar su rendimiento operativo y tomar decisiones estratégicas basadas en datos. Esta investigación se desarrolla en el marco de la institución Pharma Business Education, la primera escuela de negocio 100% online en España dedicada a facilitar el acceso de perfiles científicos a la industria farmacéutica, biotecnológica y de tecnología sanitaria.

Desde su fundación en 2022, Pharma Business Education ha experimentado un crecimiento sostenido en su comunidad de alumnos, alcanzando más de 800 matriculados y consolidando más de 155 colaboraciones con empresas del sector. Este avance ha generado nuevas necesidades de sistematización y optimización de procesos clave, en este trabajo se priorizará el comportamiento de los leads comerciales en el embudo de ventas para garantizar la eficiencia del equipo de admisión.

Para dar respuesta a esta necesidad, este trabajo propone el diseño e implementación de un modelo predictivo basado en técnicas de aprendizaje automático. Se enfocará por tanto en el área comercial (B2C), donde se pretende automatizar el sistema de priorización de leads en función del nivel de cualificación, canal de entrada, interacción digital y antecedentes de conversión.

Se emplearán, por tanto, en etapa descriptiva el modelo D.A.S.E (Descriptive Analytics Structured Exploration), en etapa predictiva, P.R.E.M.I.A. (Predictive & Machine Learning Integration Approach) y finalmente emplearemos Metodología O.P.T.I.M.A. (Optimization & Prescriptive Techniques for Intelligence Management & Action) por su enfoque integral y su reconocida aplicabilidad en proyectos analíticos orientados a negocio. El uso de Python y sus bibliotecas especializadas permitirá una implementación técnica robusta, compatible con las herramientas digitales actuales de la institución.

Los resultados preliminares identifican que más del 60 % de los leads se estanca en fases intermedias del funnel comercial debido a la falta de seguimiento y priorización adecuada. Esta situación representa una pérdida directa de oportunidades de conversión y una sobrecarga operativa para el equipo.

Este trabajo busca generar una solución estratégica basada en datos, orientada a la mejora continua de la eficiencia operativa institucional. La propuesta es extrapolable a otras escuelas de negocio especializadas y representa un avance hacia una gestión educativa más inteligente, sostenible y orientada al valor.

Palabras claves: captación, conversión, educación, farmacéutica, alumnos, funnel.

Abstract

The digital transformation and the integration of artificial intelligence in organizational processes have reshaped operational models in higher education institutions. In particular, specialized business schools face the challenge of incorporating predictive technologies to enhance operational performance and drive data-informed decision-making. This research takes place within Pharma Business Education, the first fully online business school in Spain focused on facilitating the professional entry of science graduates into the pharmaceutical, biotechnology, and health technology industries.

Since its establishment in 2022, Pharma Business Education has experienced significant growth in both its student community—exceeding 700 enrollments—and in its network of over 140 partner companies. This growth has exposed two critical process gaps: (1) the conversion of internship offers into effective student placements, and (2) the prioritization of commercial leads within the admissions funnel to ensure efficient resource allocation.

To address these challenges, this project proposes the design and deployment of two predictive models using machine learning techniques. The model, applied to the commercial admissions area (B2C), seeks to automate lead prioritization based on qualification level, acquisition channel, engagement behavior, and historical conversion data.

Methodology chosen descriptive stage the D.A.S.E model (Descriptive Analytics Structured Exploration), in predictive stage, P.R.E.M.I.A. (Predictive integration approach and machine learning) and finally we will use O.P.T.I.M.A. Methodology. (Optimization and Prescriptive Techniques for Intelligence Management and Action) for its comprehensive approach and its recognized applicability in business-oriented analytical projects.

The models will be implemented using Python and its ecosystem of data analytics libraries, fully compatible with the institution's existing digital infrastructure. Preliminary findings show that the strategic classification of internship offers has already led to a significant increase in conversion rates—from 15% to 30% within one year. Additionally, more than 60% of leads stall in intermediate funnel stages due to insufficient prioritization, indicating a pressing need for a data-driven approach.

This project aims not only to improve internal efficiency but also to establish a replicable model for predictive analytics implementation in specialized education, contributing to a more intelligent and value-oriented educational management framework.

***Keywords**: acquisition, conversion, education, pharmaceutical, internship, students, offers, companies.*

Índice de contenido del TFM

1. Introducción

1.1 Contextualización y propósito del trabajo

En los últimos años, la transformación digital y el desarrollo de la inteligencia artificial han tenido un impacto disruptivo en múltiples sectores, incluyendo el ámbito educativo y formativo. Las instituciones de enseñanza, especialmente aquellas vinculadas al desarrollo profesional, se ven obligadas a repensar sus procesos operativos, comerciales y académicos desde una perspectiva basada en datos. En este contexto, la integración de herramientas de analítica avanzada permite a las organizaciones educativas ser más eficientes, ofrecer experiencias más personalizadas y optimizar sus recursos.

Pharma Business Education nace en el año 2022 con una misión clara: facilitar el acceso de titulados universitarios en ciencias de la salud a la industria farmacéutica, biotecnológica y de tecnología sanitaria. Como primera escuela de negocio 100% online especializada en este ámbito en España, su propuesta se basa en la combinación de formación de calidad, fuerte orientación práctica y una estrecha colaboración con empresas del sector, asegurando que los alumnos que accedan a su formación 100% conseguirán oportunidades en el sector. Actualmente, la institución ha formado a más de 800 alumnos, ha establecido más de 155 convenios con empresas colaboradoras y ha desarrollado un ecosistema digital centrado en la empleabilidad y el desarrollo profesional.

Este trabajo surge ante la necesidad estratégica de optimizar un área fundamental del modelo de negocio de la institución: el embudo de conversión de leads comerciales (gestionado por el área de marketing y ventas o B2C), donde la clasificación y priorización de los contactos resulta clave para lograr una mayor eficiencia comercial.

La finalidad de este Trabajo de Fin de Máster es aplicar técnicas de analítica predictiva y machine learning para desarrollar dos modelos que permitan mejorar los indicadores clave de ambas áreas. Se emplearán, por tanto, en etapa descriptiva el modelo D.A.S.E (Descriptive Analytics Structured Exploration), en predictiva y Machine Learning P.R.E.M.I.A. Predictive & Machine Learning Integration Approach) y finalmente emplearemos Metodología O.P.T.I.M.A. (Optimization & Prescriptive Techniques for

Intelligence Management & Action) por su enfoque integral y su reconocida aplicabilidad en proyectos analíticos orientados a negocio.

1.2 Justificación

La creciente complejidad de los procesos de captación y conversión en organizaciones formativas exige soluciones innovadoras, ágiles y basadas en datos. En el caso de Pharma Business Education la estructura organizativa donde nos hemos focalizado a la hora de desarrollar este trabajo es en el área operativa: Marketing y ventas (B2C), orientada a la captación y conversión de leads en alumnos matriculados en los programas formativos.

Actualmente, se observa que en el área comercial gestiona un volumen elevado de leads, muchos de los cuales no se encuentran lo suficientemente cualificados o no avanzan por el embudo de conversión debido a la ausencia de un sistema claro de priorización y scoring predictivo.

En este escenario, se hace evidente la necesidad de incorporar herramientas que permitan extraer valor de los datos históricos disponibles. La literatura especializada (Davenport & Harris, 2007; McAfee & Brynjolfsson, 2012) señala que las organizaciones orientadas por datos —conocidas como Data-Driven— son significativamente más eficaces en la toma de decisiones estratégicas. En el ámbito educativo, informes recientes destacan el uso creciente de modelos de predicción en la gestión de procesos de admisión y captación de estudiantes (Kurzweil & Wu, 2015; Watermark, 2021).

Por otro lado, en el sector farmacéutico, donde se ubican los principales stakeholders institucionales de Pharma Business Education, la adopción de analítica avanzada ha demostrado tener un impacto positivo en la eficiencia de los procesos de recursos humanos y desarrollo de talento. Según McKinsey (2018), el uso de inteligencia artificial y técnicas de machine learning puede suponer una mejora del 30 % en la efectividad operativa de las organizaciones del sector.

Desde un punto de vista estratégico, la implementación de modelos predictivos permitirá a la institución anticiparse a comportamientos de alumnos, optimizar la asignación de

recursos del equipo humano y mejorar el ratio de conversión, en término de matriculaciones. Este enfoque no solo contribuirá a mejorar el posicionamiento de Pharma Business Education como referente digital en su sector, sino que también proporcionará un marco replicable en otras instituciones con retos similares.

En definitiva, el presente trabajo se justifica por la confluencia de tres factores clave:

- Disponibilidad de datos históricos relevantes y estructurados.
- Existencia de una necesidad operativa concreta que afecta directamente al rendimiento institucional.
- Oportunidad de generar valor a través de modelos de predicción aplicados a la toma de decisiones estratégicas.

1.3 Planteamiento del problema

Pese a la evolución positiva de Pharma Business Education en términos de volumen de alumnos formados y colaboraciones empresariales establecidas, persisten ciertos desafíos operativos críticos que limitan el potencial de escalabilidad y eficiencia institucional. En particular, se identifica una gran área de mejora:

- Falta de un sistema robusto de clasificación y priorización de leads dentro del funnel comercial.

El equipo de ventas y marketing (B2C) gestiona una gran cantidad de leads provenientes de múltiples canales (Google Ads, Meta Ads, redes sociales, eventos, formularios de contacto). Sin embargo, debido a la heterogeneidad de los perfiles captados y la ausencia de un sistema automatizado de scoring, los recursos humanos disponibles se distribuyen de manera ineficiente. Los leads más prometedores no siempre son abordados con prioridad, y una parte importante de los contactos se estanca en fases intermedias del embudo sin recibir el seguimiento adecuado.

La consecuencia directa de esta problemática implica una reducción en la conversión global, una asignación subóptima del tiempo del equipo y una posible pérdida de oportunidades tanto a nivel institucional como comercial. Este trabajo plantea la hipótesis

de que la implementación de modelos predictivos puede contribuir significativamente a resolver esta limitación.

1.4 Finalidad del estudio

La finalidad de este estudio es desarrollar una solución basada en analítica avanzada que permite transformar datos históricos en conocimiento accionable. En concreto, se busca diseñar e implementar dos modelos predictivos de machine learning, orientado a optimización del funnel de leads (área B2C).

A través de esta iniciativa, se pretende facilitar la toma de decisiones fundamentadas, aumentar la eficiencia operativa del equipo, y alinear los recursos institucionales con aquellas acciones que generen mayor retorno en términos de empleabilidad y captación. Esta propuesta se sustenta en la idea de que una gestión estratégica y automatizada de los datos puede elevar la competitividad de la institución y consolidar su posicionamiento como referente educativo digital en el sector farmacéutico.

1.5 Objetivos del trabajo

1.5.1 Objetivo general

Diseñar e implementar dos modelos predictivos basados en técnicas de machine learning para optimizar la conversión de leads comerciales en el área B2C de la institución Pharma Business Education.

1.5.2 Objetivos específicos

1. Realizar un diagnóstico detallado de los procesos actuales de captación, clasificación y conversión tanto en el área institucional como en la comercial.
2. Identificar las variables clave que impactan en la conversión de ofertas de prácticas y leads, mediante análisis exploratorios y correlacionales.
3. Diseñar un modelo de predicción para estimar la probabilidad de conversión de ofertas en prácticas efectivas, basado en variables como localización, área funcional, empresa, y perfil del alumno.

4. Validar ambos modelos a través de métricas de desempeño como AUC-ROC, F1-score, lift y precisión, comparando los resultados frente a los procesos actuales.
5. Proponer un plan de implementación y despliegue institucional de los modelos, incluyendo integración con las herramientas existentes (CRM, automatización, reporting).
6. Establecer indicadores clave (KPIs) para el seguimiento continuo del impacto de la solución.
7. Generar un marco metodológico replicable que pueda aplicarse en contextos similares dentro del ámbito educativo o del sector salud.

1.6 Breve descripción de la institución Pharma Business Education

1.6.1 Historia y misión

Pharma Business Education es una escuela de negocios digital fundada en el año 2022 con el propósito de acompañar y preparar a jóvenes titulados en ciencias de la salud para acceder con éxito a posiciones de entrada en la industria farmacéutica, biotecnológica y de tecnología sanitaria. Desde sus inicios, la institución se ha caracterizado por su enfoque 100 % online, su vocación práctica y su estrecha relación con el entorno empresarial.

En apenas tres años de funcionamiento, Pharma Business Education ha logrado consolidarse como una referencia en el sector formativo de acceso al empleo en la industria farmacéutica, con una comunidad de más de 800 alumnos formados y un porcentaje de empleabilidad superior al 60 % en los primeros meses tras la finalización del máster. Además, ha establecido más de 155 colaboraciones con empresas líderes del sector, desarrollando un ecosistema profesional que facilita la inserción laboral efectiva de sus estudiantes.

La misión institucional se centra en democratizar el acceso a la industria para perfiles con formación sanitaria, romper barreras de entrada, y ofrecer una formación adaptada a las exigencias reales del mercado. Bajo esta premisa, la institución ha desarrollado un modelo de formación en línea basado en contenidos actualizados, tutorías personalizadas, procesos

de selección simulados, y una fuerte orientación al desarrollo de habilidades profesionales e inserción laboral.

1.6.2 Proceso actual de conversión de ofertas y de conversión de leads (B2C)

El modelo operativo en el que nos vamos a centrar dentro de la organización interna de Pharma Business Education se estructura en el área de marketing y ventas (B2C).

En el área B2C, la captación de leads se realiza a través de campañas de pago (Google Ads, Meta Ads), posicionamiento orgánico (SEO), presencia activa en redes sociales (Instagram, LinkedIn, TikTok, YouTube) y participación en ferias y eventos universitarios. Los leads cualificados llegan al CRM (Hubspot) mediante formularios digitales (Typeform), donde son segmentados automáticamente según variables clave como ubicación, titulación y edad. A partir de este punto, el equipo comercial realiza un primer contacto vía llamada o WhatsApp para explorar el interés y guiar al candidato por las distintas etapas del embudo. Aquellos leads que no responden son gestionados mediante secuencias de email automatizadas o bots inteligentes con el fin de mantener el interés activo. Finalmente, los perfiles que superan el proceso de admisión formalizan su inscripción y se convierten en alumnos matriculados.

Ambas áreas comparten un reto común: la gestión eficiente de los datos y la necesidad de establecer criterios objetivos y automatizados para la toma de decisiones, tanto en la asignación de alumnos a ofertas como en la priorización de leads comerciales. En este sentido, la presente investigación se plantea como una oportunidad para dotar a la organización de herramientas predictivas que permitan dar un salto cualitativo en su funcionamiento operativo.

1.7 Recursos disponibles y entorno tecnológico

La implementación de un proyecto de analítica predictiva requiere no solo de un diagnóstico preciso del problema, sino también de la disponibilidad de los recursos adecuados para su ejecución. En este sentido, Pharma Business Education cuenta con una combinación de recursos humanos, tecnológicos y organizativos que permiten afrontar con garantías el desarrollo de los modelos propuestos.

Desde el punto de vista humano, el equipo de la institución está conformado por profesionales especializados en áreas clave como relaciones institucionales, marketing digital, captación de talento, analítica de datos y tecnología educativa. El equipo de B2C cuenta con especialistas en campañas digitales, SEO, funnel de conversión y asesoramiento educativo y profesional, incluyendo profesionales con perfiles analíticos que permiten trabajar con datos extraídos del CRM y otras plataformas.

En cuanto a los recursos tecnológicos, la institución trabaja con un stack de herramientas digitales bien consolidado. El CRM Hubspot constituye la columna vertebral de la captación y seguimiento de leads, permitiendo segmentaciones, automatizaciones, tareas y visualización del pipeline comercial. Para la captación y análisis del tráfico digital se utilizan plataformas como Google Ads, Meta Ads, Google Analytics y herramientas de SEO. La gestión de procesos internos se apoya en Notion, mientras que la web institucional está desarrollada sobre Webflow, lo que permite actualizaciones ágiles y personalización de la experiencia de usuario.

El desarrollo del modelo predictivo se llevará a cabo en el lenguaje Python, aprovechando librerías como pandas para la manipulación de datos, scikit-learn para la creación de modelos de machine learning, matplotlib y seaborn para la visualización de resultados, y herramientas de automatización como Jupyter Notebooks y Google Colab. Esta elección técnica se justifica por la robustez, escalabilidad y flexibilidad que ofrece el ecosistema Python en proyectos de ciencia de datos.

En cuanto a los recursos de datos, Pharma Business Education dispone de un repositorio histórico desde el año 2022 que incluye: registros detallados de todas las ofertas de prácticas captadas, procesos de selección realizados, convenios firmados, leads recibidos clasificados por canal y tipo, datos sociodemográficos de los candidatos, tasas de conversión por etapa del funnel, tiempos medios de contacto y duración de procesos, así como resultados de campañas de marketing. Esta base de datos representa un activo estratégico fundamental para el éxito del proyecto, permitiendo no solo la construcción de los modelos, sino también su validación empírica y análisis continuo.

Por último, la dirección de la institución ha mostrado un alto grado de compromiso con la mejora basada en datos, lo que garantiza el apoyo organizativo necesario para la implementación y escalado de las soluciones propuestas. La cultura organizativa orientada a la mejora continua y la innovación actúa como catalizador para que este tipo de proyectos puedan desplegarse con éxito dentro del marco estratégico institucional.

1.8 Diagnóstico estratégico: análisis DAFO

El análisis DAFO (Debilidades, Amenazas, Fortalezas y Oportunidades) permite obtener una visión integral del contexto interno y externo de la institución respecto a la implementación de soluciones predictivas en sus procesos operativos. A continuación, se presenta una síntesis del diagnóstico estratégico de Pharma Business Education:

Fortalezas

- Fuerte posicionamiento en un nicho especializado: formación online para el acceso a la industria farmacéutica.
- Base de datos estructurada con registros históricos de leads desde 2022.
- Elevado grado de satisfacción de los alumnos actuales, lo que genera un canal orgánico de captación vía recomendaciones (NPS superior a 40).
- Equipo multidisciplinar con competencias en áreas clave (ventas, marketing digital, analítica de datos).
- Compromiso institucional con la mejora continua y la innovación basada en datos.

Debilidades

- Ausencia de modelos predictivos en los procesos de toma de decisiones.
- Procesos de priorización y asignación realizados de forma manual, con altos costes operativos.
- Saturación del equipo de ventas, que gestiona leads sin diferenciación cualitativa.

Oportunidades

- Creciente demanda de perfiles con formación técnica y digital en el sector salud y farmacéutico.

- Avances en herramientas de automatización y analítica accesibles mediante código abierto.
- Tendencia generalizada hacia la adopción de decisiones basadas en datos en el sector educativo.
- Potencial de replicabilidad del modelo predictivo en otras líneas de negocio y sectores.
- Posibilidad de integrar soluciones de inteligencia artificial conversacional para el filtrado inicial de leads.

Amenazas

- Intensificación de la competencia en el sector educativo especializado.
- Limitaciones normativas relacionadas con la protección de datos personales (RGPD).
- Saturación publicitaria en los canales de captación digital, que podría reducir la efectividad de las campañas pagadas.
- Dificultad para mantener una personalización efectiva del trato humano si no se gestiona correctamente la automatización.

Este análisis pone de manifiesto la oportunidad estratégica que representa la implementación de modelos analíticos en Pharma Business Education, así como los desafíos que deberán ser gestionados de forma proactiva para garantizar la sostenibilidad y el impacto del proyecto.

2. Marco teórico

2.1 Introducción

El marco teórico que conforma nuestro proyecto de investigación se centra el área de conocimiento de Business Analytics y su aplicación dentro una escuela de negocio especializada en la industria farmacéutica de nombre Pharma Business Education, mediante la combinación de análisis de datos, estadística, visualización de datos, modelos de Machine Learning (ML) y reporting podremos extraer información estratégica y de alto valor para encontrar la relevancia de los datos sobre los puntos objeto de estudio en este trabajo de investigación , los cuales serán las piezas fundamentales para resolver la interrogante que desde la Gerencia de la empresa se ha originado.

Con este uso óptimo del Business Analytics (BA), se busca mejorar la conversión de leads para el departamento de B2C. Para ello, se elaborará un modelo que pueda predecir desde una etapa temprana cuál es el porcentaje de finalización para aquellos leads que aún no se han convertido a alumnos regulares. dando oportunidad al departamento de ventas para precisar el seguimiento de los leads y crear estrategias de ventas focalizadas y personalizadas con la finalidad de aumentar la conversión de dichos leads que actualmente se ubica en un 30% y con esta medida se quiere aumentar a un 45% para el primer año. La finalidad de la aplicación de estos modelos predictivos permitirá a Pharma Business Education mejorar la conversión de los leads, mejorar su porcentaje de market share y consolidarse como una escuela de negocios líder en el sector, perdurable en el tiempo y mejorando su reputación.

2.2 Business Analytics

Se denomina Business Analytics (BA) al conjunto de herramientas de tipo tecnológico que aunado al análisis de la información empresarial cuantificable nos permite obtener información valiosa para la toma de decisiones estratégicas. Esta disciplina se apoya en modelos predictivos, estadísticos y tecnológicos que dan como resultado un mejor rendimiento operacional, identificar patrones y tendencias en áreas como, marketing,

ventas, finanzas, producción, recursos humanos, adquisiciones entre otras. (Gómez, 2021; Laursen & Thorlund, 2016)

El Business Analytics se divide principalmente en 4 tipos de análisis y cada uno tiene un propósito definido: Análisis descriptivo, Análisis de Diagnóstico, Análisis Predictivo y Análisis Prescriptivo. (Marr, 2016)

Análisis Descriptivo

Tiene como principio el análisis de los datos históricos buscando entender qué ha sucedido en el pasado y presente. Consiste en preparar y analizar datos históricos para identificar patrones y tendencias. Es el tipo más simple y común entre los negocios, y es la base de toda la interpretación de datos. El análisis descriptivo responde a la pregunta "¿qué sucedió?" en un periodo determinado al resumir datos anteriores. Este tipo de análisis suele utilizarse principalmente para dar seguimiento a los indicadores clave de rendimiento (KPI) que establece cada empresa. Su ventaja principal es que ayuda a identificar áreas de mejora y oportunidades basadas en datos históricos. Se basa en técnicas de agregación y exploración de datos. (K. Ortega, 2023; J. Martel, 2018; Marr, 2016)

Análisis Diagnóstico

Se basa en entender por qué sucedieron los eventos encontrados en el análisis descriptivo. Identifica las causas que llevaron al estado actual de la organización Este tipo de análisis responde a la pregunta "¿por qué sucedió?", a partir de los descubrimientos realizados en el análisis descriptivo. Los analistas de datos profundizan en los resultados para encontrar las causas de los patrones de comportamiento, lo cual es información de gran valor para las empresas, especialmente para aquellas dedicadas al marketing y al comercio. El análisis de diagnóstico también funciona para detectar y solucionar problemas rápidamente. (K. Ortega, 2023; Laursen & Thorlund, 2016)

Análisis Predictivo

Se define como el método analítico donde al conjunto de datos se le aplican modelos estadísticos, matemáticos y de machine learning con el fin de poder predecir eventos futuros y poder anticiparse a los diferentes comportamientos que pueda tener las áreas que conforman la organización. Intenta responder a la pregunta "¿qué es probable que suceda?", a partir de datos anteriores y hacer estimaciones sobre resultados futuros. Este tipo de análisis es un poco más complejo, ya que se basa en el modelado estadístico (una herramienta basada en las matemáticas donde se combinan datos cualitativos y cuantitativos), que requiere tecnología adicional y profesionales especializados para realizar pronósticos correctamente. El análisis predictivo puede utilizarse, por ejemplo, para evaluar riesgos, pronosticar ventas y segmentar clientes. Su ventaja radica en la capacidad de prever escenarios futuros y planificar estrategias en consecuencia. (K. Ortega, 2023; Shmueli et al., 2019)

Análisis Prescriptivo

Es la técnica dentro de la analítica de datos que no solo predice los eventos, sino que nos otorga información vital para poder realizar acciones óptimas con el fin de obtener los mejores resultados posibles. Se basa en algoritmos de inteligencia artificial, optimización y modelos matemáticos para asegurar una correcta toma de decisiones. Se trata de una combinación de todos los tipos de análisis de datos anteriores, que determina qué acción tomar para eliminar un problema (actual o futuro) o aprovechar al máximo una tendencia, es decir, el análisis prescriptivo se utiliza para mejorar la toma de decisiones. También es un tanto complejo de implementar, pues utiliza herramientas de análisis y tecnologías avanzadas, como el aprendizaje automático, reglas comerciales y algoritmos, la Inteligencia Artificial (IA) es el ejemplo perfecto para esto.

Cada uno de estos tipos de análisis de datos están conectados y hasta cierto punto son interdependientes. A pesar de que cada uno tiene un propósito diferente y proporciona información particular, en ocasiones es indispensable trabajar con más de uno a la vez y

para esto se requiere de mucho conocimiento en el sector, así como una serie de habilidades como, la atención al detalle, la organización y pensamiento crítico. (K. Ortega, 2023; Shmueli et al., 2019)

2.3 Evolución del Business Analytics

El Business Analytics ha tenido un importante crecimiento y expansión a lo largo de los últimos años desde que un Informático de origen Alemán de nombre Hans Peter Luhn (IBM) publicará en el año 1958 su artículo **“** A Business Intelligence System **”** en el que hablaba de un sistema que almacenará la información en documentos textuales y definía el Business Intelligence como << la capacidad de aprender las interrelaciones de los hechos presentados de tal manera que guíen la acción hacia un objetivo deseado>>(CANO, 2018).

Un par de años más tarde, Richard Kimball (orientado a la consulta de información rápida y sencilla) y Bill Inmon (orientado al almacenamiento de grandes cantidades de datos) revolucionaron la arquitectura de bases de datos analíticas, ya que, hasta entonces, el modelado de las bases de datos se orientaba a optimizar el almacenamiento debido a su coste.

En cambio, no es hasta 1989 cuando Howard Dresner, quien describió el BI como “un término general para describir conceptos y métodos para mejorar la toma de decisiones comerciales mediante el uso de sistemas de soporte basados en hechos”

Ya por el 2000, surgen otro tipo de analítica de negocios más moderna, proactiva, rápida y eficaz, que necesita al Business Intelligence como base para sus cálculos estadísticos e inferencias. El Business Analytics**.** (CANO, 2018)

2.4 Rol estratégico del análisis de datos en las escuelas de negocio

Para nadie es un secreto que obtener información relevante de los datos para la toma de decisiones nos ofrece una ventaja sobre nuestra competencia y más aún en un entorno tan

competitivo como el educativo. Las escuelas de negocio han aumentado su participación en los últimos años ofreciéndose como una alternativa a las conocidas universidades que en algunos casos mantienen sistemas educativos obsoletos y no acordes al ritmo de vida de hoy día, con horarios poco flexibles, y con una base tecnológica deficiente Las escuelas de negocios ofrecen una solución para todos aquellos que por motivos de tiempo no pueden amoldarse a los rígidos horarios universitarios además que con la ventaja de la educación remota y asincrónica se convierten en una potente solución para el estudiante.

Es por ello que tener un óptimo análisis de datos permite, poder conocer el estado actual de la organización, identificar en tiempo real, tendencias o patrones que puedan estar perjudicando a la empresa, poder predecir su comportamiento y anticiparse a los desafíos que están por venir.

Utilizar las herramientas analíticas en una escuela de negocio entrega información estratégica necesaria para destacarse del resto, poder personalizar las estrategias de marketing, optimizar la gestión administrativa, académica, del personal entre otras. Es por ello que Pharma Business Education busca con este proyecto de investigación aumentar su capacidad de conversión de los leads (B2C) todo esto bajo una total cultura Data-Driven. (Provost & Fawcett, 2013)

2.5 Estrategias Data-Driven

El término Data-Driven hace referencia a un enfoque de gestión y toma de decisiones fundamentado en el análisis sistemático de datos. En contraposición a los modelos tradicionales basados en la intuición o la experiencia, el enfoque Data-Driven promueve la recopilación, procesamiento e interpretación de grandes volúmenes de información con el fin de generar conocimientos accionables que mejoren los procesos internos y externos de una organización. En el ámbito empresarial, las estrategias Data-Driven permiten una mejor segmentación de los públicos objetivos, una personalización más precisa de las campañas de marketing, así como una mayor capacidad de reacción ante los cambios del entorno. La clave de su efectividad reside en la calidad de los datos, en la capacidad analítica de la

organización y en la integración transversal de estos conocimientos en la toma de decisiones estratégicas. (EY España, 2024)

Cultura organizacional orientada a datos

El éxito de una estrategia Data-Driven no solo depende de la tecnología, sino también de la cultura de la organización. Las empresas que fomentan una mentalidad basada en datos y les profesan a todos sus colaboradores la importancia de una correcta manipulación y tratamiento de estos, aprenden a tomar decisiones sustentado en evidencia empírica y no en suposiciones. Promover la colaboración entre departamentos para maximizar el valor de los datos es esencial en un entorno donde el dato es la pieza fundamental. Tener una infraestructura tecnológica acorde a las necesidades de la empresa es vital para un correcto funcionamiento. (C. Ortega, 2024; McAfee et al., 2012)

2.6 Ventajas del Data-Driven

a) Minimización de los riesgos y reducción de errores.

Al contar con información confiable basada en datos limpios, y precisos, nos otorga seguridad para la toma de decisiones ya que el error se reduce al mínimo, así como los riesgos de elaborar estrategias con datos poco fiables.

b) Optimización de los procesos dentro de la organización

Plantea el mejoramiento de los procesos internos, tener información en tiempo real, eficiencia y eficacia en el manejo de los datos, esto conduce a la operatividad empresarial deseada, minimiza los costes en procesos poco útiles, reduce los tiempos de ejecuciones, se anticipa para evitar los llamados cuellos de botella, maximiza la capacidad instalada de la empresa, optimiza los recursos de personal, entre otros.

c) Mejora la gestión financiera

Con información en tiempo real nos ayuda a prever fluctuaciones del mercado y así prepararnos para eventos posteriores minimizando las posibles pérdidas.

d) Identificar nuevas oportunidades

Un entorno Data-Driven donde la información está al alcance de la mano nos brinda nuevas oportunidades de negocio, permitiéndonos evolucionar con el mercado, leer las tendencias y los cambios en definitiva ser protagonista y no un espectador.

e) Uso de KPI (Indicadores de rendimiento clave) para el seguimiento continuo de la gestión empresarial

Los "Key Performance Indicators" (KPI) por sus siglas en inglés son un aliado estratégico para mantener un seguimiento continuo de los objetivos definidos por la organización lo que mantiene el equipo motivado y comprometido al tener una medición de su gestión en tiempo real. (Eckerson, 2010)

2.7. Analítica predictiva en la industria farmacéutica

La analítica predictiva se basa en la utilización de técnicas estadísticas, algoritmos de Machine Learning y modelos matemáticos para identificar patrones en los datos históricos y prever comportamientos futuros. En la industria de negocio de escuela de farmacéutica, esta tecnología ha cobrado especial relevancia para anticipar tendencias del mercado, estimar la respuesta de los alumnos potenciales a nuevas campañas, y optimizar la distribución de recursos en función del rendimiento esperado. Entre sus aplicaciones más destacadas se encuentran la predicción de tasas de prescripción, el análisis del ciclo de conversión, y la identificación de alumnos con alto potencial de conversión. Al combinarse con modelos Data-Driven, la analítica predictiva permite una gestión proactiva del alumno-empresa, facilitando intervenciones más oportunas y efectivas.

3. Metodología

3.1 Objetivos e hipótesis

3.1.1 Objetivos

El objetivo de este Trabajo Final de Máster (TFM) es generar mediante el análisis de datos transformados, informe de propuesta para consumo de la Gerencia Pharma Business Education, que coadyuve a la toma de decisiones y a la obtención de resultados altamente positivos, significativos y comprobables, manteniendo el enfoque del estudio en los problemas de la empresa ya definidos: una baja tasa de conversión de leads a alumnos matriculados (Proceso B2C).

Por tanto, se buscará influir directamente en la transformación de los procesos Pharma Business Education aproximando su modelo hacia a la objetividad, simplicidad y automatización, basado en Data Driven y Machine Learning, esto con la idea de minimizar significativamente el esfuerzo y requerimiento humano en procesos de análisis matemático-estadístico, seguimiento y control de operaciones, y por consiguiente enfocar el esfuerzo de la Gerencia en la toma de decisiones y aplicabilidad continua de este modelo dentro de su negocio.

Para el fin propuesto trabajaremos distintas metodologías que agrupan estrategias de análisis descriptivo, predictivo y prescriptivo. Se emplearán, por tanto, en etapa descriptiva el modelo D.A.S.E (Descriptive Analytics Structured Exploration), en predictiva y Machine Learning P.R.E.M.I.A. Predictive & Machine Learning Integration Approach) y finalmente emplearemos Metodología O.P.T.I.M.A. (Optimization & Prescriptive Techniques for Intelligence Management & Action)

3.1.2 Hipótesis

Se espera que las metodologías descriptivas seleccionadas en este trabajo brinden un panorama completo y estructurado que describa de manera analítica la situación, detecte el problema y defina el peso que tienen las distintas variables sobre nuestra variable objetivo (Conversiones B2C).

Asimismo, se buscará mediante técnicas de análisis predictivo reconocer patrones, tendencias y escenarios que sirvan como récord histórico y guía que ayuden en la identificación necesaria de puntos de influencia exactos tanto internos como externos que requieran tratamiento y que deban ser sujetos a transformación y análisis en función de resolver los problemas detectados.

Finalmente se espera influir en la toma de decisiones y manejo de procesos de Pharma Business Education mediante la creación de un análisis prescriptivo modelado conforme a la empresa y que se alinee a los objetivos perseguidos.

3.2 Diseño

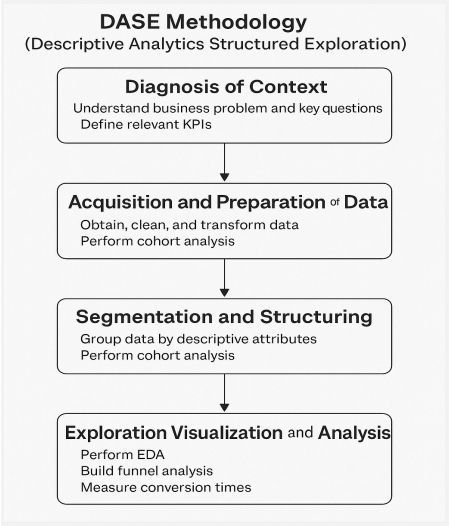
Diseño: Descriptive Analytics Structured Exploration (D.A.S.E) / Predictive & Machine Learning Integration Approach (P.R.E.M.I.A) / Optimization & Prescriptive Techniques for Intelligent Management & Action (O.P.T.I.M.A.)

3.2.1. Descriptive Analytics Structured Exploration (D.A.S.E)

La metodología D.A.S.E está diseñada para estructurar el análisis descriptivo de datos en proyectos de analítica de negocio. Integra técnicas exploratorias, de segmentación, de cohortes y de análisis de conversión bajo un enfoque secuencial y sistemático.

Esta metodología comprende 4 etapas: Diagnóstico del Contexto, Adquisición y Preparación de Datos, Segmentación y Estructuración y finalmente Exploración Visual y Analítica.

La imagen del proceso D.A.S.E. muestra cada una de estas etapas:



Fuente: OpenAI (2025). ChatGPT.

Los datasets sujetos a esta metodología, tienen origen en dos Software tipo CRM de Pharma Business Education, uno para el levantamiento, control y seguimiento de procesos de clientes y ventas (Hubspot), y el otro para la gestión de alumnos (Notion).

DASE da el marco conceptual para agrupar técnicas de análisis descriptivo Funnel, Cohortes, Segmentación, EDA y Lead Velocity.

La analítica descriptiva permite estructurar el conocimiento de los datos mediante procesos sistemáticos de limpieza, exploración visual, segmentación y análisis de patrones que explican el comportamiento actual del negocio. (Rodríguez, J. 2021).

3.2.2. Predictive & Machine Learning Integration Approach (P.R.E.M.I.A)

La metodología PREMIA está diseñada para estructurar el desarrollo e implementación de modelos predictivos y de machine learning dentro de un contexto de analítica aplicada a negocios. Se basa en buenas prácticas del ciclo de vida del modelado, pero adaptada a una visión práctica, orientada a la toma de decisiones.

El aprendizaje automático aplicado a los negocios permite construir modelos predictivos capaces de anticipar el comportamiento de clientes, optimizar procesos comerciales, segmentar mercados y apoyar decisiones estratégicas basadas en patrones descubiertos en los datos. (Fernández, A., & García, S. 2020).

Con esta aplicación metodológica, se logrará una integración estructurada de técnicas de machine learning para priorizar leads y clientes, anticipar conversiones o cancelaciones, segmentar audiencias de forma inteligente, apoyar decisiones comerciales basadas en predicciones confiables.

La imagen del proceso P.R.E.M.I.A muestra cada una de estas etapas

Fuente: OpenAI (2025). ChatGPT.

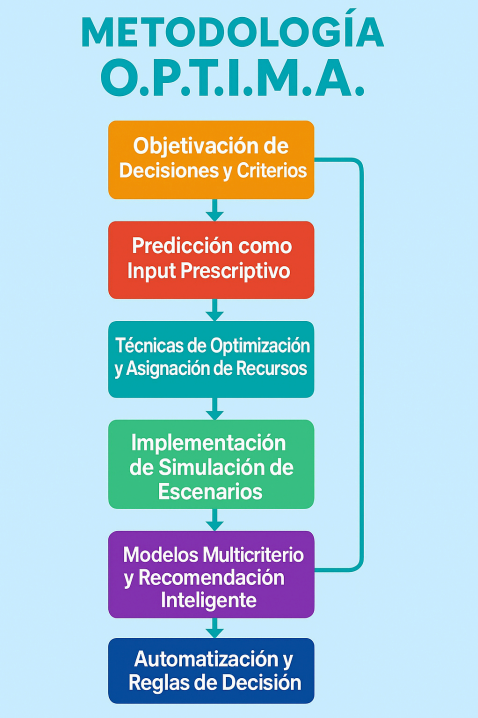
El objetivo es lograr traducir las necesidades Pharma Business Education en una tarea de machine learning concreta mediante técnicas de clasificación, regresión, segmentación. Además de disponer de un datawarehouse con base de datos aptas para entrenar modelos con alta calidad y relevancia predictiva, evaluables en calidad y utilidad comparando los resultados con métricas clave.

3.2.3 Optimization & Prescriptive Techniques for Intelligent Management & Action (O.P.T.I.M.A)

La metodología O.P.T.I.M.A proporciona un marco estructurado para aplicar técnicas de analítica prescriptiva, orientadas a la toma de decisiones estratégicas y operativas con base en modelos predictivos, simulaciones, optimización matemática y sistemas inteligentes. (López-Iturriaga, F. J. 2020).

Se busca con esta metodología, definir claramente los objetivos de negocio de Pharma Business Education, cuáles son los procesos se quieren optimizar y definir qué criterios intervendrán en la toma de decisiones. Se utilizará por tanto modelos de análisis predictivos como entrada para producir mejoras y establecer recomendaciones ejecutables por la Gerencia. Adicionalmente se encontrará la mejor combinación de acciones y recursos que permitan un mejor rendimiento.

La imagen del proceso O.P.T.I.M.A muestra cada una de estas etapas



Fuente: OpenAI (2025). ChatGPT.

Finalmente se evaluará el comportamiento del sistema de Pharma Business Education bajo distintos escenarios hipotéticos o inciertos como métodos de comprobación y análisis prospectivo.

3.3 Participantes

3.3.1 Product Owner (Propietario del producto)

Este rol debe ser asumido por un perfil que reconozca la estructura y el funcionamiento de la organización objeto de análisis, con acceso directo a la gerencia, por lo tanto, se ha establecido un canal de comunicación formal con Pharma Business Education, a través del cual se ha recopilado datos e información de sus operaciones y se ha determinado el alcance de este trabajo. En consecuencia, la responsabilidad de este rol será asumida por Inés Martínez.

El éxito de los proyectos de analítica depende en gran medida de la capacidad de integrar el conocimiento del negocio con las capacidades técnicas, lo cual requiere perfiles capaces de traducir las necesidades estratégicas en soluciones de datos. (García. A, 2018).

3.3.2 Scrum Máster (Facilitador de proyecto)

Su función principal es coordinar las reuniones y velar por el cumplimiento de los principios del marco de trabajo Scrum, debe ser representado por una figura de carácter colaborativo y auto-delegante. La persona bajo este rol buscará garantizar una adecuada organización y seguimiento metodológico, esta responsabilidad será asumida por Sergio Escudero.

Más allá de la facilitación de reuniones, el Scrum Máster debe ser el principal garante del cumplimiento de las prácticas ágiles, impulsando la mejora continua y promoviendo un entorno de trabajo colaborativo y sostenible. (Torres, M. 2022).

3.3.3 Equipo de Desarrollo

En el presente proyecto, todas las actividades técnicas y analíticas serán abordadas de manera conjunta por los integrantes del equipo de trabajo. Esto incluye tareas clave como la

limpieza y el análisis exploratorio de datos, el desarrollo e implementación de modelos estadísticos y de aprendizaje automático, así como la construcción de algoritmos predictivos. Estas funciones serán asumidas de forma colaborativa por el equipo Pharma Team, conformado por Adriana Gainza, Inés Martínez, Nelson Méndez y Sergio Escudero, quienes trabajarán de manera coordinada a lo largo de las distintas fases del Trabajo Final de Máster.

El equipo de desarrollo trabaja de manera colaborativa, compartiendo la responsabilidad sobre los resultados, eliminando los roles jerárquicos tradicionales. (Pérez, L. 2020).

3.4 Instrumentos

3.4.1 Instrumentos para la recolección y preparación de datos

Etapas: Diagnóstico del Contexto, Adquisición y Preparación de Datos (D.A.S.E)

CRM HubSpot: para la exportación de datos de leads, conversiones, fuentes e interacciones. Es un sistema de gestión de relaciones con el cliente (Customer Relationship Management) que ayuda a las empresas a organizar y optimizar su proceso de ventas, marketing y servicio al cliente. Permite gestionar información de clientes, contactos, empresas y oportunidades de venta, así como las interacciones a través de diferentes canales. (Sánchez, J. 2021).

Notion: para la extracción de datos sobre la gestión y avance de alumnos. Notion es una herramienta flexible de gestión de información que permite a los usuarios construir bases de datos personalizadas, automatizar flujos de trabajo y organizar datos de clientes o proyectos, ofreciendo funcionalidades similares a las de un CRM adaptable a las necesidades de cada organización. (Martín, L. 2022).

Google Sheets: para el almacenamiento y gestión de data sets, limpieza inicial, filtrado, validación y codificación de variables. Aplicación web de hojas de cálculo, similar a Excel, que forma parte de Google Workspace. Es una herramienta en línea para la creación y gestión colaborativa de hojas de cálculo, facilitando la organización, validación y limpieza

inicial de datos sin necesidad de instalaciones locales, lo que la hace ideal para la gestión eficiente de datasets en proyectos de análisis de datos. (Pérez, M. 2022).

Power BI/Power Query: para la transformación de datos, creación de columnas calculadas, combinación de tablas, operaciones ETL y visualizaciones y segmentaciones dinámicas. Plataforma de Business Intelligence (BI) y análisis de datos de Microsoft. Power BI es una solución de inteligencia de negocio que permite transformar, modelar y visualizar datos de manera interactiva, integrando capacidades ETL a través de Power Query, lo que facilita la preparación de los datos, la creación de columnas calculadas, la combinación de fuentes y la generación de dashboards dinámicos para la toma de decisiones. (Hernández, P. 2022).

Visual Studio Code: para la ejecución de códigos en el ejercicio del TFM, es un editor de código abierto, ligero y multiplataforma, diseñado para programadores que requieren escribir, depurar y ejecutar código en múltiples lenguajes, facilitando el desarrollo de aplicaciones, scripts y automatizaciones en entornos de análisis de datos. (Ruiz, C. 2022)

Jupiter Notebook: para la descripción textual y ejecución de códigos en el ejercicio del TFM. Es una aplicación web de código abierto que permite combinar código en vivo, narrativas descriptivas, visualizaciones y ecuaciones matemáticas, facilitando el desarrollo interactivo de proyectos de análisis de datos y aprendizaje automático. (Morales, D. 2022).

Google Colab: para la ejecución de códigos en el ejercicio del TFM, plataforma online gratuita que permite programar en Python mediante cuadernos interactivos en la nube, ofreciendo acceso a recursos de computación avanzados, facilitando el desarrollo de proyectos de análisis de datos, aprendizaje automático y ciencia de datos sin necesidad de instalación local. (Navarro, J. 2023).

Lenguaje de Programación Python: Usando librerías como pandas, numpy, Matplotlib, entre otras para el tratamiento de datos masivos y automatización del preprocesamiento, operaciones EDA. Python es un lenguaje de programación flexible y potente, ampliamente utilizado en el análisis de datos por su sintaxis sencilla y la disponibilidad de librerías

especializadas como pandas, numpy y matplotlib, que facilitan el tratamiento de grandes volúmenes de datos, su preprocesamiento, exploración visual y automatización de tareas analíticas. López, M. (2023).

3.4.2 Instrumentos para análisis predictivo (P.R.E.M.I.A)

Etapas: Definición de objetivos predictivos, modelado, validación y evaluación.

Lenguaje de programación Python: para desarrollo de modelos de clasificación, clustering, series temporales y regresiones lineal, múltiple, logística, para la toma de decisiones, optimización de recursos, mejora de la experiencia al cliente, incremento de la eficiencia operativa, identificación de nuevas oportunidades de negocio, pronóstico de demanda entre otros.

VSCode/ Jupyter Notebooks/ Google Colab: para el desarrollo interactivo de modelos y visualización de resultados.

Data Warehouse local / Google BigQuery: almacenamiento estructurado y consulta eficiente de datos históricos. Un *data Warehouse* es un sistema de almacenamiento estructurado de datos que permite consolidar información histórica de diferentes fuentes para facilitar su análisis. Soluciones en la nube como Google BigQuery ofrecen alta escalabilidad, velocidad de consulta y facilidad de integración con herramientas analíticas, optimizando los procesos de inteligencia de negocio. (Ortega, S. 2023).

3.4.3 Instrumentos para análisis prescriptivo (O.P.T.I.M.A)

Etapas: Simulación, optimización, generación de recomendaciones

Python/ scikit-learn, TensorFlow: para crear modelos de optimización avanzada. Librerías como scikit-learn y TensorFlow, integradas en el ecosistema de Python, permiten desarrollar modelos de aprendizaje automático supervisado y no supervisado, redes neuronales y técnicas avanzadas de inteligencia artificial, facilitando tanto la experimentación como la producción de soluciones predictivas aplicadas a los negocios. (Gutiérrez, P. 2023).

3.4.4. Instrumentos metodológicos y colaborativos

Notion: para la gestión de tareas y seguimiento del proyecto en metodología ágil (Scrum).
Google Drive: como repositorio de datos, reportes y documentos colaborativos. Google Drive es un servicio de almacenamiento en la nube que permite guardar, organizar y compartir archivos de forma colaborativa. Su integración con el resto de las aplicaciones de Google Workspace lo convierte en una herramienta fundamental para el trabajo en equipo, la gestión de datos y la disponibilidad de documentación en entornos de investigación y analítica de datos. (Vargas, E. 2023).

ChatGPT: para apoyo en investigación, redacción técnica, corrección de estilo, codificación, interpretación de resultados y sugerencias metodológicas. Modelos de lenguaje como ChatGPT, basados en inteligencia artificial generativa, son empleados como herramientas de apoyo en investigación académica y empresarial, facilitando la generación de texto técnico, la asistencia en codificación, la interpretación de resultados analíticos y la sugerencia de marcos metodológicos, optimizando tiempos y recursos en el desarrollo de proyectos de analítica de datos. (Ramírez, L. 2024).

3.5 Eventos

3.5.1 Sprints

El proyecto se desarrollará en ciclos cortos de trabajo denominados *sprints*, con una duración estimada de entre 3 días y una semana, dependiendo de la complejidad de las tareas asignadas en cada fase.

3.5.2 Reuniones diarias (Daily Stand-up Meetings)

Durante la ejecución del proyecto, se llevarán a cabo reuniones breves y diarias destinadas a revisar el avance, identificar posibles obstáculos y resolver inquietudes del equipo. Se utilizará WhatsApp como canal complementario, adaptándose así a la disponibilidad de los participantes.

WhatsApp es una herramienta de mensajería instantánea ampliamente utilizada en entornos profesionales como canal de comunicación, seguimiento de proyectos, coordinación de equipos y atención a clientes, favoreciendo la agilidad en la toma de decisiones y el intercambio inmediato de información. (Gómez, F. 2023).

3.5.3 Revisión y retrospectiva de sprint (Sprint Review y Retrospective).

Al finalizar cada sprint se efectuará una revisión de los entregables producidos y una reflexión colaborativa sobre el desarrollo del trabajo. Estas sesiones incluirán la retroalimentación obtenida en reuniones virtuales con los stakeholders vinculados al área académica del TFM, así como impresiones y evaluaciones proporcionadas por la empresa cliente, asegurando así la alineación del proyecto con las expectativas y necesidades de ambas partes.

3.6 Procedimiento

El desarrollo se articula en ocho etapas sucesivas que responden a un enfoque integral de análisis de datos en contexto empresarial. Cada una de estas fases se alinea con los objetivos estratégicos de Pharma Business Education, abordando los desafíos detectados en sus procesos B2C mediante el uso de herramientas de analítica avanzada y automatización.

3.6.1 Comprensión del negocio

Objetivo: comprender el funcionamiento de los procesos actuales de la empresa y los problemas asociados.
Revisión de documentación interna y entrevistas con la Gerencia.
Identificación de necesidades clave: mejora de tasas de conversión.
Definición de objetivos de negocio y KPIs relacionados.
Resultado: documento de entendimiento del contexto empresarial y objetivos analíticos.

3.6.2 Comprensión de los datos

Objetivo: conocer la disponibilidad, estructura y calidad de los datos relevantes para el análisis.

Identificación de fuentes de datos (CRM, LMS, bases académicas).
Exploración inicial de variables, estructuras y relaciones.
Evaluación de la calidad de los datos: valores nulos, duplicados, inconsistencias.
Resultado: informe técnico de auditoría de datos y diccionario de variables.

3.6.3 Preparación de los datos

Objetivo: transformar y estructurar los datos para garantizar su utilidad en el análisis.
Limpieza de datos: normalización, tratamiento de valores faltantes.
Creación de variables derivadas (feature engineering).
Estandarización de formatos y codificaciones.
Resultado: dataset limpio y estructurado para análisis y modelado.

3.6.4 Evaluación y uso del CRM

Objetivo: integrar el software CRM como herramienta clave en la recopilación y gestión de datos.
Análisis de capacidades del CRM implementado (HubSpot, Notion).
Exportación e integración de datos relevantes.
Configuración de eventos, etiquetas y funnel de conversión.
Resultado: integración funcional entre CRM y el proceso analítico.

3.6.5 Combinación e integración de datos

Objetivo: consolidar toda la información relevante en un único entorno de análisis.
Cruce de información de diferentes plataformas (CRM, sistemas educativos).
Validación de integridad en la combinación de bases de datos.
Construcción de un repositorio único para el análisis (data mart o dataset maestro).
Resultado: dataset maestro validado y documentado.

3.6.6 Modelado analítico descriptivo

Objetivo: obtener una comprensión profunda de los datos mediante técnicas exploratorias.
Visualización de tendencias, tasas y comportamientos.

Segmentación de leads según variables clave.
Estudio de correlaciones e identificación de patrones.
Resultado: informe descriptivo y hallazgos clave.

3.6.7 Modelado predictivo

Objetivo: construir modelos de predicción que anticipen comportamientos clave.
Selección de algoritmos de machine learning apropiados.
Entrenamiento, validación y evaluación de modelos.
Interpretación de resultados y análisis de importancia de variables.
Resultado: modelos predictivos listos para aplicación operativa.

3.6.8 Despliegue e integración del modelo

Objetivo: poner en práctica los modelos desarrollados para apoyar la toma de decisiones.
Generación de recomendaciones automatizadas.
Creación de reportes y visualizaciones operativas para la Gerencia.
Diseño de escenarios y estrategias prescriptivas y prospectivas basadas en los modelos.
Resultado

4. EDA

4.1 Descripción del conjunto de datos:

Para poder realizar el modelo más adecuado, nuestro primer paso fue seleccionar qué datasets eran los necesarios para poder entrenar el modelo adecuadamente y unificarlos en un dataset de datos de entrenamiento. De los datasets disponibles en la empresa, seleccionamos:

Alumnos.csv (752 registros iniciales): dataset que contiene la información de todos los alumnos de la organización, con propiedades que nos permiten poder identificar patrones posteriormente de por qué convierte o no un lead. Las propiedades seleccionadas fueron:

Nivel de estudios: para poder acceder a nuestra formación deben tener mínimo el grado universitario. Si que es verdad que tal vez hay un mayor patrón de perfiles de doctorado o grado y máster que de grado y por eso añadimos esta propiedad al modelo.
Especialización (carrera estudiada): queremos identificar si hay un patrón concreto asociado a un aumento de la probabilidad de conversión debido a haber estudiado una carrera en concreto.

Nivel de inglés: para acceder a las formaciones de la empresa mínimo deben tener un B2

Comunidad autónoma de residencia: queremos validar si influye el lugar de residencia en la toma de decisiones para acceder a nuestra formación ya que las oportunidades de prácticas mayoritarias están en Madrid, Barcelona y Valencia.

Rango de edad: esta propiedad nos permite saber con qué rango de edad están accediendo al máster.

Canal de captación: esta propiedad nos permite identificar si existen canales más rentables debido a su alta conversión que otros.

Contactos_B2C.csv (1302 registros iniciales) y Negocios_B2C.csv (6399 registros iniciales): seleccionamos estos datasets con el objetivo de obtener los potenciales alumnos del funnel del CRM que utiliza la empresa (Hubspot) que no convirtieron. Para ello teníamos que conseguir filtrar los leads que estuviesen en la etapa de negocio del pipeline de B2C llamada "No volver a contactar" ya que podemos considerarlos como leads perdidos o no convertidos. Para eso, tuvimos que trabajar con estos dos datasets:

Contactos B2C. csv: dataset con la información de todos los contactos presentes en el funnel de ventas, incluyendo todas las propiedades de: nivel de estudios, especialización, nivel de inglés, rango de edad, comunidad autónoma y canal de captación. Seleccionamos las mismas propiedades que en el csv de Alumnos_convertidos para poder trabajar desde las mismas variables para el modelo.

Negocios_B2C.csv: dataset con la información de todos los negocios asociados a los contactos del dataset Contactos_B2C, presentes en el funnel de ventas del CRM en uso que es Hubspot. Hemos descargado estos dos datasets por separado porque el CRM lo descarga así y cada una tiene información que nos interesa. Contactos_B2C contiene las propiedades de los contactos como: edad, comunidad, especialización, etc. y Negocios B2C contiene la etapa de negocio en la que se encuentra el potencial alumno, concretamente nos indica aquellos que están en "No volver a contactar".

4.2 Limpieza y Transformación de los Datos

Antes de proceder al análisis exploratorio de los datos y la posterior construcción del modelo predictivo, fue necesario realizar un proceso de limpieza y transformación de los

datos para conseguir el dataset de datos de entrenamiento conformado por los registros de Alumnos_convertidos y los registros de la fusión de los datasets Contactos_B2C y Negocios_B2C filtrados por la etapa de "No volver a contactar". Para ello, procedimos a través del lenguaje Python en Jupyter Notebook:

Limpieza de propiedades: este paso inicial se realizó manualmente en el documento Google Sheets revisando cada una de las columnas para ver cuáles nos interesan.

Unificación de propiedades: se revisaron las propiedades existentes de cada una de las bases de datos y si había duplicidades de propiedades las eliminamos. También relacionado con el nombre de las propiedades, se encontraron similitudes entre "correo" e "email" y se procedió a unificarlo. También los registros contenían datos escritos de diferente forma, por ejemplo, en la propiedad de comunidad, y procedimos a unificar cada uno de los nombres de las comunidades autónomas.

Anonimización del dataset: cómo los datos del dataset son pertenecientes a usuarios reales, se decidió anonimizar los datasets, se eliminó la columna de Nombre y Apellidos y se creó un ID único, llamado id_lead.

Limpieza de los valores nulos o imputación de valores faltantes: Aquellas columnas que contenían valores faltantes fueron sometidas a un proceso de imputación, en el cual se completaron los valores nulos con la media de su característica correspondiente. Esto se debe a que cómo había un alto % de valores nulos, consideramos de mayor valor imputar a los valores nulos, valores reales con la distribución existente en los valores no nulos. Se ponderó así para que hubiese una distribución equivalente y poder obtener mejores resultados al poder disponer de más registros en lugar de eliminarlos. En algunos casos, se eliminaron aquellos registros donde la falta de datos era excesiva o no se podía realizar una imputación confiable.

Eliminación de registros duplicados: Algunos contactos y negocios fueron registrados varias veces debido a la naturaleza del sistema CRM. Estos duplicados se eliminaron para

evitar que el modelo entrenara con datos redundantes, lo que podría sesgar los resultados y disminuir la calidad de las predicciones.

Estandarización de los valores de cada una de las propiedades: Los datos originales contenían varias columnas con información categórica que no podía ser utilizada directamente en el modelo. Por ejemplo, variables como el nivel de inglés, canal de captación o comunidad autónoma estaban representadas como cadenas de texto. Para poder utilizarlas en el modelo de machine learning, se llevó a cabo una conversión de estas variables a formato numérico mediante un proceso denominado codificación, donde cada categoría fue reemplazada por un valor numérico.
Obtención del dataset de leads_no_convertidos: para la obtención del dataset de leads_no_convertidos que luego servirá para la consecución del dataset de entrenamiento, se realizaron los siguientes pasos:

Se efectuó una unión *left join* sobre el ID de registro, de modo que se conservaron todas las variables previamente seleccionadas del conjunto Contactos_B2C y se añadieron las columnas "Etapa de negocio" y "Canal de captación" procedentes de Negocios_B2C.

Filtrado por la propiedad de "Etapa de negocio" con el valor: "No volver a contactar". Para la consecución de un dataset que solamente muestre los registros que no han convertido, realizamos este filtrado.
Obtención del dataset de entrenamiento (dataset_final_modelo.csv) a través del join de leads_no_convertidos y alumnos_convertidos. Este dataset unificado y procesado fue el que se utilizó para realizar el análisis exploratorio y para entrenar los modelos predictivos que posteriormente se detallan en el trabajo.

4.3 Desarrollo del análisis exploratorio de datos (EDA)

El Análisis exploratorio de datos (EDA) es un paso fundamental en el proceso de desarrollo de modelos predictivos, ya que permite identificar las variables que son más relevantes para la predicción de una determinada salida. En este caso, nuestro objetivo es predecir la conversión de leads en alumnos mediante un modelo de regresión logística. A través del

EDA, analizamos distintas distribuciones de variables, correlaciones y relaciones entre características que influyen directamente en la calidad y desempeño del modelo predictivo. A continuación, se presentan los resultados de los gráficos obtenidos, sus conclusiones y cómo estas pueden ser incorporadas en el modelo de regresión logística.

Para este análisis, se utilizó el dataset final de entrenamiento, denominado dataset_final_modelo.csv, el cual contiene un total de 1278 registros. Este conjunto de datos integra información de los leads, con variables que abarcan tanto características personales como información asociada a la interacción de los leads con el funnel de ventas.

4.3.1. Descripción general del dataset

El dataset utilizado contiene las siguientes propiedades clave:

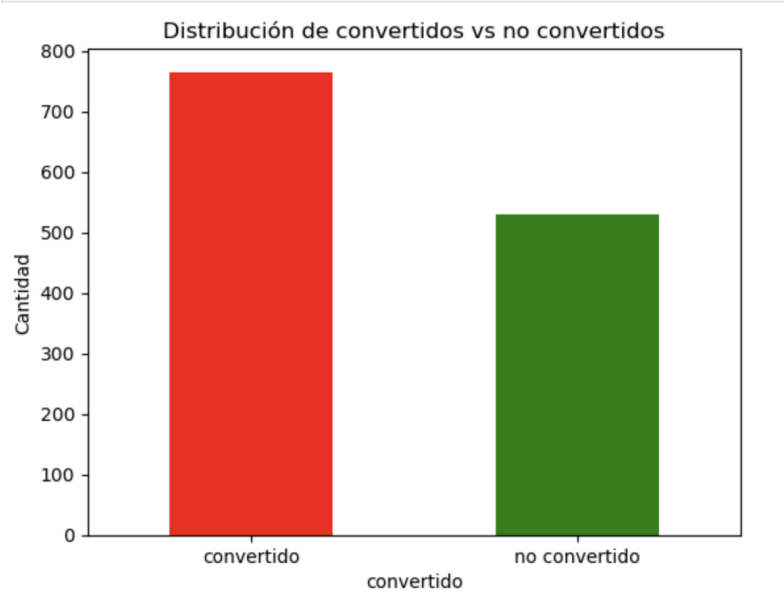
- **id_lead:** identificador único para cada lead.
- **canal de captación y canal de captación (num):** Descripción y representación numérica del canal a través del cual el lead fue adquirido (por ejemplo, Facebook Ads, Google Ads, etc.).
- **edad y rango edad**: Edad del lead y su rango etario.
- **especialización**: Área de estudios o profesión de cada lead.
- **nivel de estudios y nivel de inglés:** Clasificación del nivel educativo y el dominio del inglés de cada lead.
- **id de registro y etapa del negocio:** Identificador único para cada registro y la fase del funnel de ventas en la que se encuentra el lead.
- **comunidad autónoma:** Ubicación geográfica del lead dentro de España.
- **convertido y convertido_bin**: Indicador binario que señala si el lead se convirtió en alumno o no.

Este dataset proporciona una visión completa de cada lead, lo cual es esencial para comprender las dinámicas detrás de la conversión y, en última instancia, para predecir el comportamiento de otros leads.

4.3.2. Distribución de la Variable "Convertido"

El primer análisis que realizamos fue la distribución de la variable convertido, que indica si un lead se convirtió o no en alumno del máster. A través de un gráfico de barras, se visualizó la cantidad de leads que fueron clasificados como convertidos y no convertidos.

Gráfico 1: Distribución de convertidos vs no convertidos

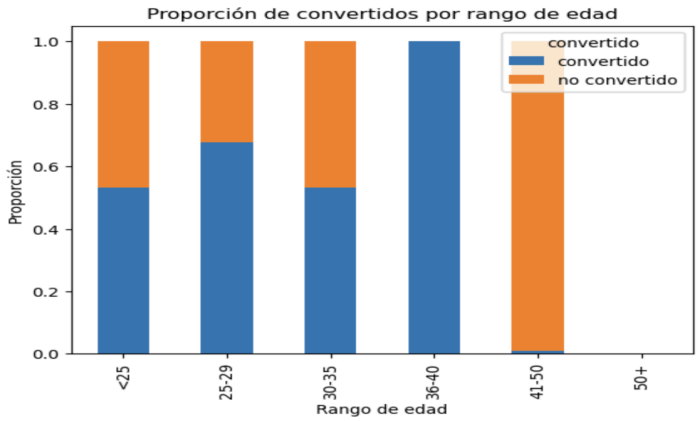


En este gráfico, se observa una distribución ligeramente desbalanceada, con un número menor de leads no convertidos (en verde) en comparación con los convertidos (en rojo). Esto puede generar que el modelo tienda más a predecir valores convertidos que no convertidos ya que no tiene la misma cantidad de información para trabajar.

4.3.3. Análisis por rango de edad

Una de las preguntas clave en este análisis fue: ¿Qué grupos de edad tienen más probabilidades de convertirse en alumnos? Para responder a esta pregunta, se realizó un gráfico que muestra la proporción de convertidos por rango de edad.

Gráfico 2: Proporción de convertidos por rango de edad



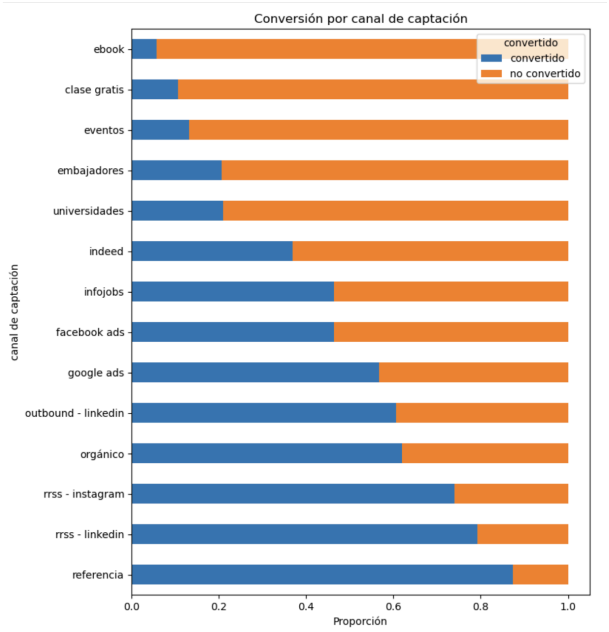
En este gráfico, se observa que, en general, los grupos de edad más jóvenes (menos de 25 años), los de entre 25 y 29 años y los de 30 a 35, presentan una mayor proporción de conversiones, lo cual es coherente ya que las formaciones ofertadas por la empresa están eminentemente dirigidas a perfiles recién egresados o que hayan defendido la tesis recientemente. Sin embargo, es importante notar que, a pesar de que estos grupos tienen más conversiones, la proporción de no convertidos en estos rangos de edad sigue siendo significativa. Por otro lado, los rangos de edad mayores (41-50 a 50+) muestran una proporción de conversión más baja. Esta información es útil, ya que puede ayudar a personalizar estrategias de marketing dirigidas a los rangos etarios con mayor potencial de conversión. En el caso de la tasa de conversión de 36 a 40, al comentarlo con la empresa, parece que hay una alta conversión, pero un bajo número de perfiles con interés en realizar esta formación.

4.4 Análisis por canal de captación

Otro análisis importante fue determinar qué canales de captación son más efectivos para la conversión de leads. Utilizamos un gráfico de barras apiladas para mostrar la conversión

por canal de captación, donde cada barra representa un canal específico y se apilan las proporciones de convertidos y no convertidos.

Gráfico 3: Conversión por canal de captación

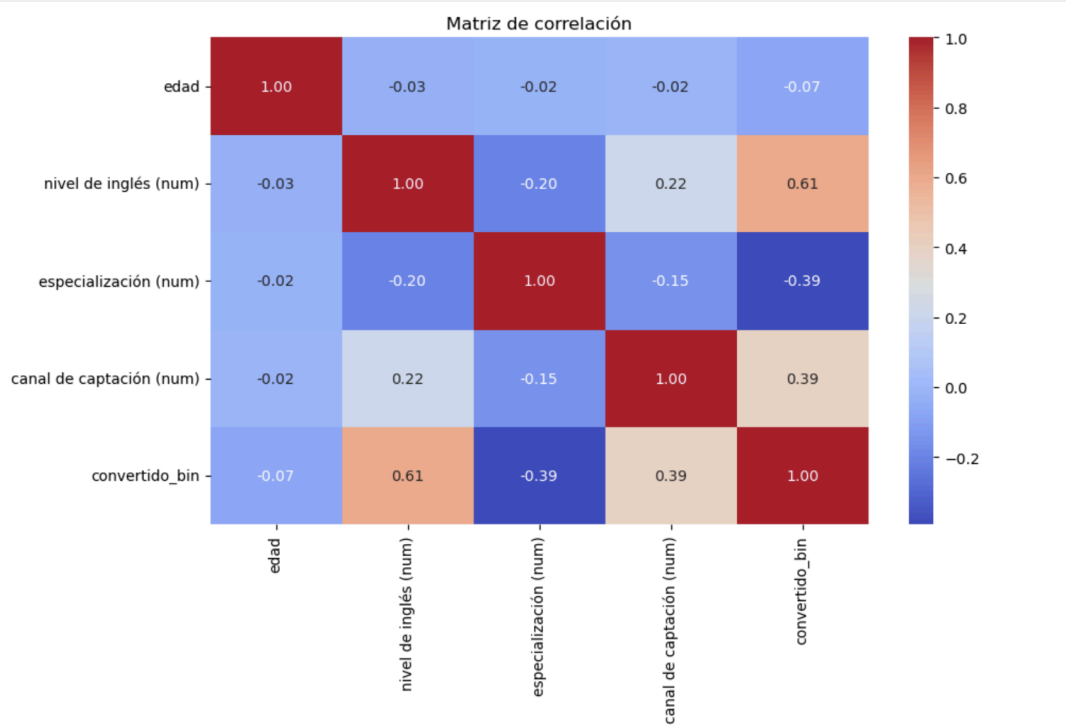


Los resultados muestran que los canales de referencias, redes sociales orgánicas (LinkedIn e Instagram), orgánico (web), estrategias outbound y Paid media (Google ads y Facebook ads) tienen una alta proporción de conversión, lo que sugiere que los leads que provienen de estos canales tienen una mayor probabilidad de convertirse en alumnos. En contraste, canales como ebooks informativos enviados a través del email, CTAs como clases gratis o la asistencia a eventos presenciales muestran una proporción significativamente mayor de leads no convertidos. Este tipo de análisis es esencial para ajustar las estrategias de captación de leads, enfocándose en los canales que generan un mayor retorno de inversión en términos de conversión.

4.5. Matriz de correlación

Para comprender las relaciones entre las variables numéricas del dataset, se generó una matriz de correlación. Este análisis reveló cómo se relacionan las variables clave, como edad, nivel de inglés, especialización y canal de captación, con la variable objetivo convertido_bin, que indica si el lead se convirtió o no.

Gráfico 4: Matriz de correlación



En la matriz de correlación, se observa que la variable nivel de inglés tiene una correlación positiva significativa con la variable convertido_bin (0.61), lo que sugiere que los leads con un mejor nivel de inglés tienen más probabilidades de convertirse en alumnos. De manera similar, la especialización también muestra una correlación moderada con la conversión (0.39), lo que indica que ciertas especializaciones podrían ser más atractivas para los programas ofrecidos por la institución. Las variables edad y canal de captación tienen correlaciones más débiles con la conversión, lo que sugiere que, aunque estas variables influyen, no son tan determinantes como el nivel de inglés o la especialización.

4.6. Funnel de porcentaje (%) de conversión por característica del lead.

El siguiente tratamiento de la información, nos brinda una fotografía sobre la distribución de las variables y los atributos que predominan entre los leads que finalmente convierten en alumnos, más no explica por sí sólo la causalidad, ni la influencia o peso que cada una de estas características tiene en la conversión.

Para ello construimos dos piezas complementarias:

Tabla de modas

Identifica, para cada característica clave (canal, región, especialización, etc.), el valor que más se repite en todo el conjunto de leads.

Mide cuántos de los 765 leads que acaban en la etapa *Alumno* (100 %) comparten ese valor y qué porcentaje suponen dentro de ese universo convertido.

Embudo (funnel) jerárquico

Representa visualmente cómo los leads se van "filtrando" a medida que atraviesan las características dominantes desde la base todos pasando a canal referencia →Comunidad Madrid → especialidad Farmacia, etc.
Prioriza atributos de repetición en la conversión.
Permite ver de un vistazo dónde se produce la mayor reducción de volumen y, por tanto, dónde podría haber mayor margen de optimización.
Detecta cuellos de botella
Permite identificar la audiencia objetivo.
Permite identificar riesgos y oportunidades.

A continuación, se muestra la tabla base para la generación del funnel descriptivo.

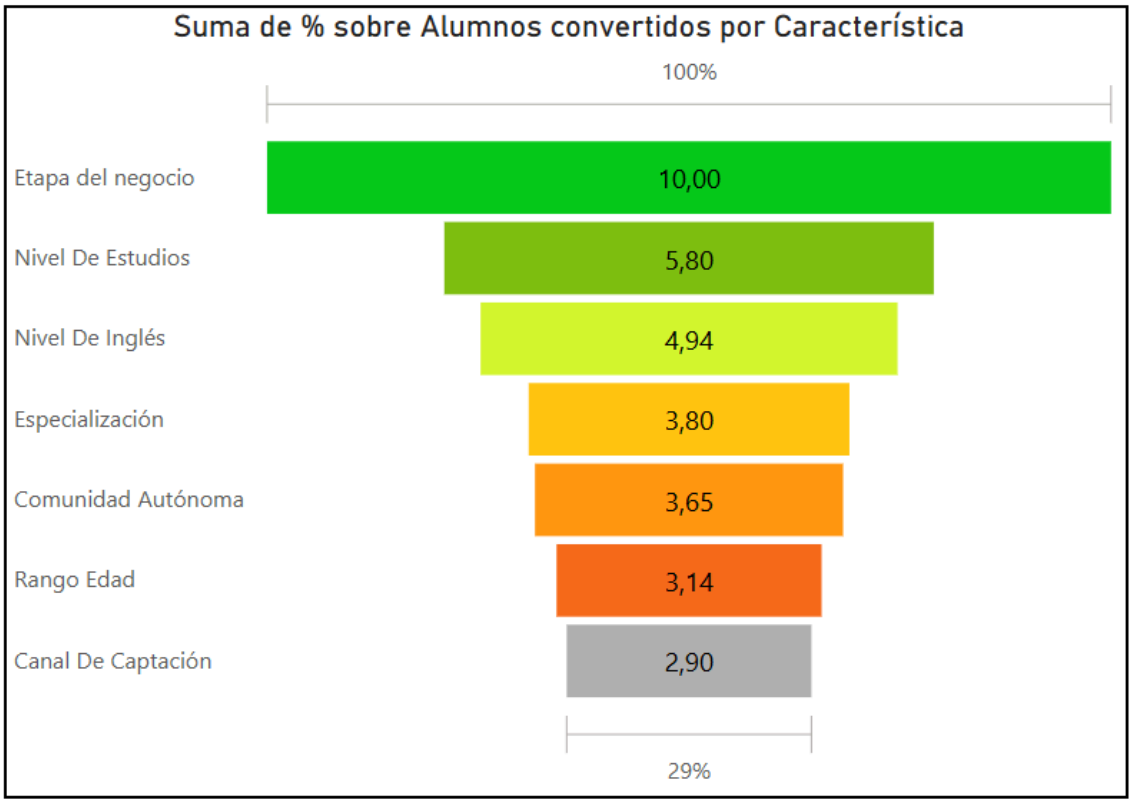
Característica	Valor más frecuente	Repeticiones	Convertidos con ese valor	% sobre Alumnos convertidos
Etapa del negocio	Alumno	765	765	100%
Rango de edad	30 – 35 años	452	240	31.4 %
Canal de captación	Referencia	254	222	29.0 %
Comunidad autónoma	Comunidad de Madrid	311	279	36.5 %
Especialización	Farmacia	291	291	38.0 %
Nivel de estudios	Grado universitario	444	444	58.0 %
Nivel de inglés	B2	479	378	49.4 %

Total leads contactados: 1295

Resumen general de la tabla:

Número de repeticiones: moda en los 1.295 leads.
Número de leads convertidos con el valor más frecuente valor: leads que se convirtieron (Alumnos) y comparten una misma característica.
Alumnos convertidos: 765 alumnos.

Gráfico 5: Funnel de porcentaje (%) de conversión por característica del lead.



Por moda o valor más repetido, el perfil estimado del alumno convertido sería un profesional de entre 30 a 35 años, egresado universitario, con B2 de inglés, especializado en Farmacia, residente (o con IP) en Madrid y captado por recomendación (referencia).

Limitaciones

La columna rango edad presenta un solo valor dominante (30-35).
El funnel muestra la distribución, pero no la causalidad; podría sería necesario en adelante generar un test tipo (A/B) para validar ajustes en el proceso.

4.7. Conclusiones del EDA

El análisis exploratorio de datos reveló información valiosa sobre las características de los leads y los factores que influyen en su conversión. Las variables que tuvieron un mayor impacto fueron nivel de estudios (egresado universitario), nivel de inglés (B2) y especialización (Farmacia), con una correlación positiva significativa con la conversión. Además, se identificó que el canal de captación por referencias, tienen un rendimiento superior en términos de conversión, lo que sugiere que las estrategias de marketing podrían beneficiarse al centrarse más en estos canales.

En términos de preparación para el modelado predictivo, es crucial centrarse en las variables con mayor correlación con la conversión, como nivel de inglés, especialización y canal de captación, ya que estas parecen ser los principales predictores de éxito en la conversión de leads. Este análisis de datos exploratorio ha proporcionado una base sólida para la siguiente etapa de desarrollo del modelo, lo que permitirá optimizar los esfuerzos de marketing y ventas de la institución.

5. Resultados

5.1 Desempeño del modelo de predicción de conversión.

Teniendo como origen los diferentes datos suministrados por la empresa Pharma Business Education, relacionados al proceso B2C, se ejecutaron los pasos descritos en la metodología, iniciando con el análisis EDA del conjunto de datos, lo que permitió su comprensión, preparación y transformación. Se logró obtener en este proceso, un dataset integrado, sincerado, estandarizado y estructurado que sirvió de base para la ejecución del modelo de predicción.

Para definir el modelo a utilizar, se consideraron las necesidades del cliente, base sobre la cual se identificó que la variable objetivo Y sujeta a predicción (target) es la conversión o no conversión del lead (usuarios) a partir del análisis de variables influyentes o no X (features), se escogió en consecuencia, aplicar un modelo de Regresión Logística, porque es una opción predictiva que funciona bien cuando el objetivo es clasificar en dos clases (clasificación binaria) en este caso (convierte o no convierte).

La elección de este enfoque se justifica en la creciente aplicación de modelos predictivos en contextos educativos y de gestión institucional, donde la identificación anticipada de usuarios con alta probabilidad de conversión permite optimizar estrategias de marketing, admisión y retención (Morales et al., 2021; Romero & Ventura, 2020).

Se implementó este modelo de aprendizaje automático supervisado, aplicado a un conjunto de datos previamente procesado:

Regresión Logística (Logistic Regression) bajo la configuración: imputación de valores faltantes mediante la función SimpleImputer (strategy = mean) de sklearn.impute.

La elección de este modelo responde a varios criterios metodológicos:

Permiten establecer una línea base de rendimiento con algoritmos de complejidad y coste computacional moderado.

Es interpretable, lo cual es deseable en el contexto donde se requiere justificar las decisiones a partir de los modelos (Molnar, 2022).

La implementación es robusta y está bien soportada en bibliotecas estándar como Scikit-Learn (Pedregosa et al., 2011).

Para valorar el rendimiento del modelo, se han utilizado las siguientes métricas, adecuadas para problemas de clasificación binaria:

Accuracy (precisión global): proporción de predicciones correctas sobre el total.

Precisión, Recall y F1-score por clase.

Precisión: proporción de verdaderos positivos entre todos los casos clasificados como positivos.

Recall (sensibilidad): proporción de verdaderos positivos entre todos los positivos reales.

F1-score: media armónica entre precisión y recall, útil cuando existe desbalance de clases.

Estas métricas se han calculado tanto en el conjunto de entrenamiento como en el conjunto de prueba, permitiendo identificar posibles signos de sobreajuste o bajo poder de generalización.

5.2 Interpretación y visualización de resultados luego de aplicar modelo.

5.2.1 Interpretación de resultados

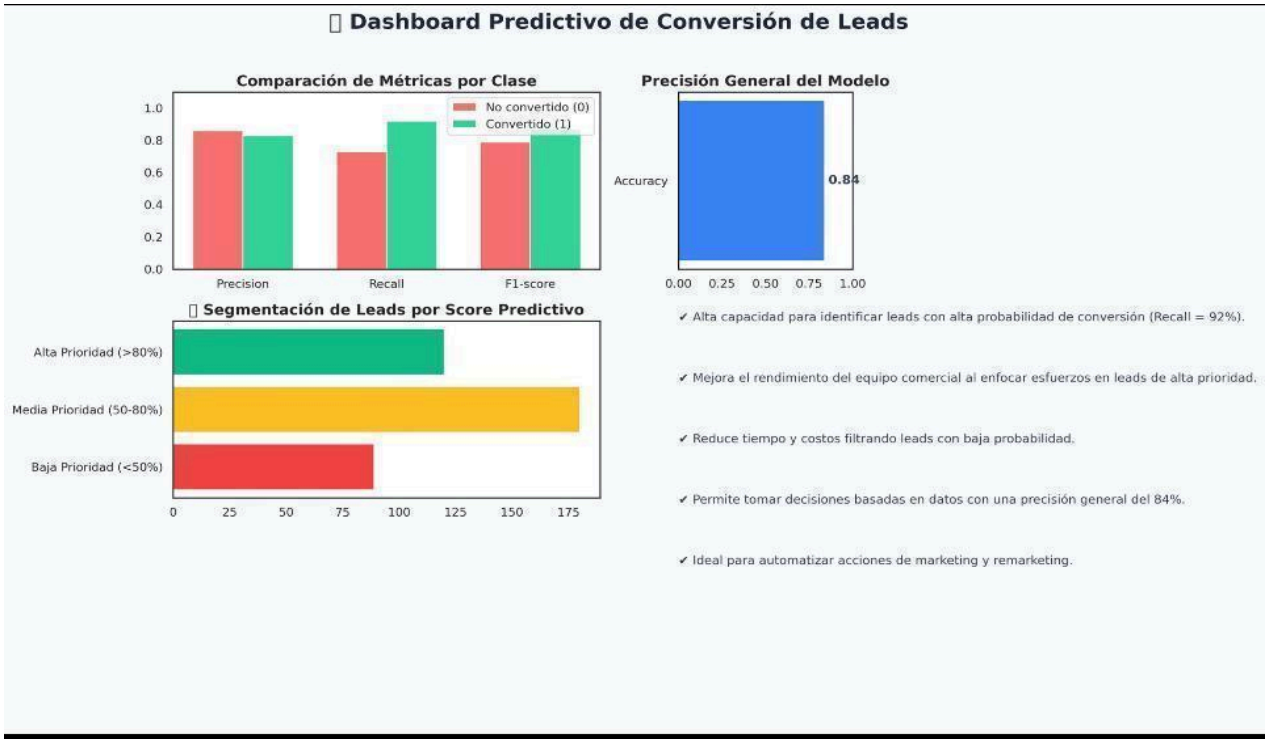
Precisión general F1 Score (84%): el modelo acierta en el 84% de los casos al clasificar leads cómo convertidos o no. Esto representa una tasa de éxito sólida para la toma de decisiones automatizadas o semiautomatizadas. Equilibrio sólido entre detección y precisión.

Precisión (0.86) para clase 0: cuando predice que alguien no se va a convertir, Clase 0 - No convertido (negativos) acierta el 86% de las veces. Esto es útil para evitar insistencias innecesarias, filtrar leads fríos y optimizar el esfuerzo de los equipos de admisión o ventas.

Recall (0.73) para clase 0: sin embargo, solo detecta el 73% de los casos reales de no conversión. Está dejando escapar un 27%.

Precisión (0.83): cuando predice que alguien se va a convertir Clase 1 - Convertido (positivos) acierta el 83% de las veces.

Recall (0.92) para clase 1: detecta el 92% de todos los leads realmente convertidos, por lo tanto, es ideal para anticipar ventas o matrículas, priorizar acciones sobre leads valiosos y diseñar campañas de retargeting, logrando aumentar la eficiencia comercial y reduce el coste de oportunidad de perder alumnos potenciales.



5.2.2 Visualización de resultados una vez aplicado el modelo.

	precision	recall	f1-score	support
0	0.86	0.73	0.79	159
1	0.83	0.92	0.87	230
accuracy			0.84	389
macro avg	0.84	0.82	0.83	389
weighted avg	0.84	0.84	0.84	389

	feature	coeficiente	impacto_absoluto
2	nivel_ingles_num	1.758257	1.758257
1	rango_edad_num	1.373626	1.373626
0	edad	-1.243272	1.243272
3	especializacion_num	-1.231004	1.231004
4	nivel_estudios_num	0.196274	0.196274

En el modelo de regresión logística, el nivel de inglés emerge como el factor más determinante: por cada punto que aumenta la variable nivel_ingles_num, la probabilidad de conversión se multiplica en torno 5,8 veces), lo que lo convierte en el impulsor más fuerte del éxito. Le sigue el rango de edad (coef. 1,37), indicando que ciertos tramos etarios según la codificación numérica empleada también incrementan notablemente la probabilidad de convertirse.

En contraste, la edad absoluta y la especialización presentan coeficientes negativos (-1,24 y -1,23); a medida que sus valores aumentan, la propensión a convertir disminuye en proporción similar, lo que sugiere que perfiles más veteranos o especialidades alejadas de la moda del dataset reducen la eficacia. Por último, el nivel de estudios muestra un efecto positivo, pero mucho más débil (0,20), aportando un aumento marginal en la probabilidad de conversión. En síntesis, optimizar la captación hacia leads con mayor dominio de inglés y dentro de los rangos de edad más favorables tendrá el mayor impacto, mientras que controlar la mezcla de especializaciones y el sesgo de edad ayudará a evitar caídas en la tasa de conversión.

6. Discusión

6.1 Análisis crítico de los hallazgos

Los modelos predictivos pueden servir como soporte para la mejora de procesos institucionales:

- Equipos de admisión: pueden priorizar candidatos con mayor potencial de conversión, racionalizando el tiempo y los recursos invertidos.
- Desarrollo docente: al analizar qué perfiles responden mejor a ciertas metodologías, se pueden ajustar los enfoques pedagógicos de forma más dirigida.

De esta forma, el análisis predictivo trasciende la conversión inicial y se convierte en un instrumento de inteligencia institucional, alineado con la misión educativa de la organización.

A pesar del buen rendimiento, es necesario contextualizar los resultados dentro de un marco crítico y estratégico:

- Dependencia de la calidad del dataset: el modelo se alimenta de variables históricas de contacto y comportamiento de leads. Si la calidad o integridad de estos datos varía (por ejemplo, fechas erróneas, registros incompletos o sesgos por canal), la precisión futura del modelo puede verse afectada. Se recomienda implementar procesos de validación y limpieza continua de los datos.

- Foco en Recall puede generar presión operativa, aunque tener un alto recall para los leads convertidos es positivo, también implica que más leads serán clasificados como “prioritarios”, lo que puede sobrecargar al equipo comercial si no se segmenta correctamente. Aquí es fundamental combinar el modelo con reglas de negocio y capacidad operativa.

- Generalización vs. Sobreajuste: si bien los resultados son prometedores en el conjunto de prueba, debe vigilarse la posibilidad de sobreajuste (overfitting). Se recomienda implementar técnicas como la validación cruzada, regularización y

evaluación periódica en producción para asegurar que el modelo mantiene su rendimiento con nuevos leads.

- Interpretabilidad del modelo para la toma de decisiones: la regresión logística tiene la ventaja de ser interpretable frente a modelos más complejos como los árboles de decisión o redes neuronales. Esto permite que los equipos de negocio entiendan el "por qué" detrás de una predicción, lo que fortalece la confianza en la herramienta y facilita su adopción.

6.2 Limitaciones del estudio

El modelo debe usarse como una herramienta de apoyo, no como sustituto de la toma de decisiones humanas. Algunas formas de integración incluyen:

- Soporte a la admisión personalizada según perfiles predictivos.
- Adaptación de planes de acogida y acciones de tutoría en función de las necesidades anticipadas.
- Alimentación de sistemas de seguimiento o retención (early warning systems).

La implementación debe ir acompañada de una política clara sobre uso de datos y gobernanza institucional.

Para mantener la vigencia del modelo, se sugiere:

- Crear un flujo automatizado de actualización de datos (ETL) que mantenga el sistema actualizado periódicamente.
- Establecer un entorno reproducible, basado en herramientas como Jupyter Notebooks, Docker o ML flow, que permita auditar, replicar o transferir el modelo a otros equipos o áreas institucionales.

Las variables utilizadas en el modelo provienen en su mayoría de categorías sociodemográficas (edad, nacionalidad), académicas (titulación previa) y lingüísticas (nivel de idioma). Si bien estas dimensiones son relevantes, no se incorporan variables de comportamiento, historial digital o motivación, que suelen ser altamente predictivas en problemas de conversión de usuarios (Baars et al., 2020).

6.3 Implicaciones para la gestión de talento y marketing educativo

Gestión del talento: entendida como el proceso integral de captación, admisión, orientación y retención de estudiantes.

Marketing educativo: vinculado a la planificación, segmentación y ejecución de campañas para atraer y convertir potenciales estudiantes.

La regresión logística y otros modelos han revelado que ciertas variables sociodemográficas (como la edad, el nivel de inglés, la formación previa o el país de origen) influyen en la probabilidad de conversión.

En concreto, este enfoque permite:

- Identificar perfiles con baja probabilidad de conversión, para ofrecer apoyo adicional (por ejemplo, tutorías reforzadas, sesiones informativas, orientación temprana).
- Diseñar itinerarios de onboarding personalizados, ajustados al perfil de ingreso de cada estudiante.
- Asignar recursos de forma más eficiente, priorizando casos con mayor riesgo de no conversión o abandono inicial.

Este tipo de segmentación predictiva se ha mostrado eficaz en estudios previos sobre retención universitaria y predicción de abandono (Arnold & Pistilli, 2012; Jayaprakash et al., 2014).

7. Conclusiones

7.1 Conclusiones generales

El estudio ha abordado el uso de la regresión logística como técnica predictiva para anticipar la probabilidad de conversión de leads (usuarios) conforme a sus características, donde dicha conversión representa un comportamiento deseado, en este caso la matriculación a los diferentes programas de estudio que ofrece la empresa. Se seleccionó un modelo clásico y bien establecido en la literatura técnica, aplicado sobre un conjunto de datos numéricos previamente tratados y escalados.

Al aplicarse el modelo de regresión logística con y sin imputación de valores faltantes, se observó que la calidad predictiva se mantuvo estable, aunque se destaca una ligera mejora cuando los valores nulos fueron tratados tal como se describe en el apartado 4.2.

Ventajas observadas:

- Alta interpretabilidad: los coeficientes del modelo permiten estimar la dirección e intensidad del efecto de cada variable predictora sobre la probabilidad de conversión.
- Robustez contra el sobreajuste: los resultados en train y test fueron consistentes, lo cual sugiere una buena generalización.
- Simplitud computacional: la eficiencia del modelo lo hace ideal para escenarios con recursos limitados o necesidad de despliegue rápido.

Los resultados reflejan que el modelo predictivo permite optimizar los procesos de captación y conversión. Su implementación no solo puede mejorar la asignación de recursos comerciales, sino también automatizar campañas más inteligentes, personalizadas y oportunas. No obstante, su uso debe ir acompañado de un enfoque de mejora continua, gobernanza del dato y alineamiento con las capacidades reales del equipo comercial y de marketing.

La integración de los modelos predictivos con los sistemas de gestión del marketing institucional puede facilitar:

- El análisis del ROI por campaña y segmento, en este sentido podría aumentar las campañas hasta un 30% y reducir el coste de adquisición por cliente.
- La trazabilidad del embudo de conversión, desde el primer contacto hasta la matrícula efectiva.
- La identificación de cuellos de botella, donde usuarios con alta probabilidad no completan la conversión por causas operativas o comunicativas.

Estas capacidades permiten establecer un círculo virtuoso de análisis, acción y retroalimentación, que fortalece la toma de decisiones estratégicas y optimiza el uso de recursos institucionales.

Antes de aplicar el modelo en decisiones sensibles, es indispensable:

- Realizar auditorías de equidad algorítmica, analizando el rendimiento por subgrupo (edad, género, nacionalidad).
- Aplicar métricas de fairness como Demographic Parity o Equal Opportunity (Barrocas et al., 2019).
- Establecer procedimientos de revisión ante resultados adversos.

7.2 Contribuciones al conocimiento aplicado y a los ODS (Objetivos de Desarrollo Sostenible)

El análisis ha constituido un ejercicio riguroso en la implementación de modelos clásicos de predicción, existen diversas áreas de mejora técnica y metodológica. La incorporación de validación cruzada, la expansión de modelos, el refinamiento de las variables predictoras y el análisis segmentado son líneas prioritarias para garantizar:

- Mayor precisión predictiva.
- Mayor estabilidad y reproducibilidad.
- Mayor equidad y utilidad práctica del modelo en contextos institucionales reales.

Los perfiles con mayor probabilidad de conversión deben emplearse para:

- Diseñar campañas dirigidas por segmento.
- Asignar presupuestos publicitarios de forma más eficaz.

- Identificar canales y mensajes con mayor rendimiento (marketing multicanal basado en datos).

A su vez, se deben establecer dashboards de control para monitorizar la evolución de las campañas, las tasas de conversión y la eficiencia por segmento. Por otro lado, el modelo de predicción aplicado contribuye a la Sostenibilidad, partiendo de las siguientes conclusiones:

Sostenibilidad Económica: uso más eficiente de los recursos. El modelo permite identificar con antelación qué leads tienen mayor probabilidad de conversión, lo que se traduce en:

- Reducción de costos operativos en llamadas, correos y seguimientos improductivos.
- Asignación eficiente de los equipos humanos, que se enfocan solo en leads viables.
- Optimización del presupuesto de marketing, al evitar campañas amplias y poco focalizadas.

Esto genera un ahorro tangible y sostenido en el tiempo, haciendo que el modelo sea financieramente sostenible para la organización.

Sostenibilidad Ambiental: disminución del uso innecesario de recursos físicos

Aunque el modelo actúa en un entorno digital, su implementación puede tener impactos indirectos en el medioambiente:

- Menor impresión de material promocional innecesario al focalizar mejor la comunicación.
- Reducción de desplazamientos físicos de asesores comerciales al priorizar leads viables.
- Fomenta un modelo paperless y digitalizado en procesos comerciales y académicos.

La eficiencia digital reduce la huella de carbono indirecta, alineándose con prácticas sostenibles.

Sostenibilidad Social: mejora de la experiencia y acceso del usuario

El modelo también aporta al bienestar del usuario final (el alumno interesado), al permitir:

- Procesos de admisión más ágiles y personalizados, evitando saturación de contactos.
- Mayor equidad, al detectar de forma objetiva perfiles con potencial real, sin sesgos humanos.
- Mejor aprovechamiento del tiempo del lead, lo cual es respetuoso con su experiencia.

Esto mejora la relación de confianza y transparencia entre institución y alumno, pilar clave en la sostenibilidad social.

7.3 Posibles líneas de continuidad del proyecto

Las recomendaciones aquí expuestas proporcionan una hoja de ruta clara y viable para escalar el modelo predictivo desarrollado hacia un entorno de aplicación real, ético y sostenible. Invertir en estas acciones no solo mejorará el rendimiento técnico, sino también el impacto institucional y la aceptación del sistema por parte de la comunidad educativa.

La incorporación de modelos predictivos en la gestión de talento y el marketing educativo ofrece beneficios concretos y cuantificables:

- Mejora la eficiencia en la admisión.
- Facilita una orientación más personalizada.
- Optimiza el uso de recursos.
- Aumenta el rendimiento de las campañas.

Sin embargo, su implementación debe estar guiada por principios éticos, transparencia y gobernanza institucional. No se trata únicamente de predecir quién se convierte, sino de entender por qué, y actuar con responsabilidad para que cada usuario tenga las mejores condiciones para lograrlo.

El modelo permite identificar patrones de características asociados a la conversión, lo cual puede emplearse para mejorar la segmentación en campañas publicitarias, tanto en canales tradicionales como digitales.

Beneficios de esta aproximación:

- Maximizar el retorno de la inversión (ROI) al focalizar esfuerzos en segmentos con alta propensión a convertir.
- Reducir costes de captación evitando impactar a perfiles con baja afinidad, sin caer en prácticas discriminatorias.
- Aumentar la tasa de conversión a través de estrategias basadas en datos empíricos y no solo en intuiciones de mercado.

Este uso de modelos predictivos para la personalización de campañas es coherente con enfoques de marketing basado en datos (data-driven marketing), ampliamente reconocidos en la literatura especializada (Chaffey & Smith, 2017).

Para mejorar la capacidad predictiva sin sacrificar interpretabilidad, se propone incorporar:

- XGBoost (Chen & Guestrin, 2016): especialmente eficaz para datos estructurados.
- LightGBM (Ke et al., 2017): optimizado para grandes volúmenes
- CatBoost (Dorogush et al., 2018): útil para manejar variables categóricas sin codificación previa.

Estos algoritmos pueden combinarse con herramientas de interpretabilidad como:

- SHAP (SHapley Additive exPlanations): basado en teoría de juegos para explicar predicciones individuales.
- LIME (Local Interpretable Model-Agnostic Explanations): aproximación local para interpretar predicciones complejas.

Para asegurar el uso responsable y eficaz del modelo, se recomienda:

- Capacitar al personal en fundamentos de analítica, interpretación de modelos y toma de decisiones basada en datos.
- Fomentar una cultura institucional de datos, donde se valore la evidencia, pero se mantenga el enfoque humano.

8. Referencias bibliográficas

Eckerson, W. W. (2010). *Performance Dashboards: Measuring, Monitoring, and Managing Your Business*. John Wiley & Sons.

J. Martel. (2018, agosto 8). Business analytics, qué es y en qué consiste. ¡TE LO EXPLICAMOS! *ITELLIGENT*. https://itelligent.es/que-es-business-analytics/

Laursen, G. H. N., & Thorlund, J. (2016). *Business Analytics for Managers: Taking Business Intelligence Beyond Reporting*. John Wiley & Sons.

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

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: The management revolution. *Harvard business review*, *90*(10), 60-68.

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

Shmueli, G., Bruce, P. C., Gedeck, P., & Patel, N. R. (2019). *Data Mining for Business Analytics: Concepts, Techniques and Applications in Python*. John Wiley & Sons.

CANO, I. R. (2018, julio 4). *Historia y evolución de la analítica de negocio*. Viewnext.com. https://www.viewnext.com/historia-y-evolucion-de-la-analitica-de-negocio/

Chen, M. (2024, septiembre 23). *Big Data, grandes posibilidades: Cómo extraer el máximo valor*. https://www.oracle.com/es/big-data/what-is-big-data/

EY España. (2024, enero 24). *El poder de los datos: Cómo obtener una ventaja competitiva con una estrategia data-driven*. https://www.ey.com/es_es/the-cfo-agenda/el-poder-datos-como-obtener-ventaja-competitiva-con-estrategia-data-driven

Gómez, C. A. O. (2021, octubre 12). El Futuro de Business Intelligence (BI) y Business Analytics (BA). *Revista Empresarial & Laboral*. https://revistaempresarial.com/tecnologia/inteligencia-de-negocios/el-futuro-de-business-intelligence-bi-y-business-analytics-ba/

Google Inc. (2025). *¿Qué es el cloud computing? Google Cloud*. Google Cloud. https://cloud.google.com/learn/what-is-cloud-computing

Ortega, C. (2024, marzo 19). *Cultura de datos: Qué es, importancia y cómo crearla. QuestionPro*.
https://www.questionpro.com/blog/es/cultura-de-datos/
Ortega, K. (2023, enero 9). *¿Cuáles son los tipos de análisis de datos que existen?* Saint Leo University.
https://worldcampus.saintleo.edu/blog/cuantos-tipos-de-analisis-de-datos-hay-cuales-son-los-tipos-de-analisis-de-datos
García, A. (2018). *Analítica de datos para la toma de decisiones empresariales*. Editorial UOC.
https://www.editorialuoc.com/analitica-de-datos-para-la-toma-de-decisiones-empresariales
Torres, M. (2022). *Gestión ágil de proyectos con Scrum: Prácticas, herramientas y casos de estudio*. UOC.
https://www.editorialuoc.com/gestion-agil-de-proyectos-con-scrum-practicas-herramientas-y-casos-de-estudio
Pérez, L. (2020). *Agilidad y Scrum: Claves para liderar equipos de alto rendimiento. Anaya.*
https://www.anayamultimedia.es/libro/agilidad-y-scrum-claves-para-liderar-equipos-de-alto-rendimiento_121273/
López-Iturriaga, F. J. (2020). *Analítica de datos para la toma de decisiones empresariales: De la analítica descriptiva a la prescriptiva.* Ediciones Pirámide.
https://www.edicionespiramide.es/libro.php?id=6216802
Fernández, A., & García, S. (2020). *Aprendizaje automático: Conceptos y aplicaciones en la empresa.* Editorial UOC.
https://www.editorialuoc.com/aprendizaje-automatico-conceptos-y-aplicaciones-en-la-empresa
Rodríguez, J. (2021). *Analítica descriptiva de datos para la toma de decisiones empresariales*. Editorial UOC.
https://www.editorialuoc.com/analitica-descriptiva-de-datos-para-la-toma-de-decisiones-empresariales
Sánchez, J. (2021). *CRM: Estrategias y sistemas de gestión de relaciones con clientes*. Editorial UOC.

https://www.editorialuoc.com/crm-estrategias-y-sistemas-de-gestion-de-relaciones-con-clientes

Martín, L. (2022). *Productividad y gestión de proyectos con Notion*. Anaya Multimedia. https://www.anayamultimedia.es/libro/productividad-y-gestion-de-proyectos-con-notion_137001/

Pérez, M. (2022). *Herramientas digitales para la analítica de datos.* Editorial UOC. https://www.editorialuoc.com/herramientas-digitales-para-la-analitica-de-datos

Hernández, P. (2022). *Business Intelligence con Power BI: Análisis, modelado y visualización de datos*. Anaya Multimedia. https://www.anayamultimedia.es/libro/business-intelligence-con-power-bi-analisis-modelado-y-visualizacion-de-datos_137465/

Ruiz, C. (2022). *Programación para análisis de datos con Python y Visual Studio Code.* Anaya Multimedia. https://www.anayamultimedia.es/libro/programacion-para-analisis-de-datos-con-python-y-visual-studio-code_137340/

Morales, D. (2022). *Análisis de datos y machine learning con Python y Jupyter Notebook.* Anaya Multimedia. https://www.anayamultimedia.es/libro/analisis-de-datos-y-machine-learning-con-python-y-jupyter-notebook_137510/

Navarro, J. (2023). *Machine Learning práctico con Google Colab y Python.* Marcombo. https://www.marcombo.com/machine-learning-practico-con-google-colab-y-python

López, M. (2023). *Python para análisis de datos: Fundamentos, librerías y aplicaciones prácticas.* Anaya Multimedia. https://www.anayamultimedia.es/libro/python-para-analisis-de-datos-fundamentos-librerias-y-aplicaciones-practicas_137603/

Ortega, S. (2023). *Big Data y almacenamiento en la nube: Google BigQuery y otros sistemas de datos.* Marcombo. https://www.marcombo.com/big-data-y-almacenamiento-en-la-nube-google-bigquery-y-otros-sistemas-de-datos

Ramírez, L. (2024). Aplicaciones de la inteligencia artificial generativa en la investigación y la analítica de datos. UOC.

https://www.editorialuoc.com/aplicaciones-de-la-inteligencia-artificial-generativa-en-la-investigacion-y-la-analitica-de-datos

Vargas, E. (2023). Herramientas digitales colaborativas para la gestión de proyectos y datos. UOC.

https://www.editorialuoc.com/herramientas-digitales-colaborativas-para-la-gestion-de-proyectos-y-datos

Gómez, F. (2023). Aplicaciones de mensajería en la gestión empresarial: Estrategias y buenas prácticas. UOC.

https://www.editorialuoc.com/aplicaciones-de-mensajeria-en-la-gestion-empresarial-estrategias-y-buenas-practicas

Aguiar, E., Ambrose, G. A., Chawla, N. V., Goodrich, V., & Brockman, J. (2015). Engagement vs Performance: Using electronic portfolios to predict first semester engineering student persistence. Journal of Learning Analytics, 2(3), 7–33.

Arnold, K. E., & Pistilli, M. D. (2012). *Course Signals at Purdue: Using learning analytics to increase student success*. Proceedings of the 2nd International Conference on Learning Analytics and Knowledge, 267–270.

Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and Machine Learning: Limitations and Opportunities*. http://fairmlbook.org

Bergstra, J., & Bengio, Y. (2012). *Random search for hyper-parameter optimization*. Journal of Machine Learning Research, 13(1), 281–305.

Berk, R. A. (2008). *Statistical learning from a regression perspective*. Springer.

Chaffey, D., & Smith, P. R. (2017). *Digital marketing excellence: Planning, optimizing and integrating online marketing*. Routledge.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). *SMOTE: Synthetic Minority Over-sampling Technique*. Journal of Artificial Intelligence Research, 16, 321–357.

Chen, T., & GuestrinSiemens, G., & Long, P. (2011). *Penetrating the fog: Analytics in learning and education*. EDUCAUSE Review, 46(5), 30–40.

UNESCO. (2021). *Recommendation on the Ethics of Artificial Intelligence*. https://unesdoc.unesco.org/ark:/48223/pf0000379920

Zheng, A., & Casari, A. (2018). *Feature engineering for machine learning: Principles and techniques for data scientists*. O'Reilly Media.

C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Dorogush, A. V., Ershov, V., & Gulin, A. (2018). *CatBoost: gradient boosting with categorical features support*. arXiv preprint arXiv:1810.11363.

He, H., & García, E. A. (2009). *Learning from imbalanced data*. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263–1284.

Jayaprakash, S. M., Moody, E. W., Lauría, E. J. M., Regan, J. R., & Baron, J. D. (2014). *Early alert of academically at-risk students: An open source analytics initiative*. Journal of Learning Analytics, 1(1), 6–47.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). *LightGBM: A highly efficient gradient boosting decision tree*. Advances in Neural Information Processing Systems, 30.

Kohavi, R. (1995). *A study of cross-validation and bootstrap for accuracy estimation and model selection*. International Joint Conference on Artificial Intelligence, 14(2), 1137–1145.

Lundberg, S. M., & Lee, S. I. (2017). *A unified approach to interpreting model predictions*. Advances in Neural Information Processing Systems, 30.

Molnar, C. (2022). *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/

Peppers, D., & Rogers, M. (2016). *Managing Customer Relationships: A Strategic Framework*. John Wiley & Sons.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). *"Why should I trust you?": Explaining the predictions of any classifier*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Romero, C., & Ventura, S. (2020). *Educational data mining and learning analytics: An updated survey*. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 10(3), e1355

9. Anexos

Código Python Modelo predictivo empleado

[16]:
```python
from sklearn.linear_model import LogisticRegression
from sklearn.impute import SimpleImputer
from sklearn.metrics import classification_report
import numpy as np

# Create an imputer to fill missing values with the mean
imputer = SimpleImputer(strategy='mean')

# Apply imputer to training data
X_train_imputed = imputer.fit_transform(X_train_scaled)

# Apply the same imputer to test data
X_test_imputed = imputer.transform(X_test_scaled)

# Now train the model with the imputed data
modelo = LogisticRegression()
modelo.fit(X_train_imputed, y_train)

# Make predictions
y_pred = modelo.predict(X_test_imputed)
print(classification_report(y_test, y_pred))

# Alternative approach: Drop rows with NaN values
# X_train_no_nan = X_train_scaled.dropna()
# y_train_no_nan = y_train[X_train_scaled.dropna().index]
# X_test_no_nan = X_test_scaled.dropna()
# y_test_no_nan = y_test[X_test_scaled.dropna().index]
```

```
              precision    recall  f1-score   support

           0       0.86      0.73      0.79       159
           1       0.83      0.92      0.87       230

    accuracy                           0.84       389
   macro avg       0.84      0.82      0.83       389
weighted avg       0.84      0.84      0.84       389
```

:
```python
import pandas as pd
import numpy as np

# Obtener los coeficientes del modelo
coeficientes = modelo.coef_[0]
nombres_columnas = X.columns

# Unir en un DataFrame
importancia = pd.DataFrame({
    'feature': nombres_columnas,
    'coeficiente': coeficientes,
    'impacto_absoluto': np.abs(coeficientes)
}).sort_values(by='impacto_absoluto', ascending=False)

print(importancia)
```

```
              feature  coeficiente  impacto_absoluto
2     nivel_ingles_num     1.758257          1.758257
1       rango_edad_num     1.373626          1.373626
0                 edad    -1.243272          1.243272
3  especializacion_num    -1.231004          1.231004
4   nivel_estudios_num     0.196274          0.196274
```

:
```python
df.to_csv("/Users/inesmartinez/Desktop/BBDD para TFM/dataset_enriquecido_para_modelo.csv", index=False)
```