Pontificia Universidad
JAVERIANA
Cali

USO DE IA EN SALESFORCE PARA DESARROLLAR UN MODELO PREDICTIVO DE MATRÍCULA DE ASPIRANTES

José Manuel García López
Miguel Angel Sanchez Paez

Proyecto de grado entregado para obtener el título de
Ingeniería de Sistemas y Computación

Dirigida por
PhD. Luisa Fernanda Rincón Pérez

Pontificia Universidad Javeriana Cali
Facultad de Ingeniería y Ciencias
Ingeniería de Sistemas y Computación
Santiago de Cali
20 de Julio de 2025

Abstract

Since its adoption in 2016, Salesforce has become a key tool for Pontificia Universidad Javeriana Cali, enabling comprehensive relationship management with students and prospective applicants in areas such as outreach, admissions, financial and academic services, and alumni engagement. Despite this consolidation of historical data, the university's use of artificial intelligence (AI) capabilities remains limited when it comes to optimizing processes—particularly student enrollment. This project aims to implement a predictive model that identifies and parametrizes the key variables influencing enrollment decisions, leveraging Salesforce's AI tools to analyze the historical data stored on the platform. The methodology includes data preparation and cleaning, model training, and iterative evaluation to improve accuracy. As a result, a model was developed that effectively anticipates enrollment outcomes, providing strategic insights to support data-driven decision-making and reinforcing the university's commitment to innovation and the use of advanced technologies in educational management.

Keywords: Machine Learning, Artificial intelligence, CRISP-DM methodology, University Enrollment, Big Data, Data Analysis, Predictive Model, Salesforce.

Resumen

Desde su adopción en 2016, Salesforce se ha posicionado como una herramienta fundamental para la Pontificia Universidad Javeriana Cali, facilitando la gestión integral de relaciones con estudiantes y aspirantes en áreas como atracción, admisiones, servicios financieros y académicos, y seguimiento de egresados. A pesar de esta consolidación de datos históricos, el aprovechamiento de las capacidades de inteligencia artificial (IA) por parte de la universidad aún es limitado al momento de optimizar procesos, entre ellos, la captación de nuevos estudiantes. Este proyecto busca implementar un modelo predictivo que identifique y parametrice las variables clave en la decisión de matrícula, aprovechando las herramientas de IA de Salesforce para analizar los datos históricos almacenados en la plataforma. La metodología incluye la preparación y limpieza de los datos, el entrenamiento del modelo y una evaluación iterativa para mejorar su precisión. Con esta iniciativa, se obtuvo un modelo que anticipa de manera efectiva la matrícula, proporcionando conocimientos estratégicos que promuevan la toma de decisiones basada en datos y refuercen el compromiso de la universidad con la innovación y el uso de tecnologías avanzadas en la gestión educativa.

Palabras Clave: Aprendizaje automático, Inteligencia artificial, Metodología CRISP-DM, Matrícula universitaria, Big Data, Análisis de datos, Modelo predictivo, Salesforce.

Índice general

Índice de cuadros

Índice de figuras

Introducción

La transformación digital en las instituciones de educación superior ha permitido una evolución significativa en la manera en que se gestionan los procesos académicos y administrativos. En este contexto, la Pontificia Universidad Javeriana Cali ha incorporado desde 2016 la plataforma Salesforce como su sistema de gestión de relaciones con aspirantes, estudiantes y egresados. Esta herramienta ha sido fundamental para consolidar información clave en procesos como atracción, admisión y seguimiento estudiantil, lo que representa una valiosa fuente de datos para la toma de decisiones estratégicas.

A pesar de contar con un volumen considerable de datos históricos, la universidad aún no ha aprovechado plenamente el potencial de las herramientas de inteligencia artificial (IA) integradas en Salesforce, como Prediction Builder. Estas tecnologías permiten desarrollar modelos capaces de anticipar comportamientos, lo cual resulta particularmente útil en el proceso de matrícula de nuevos estudiantes, donde el tiempo y los recursos disponibles son limitados.

Este trabajo de grado propone el desarrollo e implementación de un modelo predictivo que estime la probabilidad de matrícula de los aspirantes, utilizando los datos almacenados en Salesforce. Para ello, se adopta la metodología CRISP-DM como guía estructural para la comprensión del negocio, el análisis de datos, la implementación y evaluación del modelo. La iniciativa busca ofrecer una herramienta que apoye la toma de decisiones basada en datos y promueva una mayor eficiencia en los procesos de captación, reafirmando el compromiso de la Universidad con la innovación y el uso estratégico de la tecnología.

Capítulo 1

Descripción del Problema

1.1. Planteamiento del Problema

En la Pontificia Universidad Javeriana Cali, el proceso de captación de nuevos estudiantes es un esfuerzo multidisciplinario que involucra a varias áreas, como promoción institucional y mercadeo. En dicho proceso, se busca atraer a los aspirantes y convertirlos en estudiantes matriculados para el próximo periodo académico, a través de diversas campañas, llamadas, correos electrónicos y otros medios de comunicación. Sin embargo, este proceso puede ser complejo y demanda una considerable inversión de tiempo y esfuerzo, lo que dificulta la planificación y gestión tanto del personal como del presupuesto disponible.

Desde el año 2016, la Pontificia Universidad Javeriana Cali ha adoptado Salesforce como su principal plataforma de gestión de relaciones con los clientes (CRM). Esta plataforma ofrece un conjunto de soluciones integradas basadas en la nube para gestionar la relación con los clientes, incluyendo herramientas de ventas, marketing, servicio al cliente y análisis de datos. Con la implementación de Salesforce, la Universidad procura mejorar la eficiencia operativa, optimizar la comunicación y obtener una visión más integral de sus estudiantes y procesos administrativos. Actualmente, la Universidad cuenta con un histórico de datos en Salesforce de más de 200 mil registros que incluyen tanto a los aspirantes que consolidaron el proceso de admisión para ingresar a una nueva cohorte de los programas ofrecidos como a aquellos que no completaron el proceso de matrícula. Esta información representa una gran oportunidad para fortalecer las estrategias de seguimiento y retención estudiantil, especialmente en momentos críticos del proceso de admisión.

Especialmente durante la época de cierre de matrícula, donde el desafío se intensifica: el equipo de promoción institucional se esfuerza por contactar a cada aspirante, ya sea por correo o por llamada, con el objetivo de asegurar su matrícula. Este proceso resulta muy demandante y se ve limitado por la disponibilidad de recursos humanos y tecnológicos, entre ellos, la falta de un método eficaz para prever cuántos

de estos aspirantes terminarán matriculándose.

Salesforce ha desarrollado herramientas que integran soluciones de inteligencia artificial para optimizar diversos procesos, incluyendo la capacidad de realizar predicciones basadas en datos históricos. Sin embargo, en el caso de la Pontificia Universidad Javeriana Cali, aunque los beneficios potenciales son sabidos, estas herramientas aún no han sido aprovechadas. Esto se debe a que su implementación requiere un proceso riguroso de preparación y ajuste, tanto antes como después de la creación y desarrollo del modelo predictivo.

1.1.1. Formulación

La pregunta ahora es: **¿Cómo pronosticar cuáles aspirantes pueden pasar a ser estudiantes para el siguiente periodo de matrícula aprovechando las herramientas de IA que ofrece Salesforce?**

1.1.2. Sistematización

Para abordar de manera integral el problema de investigación, se plantean las siguientes subpreguntas:

- ¿Cuáles son las variables que influyen en la decisión de un estudiante al matricularse en la Pontificia Universidad Javeriana Cali?
- ¿Cuál es el estado actual de los datos disponibles y qué ajustes son necesarios para mejorar su calidad y utilidad para el análisis predictivo?
- ¿Cómo utilizar las herramientas de IA de Salesforce para predecir a partir de las variables y optimizar las predicciones sobre la matrícula de aspirantes?
- ¿Qué tan precisas son las predicciones realizadas por el modelo?
- ¿Qué ajustes en la recopilación de datos son necesarios para mejorar la efectividad del modelo predictivo?

1.2. Objetivos

1.2.1. Objetivo General

Desarrollar un modelo predictivo utilizando las herramientas de IA disponibles en Salesforce para estimar la probabilidad de que un aspirante se matricule en la Pontificia Universidad Javeriana Cali.

1.2.2. Objetivos Específicos

- Identificar las variables que influyen en la decisión de matricula de los estudiantes, utilizando datos históricos y análisis de tendencias.
- Evaluar y preparar los datos existentes en la Universidad Javeriana Cali para su uso en análisis predictivo, identificando las deficiencias y realizando los ajustes necesarios para mejorar su calidad.
- Parametrizar las variables relevantes y generar un modelo predictivo sobre la matrícula de los estudiantes.
- Evaluar los resultados de las predicciones.
- Realizar los ajustes necesarios para obtener mejores resultados.

1.3. Justificación

La implementación de inteligencia artificial en los procesos de negocio de instituciones universitarias, específicamente este modelo predictivo relacionado con la probabilidad de que un estudiante se matricule, representa un avance en la optimización de recursos y en la toma de decisiones estratégicas y basada en datos.

Al contar con una plataforma de CRM robusta como Salesforce, que ya está integrada en los procesos de la universidad y con el acceso a datos tanto históricos como actuales de los estudiantes, se dispone de una base realmente sólida para el desarrollo de un modelo predictivo. Además, el uso de las herramientas de IA ya disponibles en el ecosistema de Salesforce, reduce la necesidad de inversiones adicionales significativas en tecnología. Por otra parte, el conocimiento acumulado por el personal administrativo y técnico en el manejo de la plataforma, combinado con el apoyo de Salesforce para la configuración y optimización de sus herramientas, permite que la implementación no solo sea realizable dentro del plazo dado para el proyecto, sino también adaptable en el futuro.

En lo que respecta al impacto, es importante destacar que el objetivo del proyecto no es implementar cambios radicales en el área de promoción institucional, sino más bien proporcionar herramientas analíticas que puedan guiar futuras estrategias. La investigación para el desarrollo del modelo predictivo permitirá identificar patrones y tendencias en los datos de los estudiantes, facilitando recomendaciones basadas en evidencia para optimizar las campañas de captación. De esta manera, el proyecto contribuye en la toma de decisiones basada en datos, promoviendo una mayor eficacia en la gestión de la Universidad. Aunque algunos podrían argumentar que la implementación de un modelo predictivo no atribuye un gran valor debido a que no se puede hacer mejoras instantáneas, el valor de este proyecto radica en su capacidad para generar conocimientos accionables que pueden ser considerados por los tomadores de decisiones, asegurando que cualquier ajuste futuro sea respaldado por datos sólidos y análisis riguroso.

Por último, la utilidad del proyecto se manifiesta en varios niveles. Para la universidad, un modelo predictivo de este ámbito no solo lleva a la optimización de los procesos de captación, sino que también contribuye a la búsqueda continua de la Universidad de ser una institución innovadora, ya que el desarrollo de este modelo predictivo sirve como una herramienta de aprendizaje para futuras implementaciones de IA para resolver problemas u optimizar procesos en otras áreas fundamentales, como Apoyo Financiero o Educación Continua. Aunque se podría argumentar que el modelo podría no ser útil en contextos o situaciones no previstas inicialmente, la implementación del modelo predictivo con ayuda de 'Einstein Discovery' permite realizar ajustes en tiempo real, permitiendo que el modelo se adapte rápidamente a nuevas tendencias o cambios en el comportamiento de los estudiantes, asegurando su relevancia continua y evitando la necesidad de reconstruir el modelo desde cero.

1.4. Delimitaciones y Alcances

1.4.1. Alcances

El proyecto tiene como objetivo desarrollar e implementar un modelo predictivo basado en Salesforce Einstein para determinar la probabilidad de matrícula de los aspirantes en la Pontificia Universidad Javeriana Cali. Dentro de este alcance se incluyó:

- El uso de datos históricos de aspirantes de pregrado específicamente de la Facultad de Ingeniería y Ciencias, con el fin de identificar patrones relevantes en el proceso de matrícula.
- El entrenamiento del modelo utilizando herramientas de inteligencia artificial ofrecidas por el ecosistema Salesforce, específicamente Prediction Builder, donde se ajustó y parametrizó las variables relevantes para el proceso de matrícula.
- La implementación del modelo en un entorno de pruebas controlado (QA) de Salesforce, validando su precisión mediante la comparación de predicciones con los resultados reales.
- Este proyecto se limitó al análisis de datos almacenados en el CRM institucional de Salesforce, sin intervenir otros sistemas académicos o administrativos.
- El modelo se desarrolló en un entorno de pruebas y su implementación en producción esta fuera de los alcances del proyecto.

1.4.2. Limitaciones

- **Cobertura de datos limitada:** El análisis se basó en los datos históricos disponibles desde septiembre de 2020, fecha a partir de la cual la universidad comenzó a utilizar el objeto Aplicación en Salesforce, el cual fue esencial para este proyecto. Registros anteriores a esta fecha no se consideraron por no contar con la estructura ni el nivel de detalle requerido.
- **Uso de licencia temporal:** El desarrollo del proyecto se realizó utilizando una versión funcional de *Prediction Builder*, disponible a través de una prueba gratuita de la licencia **CRM Analytics Plus**. Esta licencia tiene un costo comercial de **165 USD mensuales**, pero como parte del programa de aprendizaje de Salesforce, se permitió el acceso temporal sin costo para completar una de las guías prácticas. La continuidad del uso del modelo en un entorno operativo requerirá la adquisición de dicha licencia.

Capítulo 2

Marco Teórico y Trabajos Relacionados

2.1. Marco Teórico

Para entender el desarrollo de un modelo predictivo de matrícula de aspirantes en la Pontificia Universidad Javeriana Cali, es importante tener claridad sobre una serie de conceptos técnicos clave. El proyecto se centra en el uso de Salesforce, un sistema de Customer Relationship Management (CRM) ampliamente utilizado en el ámbito educativo y empresarial para gestionar relaciones con clientes o usuarios. Dentro de Salesforce, la herramienta de E permitirá aplicar técnicas de inteligencia artificial (IA) para crear un modelo predictivo basado en los datos históricos de aspirantes y estudiantes. Estos datos forman parte de lo que se denomina Big Data, un conjunto masivo de información que requiere técnicas avanzadas de procesamiento, como el Data Mining, para identificar patrones útiles para la predicción. Finalmente, el uso de modelos predictivos permitirá estimar las probabilidades de que un aspirante se convierta en estudiante, optimizando así los esfuerzos de captación de la universidad.

Customer Relationship Management (CRM) y Salesforce

Un CRM (Customer Relationship Management) es una plataforma que centraliza y gestiona las interacciones con clientes actuales y potenciales. El CRM permite almacenar información relevante, desde datos demográficos hasta interacciones previas con la organización, lo que facilita la personalización de estrategias de marketing y ventas.

En este proyecto, Salesforce, una plataforma líder en CRM en la nube, juega un papel esencial pues es el CRM utilizado por la Universidad Javeriana Cali. Además, se caracteriza por ser modular y escalable, lo que permite a las instituciones educativas integrar diversas funciones de ventas, servicio al cliente, marketing y análisis de

datos en una sola plataforma [2]. Estos sistemas no solo optimizan la interacción con los usuarios, sino que también permiten que las instituciones educativas gestionen el proceso de captación de nuevos estudiantes mediante el análisis y la organización de grandes volúmenes de información.

Inteligencia Artificial en Salesforce, Prediction Builder y Analytics Studio

Dentro del ecosistema de Salesforce, Einstein AI es un conjunto de soluciones de Inteligencia Artificial (IA) diseñadas para facilitar la automatización de procesos comerciales y la toma de decisiones basada en datos. Einstein AI utiliza algoritmos de machine learning y procesamiento del lenguaje natural para generar insights que no son fácilmente visibles mediante el análisis tradicional.

Prediction Builder es una herramienta de IA desarrollada por Salesforce que permite a las organizaciones analizar grandes volúmenes de datos y generar predicciones basadas en probabilidades mediante algoritmos avanzados de machine learning [3].

Analytics Studio es una herramienta de Salesforce orientada a la gestión y análisis de datos, diseñada para ayudar a los usuarios a comprender mejor sus datos, descubrir patrones y proporcionar recomendaciones basadas en estos hallazgos.

En el contexto de este proyecto, se trabaja en conjunto con Prediction Builder y Analytics Studio para entrenar modelos predictivos a partir de datos históricos de aspirantes y estudiantes, buscando identificar aquellos factores que influyen en la conversión de un aspirante en estudiante matriculado.

Manejo de Big Data

El concepto de Big Data se refiere a conjuntos de datos cuyo volumen, velocidad y variedad requieren herramientas tecnológicas especializadas para su procesamiento y análisis. En el contexto de la Universidad Javeriana Cali, el Big Data proviene de fuentes diversas como interacciones digitales, respuestas a campañas de marketing, redes sociales, y el historial de comunicación entre los aspirantes y la universidad. Técnicamente, Salesforce tiene la capacidad de manejar datos estructurados y no estructurados, como registros de interacciones, clics en campañas publicitarias, y correos electrónicos. Estos datos requieren almacenamiento en bases de datos distribuidas y escalables, como las ofrecidas por Salesforce con su infraestructura basada en la nube.
En este proyecto, el manejo de Big Data fue clave para la construcción de un modelo predictivo, ya que proporciona la base de información sobre la cual se identificarán patrones relevantes. Según Marr [4], una de las principales ventajas del Big Data es su capacidad para descubrir tendencias ocultas y generar predicciones más precisas,

algo crucial en la predicción de la matrícula universitaria.

Data Mining

El Data Mining es el proceso de extraer conocimiento útil a partir de grandes volúmenes de datos mediante técnicas analíticas avanzadas. En el marco de este proyecto, el Data Mining será fundamental para identificar patrones en los comportamientos de los aspirantes que influyen en su decisión final de matricularse. Las principales técnicas utilizadas incluyen la clasificación, que organiza los datos en categorías predefinidas para predecir resultados específicos, como la probabilidad de que un aspirante se matricule; el agrupamiento, que forma grupos de datos con características similares para identificar segmentos de aspirantes con comportamientos comunes; y las reglas de asociación, que revelan relaciones entre variables y permiten identificar combinaciones de acciones que ocurren con frecuencia, lo cual es útil para anticipar comportamientos de inscripción [5].

Modelos Predictivos

Los modelos predictivos son herramientas que permiten estimar resultados futuros basándose en el análisis de datos históricos. En este caso, el proyecto tiene como objetivo desarrollar un modelo que prediga la probabilidad de que un aspirante se matricule, utilizando técnicas de Machine Learning. Estos modelos son útiles en entornos donde la toma de decisiones puede mejorarse mediante la anticipación de comportamientos futuros. Según Shmueli y Koppius [6], los modelos predictivos son especialmente valiosos cuando se trata de identificar tendencias y patrones que pueden guiar las estrategias organizacionales, como es el caso de las campañas de captación de estudiantes.

Metodología CRISP-DM

La metodología CRISP-DM (Cross Industry Standard Process for Data Mining) es un modelo de procesos ampliamente adoptado en proyectos de minería de datos y ciencia de datos. Fue desarrollada a finales de los años 90 por un consorcio liderado por SPSS, NCR y Daimler-Benz [7]. Esta metodología se ha consolidado como un estándar industrial gracias a su enfoque estructurado, cíclico y adaptable a diferentes dominios, incluido el sector educativo.

CRISP-DM está compuesto por seis fases principales:

- **COMPRENSIÓN DEL NEGOCIO:** El primer paso consiste en comprender los objetivos estratégicos de la organización, identificar el problema desde

una perspectiva de negocio y traducirlo en un objetivo técnico de análisis de datos. Esta etapa también define los criterios de éxito del proyecto y establece un plan de trabajo. En este proyecto, se busca alinear el modelo predictivo con los procesos de admisión de aspirantes soportados en Salesforce.

- **COMPRENSIÓN DE LOS DATOS:** Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con los datos, identificar su calidad y establecer relaciones que permitan formular hipótesis iniciales. Si se trabaja sobre una base de datos operativa, como en Salesforce, es recomendable trabajar con una copia para evitar sobrecargar el entorno productivo.
- **PREPARACIÓN DE LOS DATOS:** Una vez comprendidos los datos, se realiza su preparación para adaptarlos a las técnicas de minería de datos. Esta fase incluye la selección de atributos, la limpieza de registros, el tratamiento de valores faltantes, la codificación de variables categóricas, la integración de distintas fuentes y la transformación del formato. En esta investigación, se extraen y transforman datos de objetos como Contact, Campaign, Application y Opportunity en Salesforce.
- **MODELAMIENTO:** Con los datos preparados, se procede a aplicar técnicas estadísticas y algoritmos de aprendizaje automático para generar modelos predictivos. En esta etapa se prueban diferentes enfoques y configuraciones con el fin de obtener el mejor desempeño posible. Se utiliza validación cruzada para garantizar la robustez del modelo.
- **EVALUACIÓN:** Esta etapa consiste en analizar el modelo construido para determinar si cumple con los objetivos del negocio. Se revisan métricas como precisión, recall, F1-score, y la curva ROC. Si el modelo no alcanza el rendimiento deseado o no satisface los requisitos estratégicos, se puede retornar a fases anteriores.
- **IMPLEMENTACIÓN:** Finalmente, se implementa el modelo en el entorno operativo. El despliegue puede implicar su integración en sistemas existentes, como Salesforce, o la generación de reportes y dashboards para usuarios finales. El objetivo es transformar el conocimiento extraído en acciones que apoyen la toma de decisiones, como en el caso de predicción de matrícula de aspirantes.

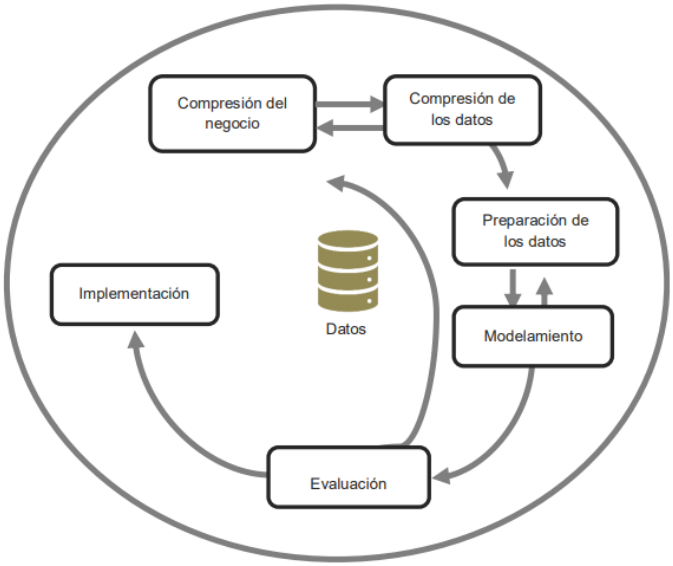

Figura 2.1: El proceso de la Metodología CRISP-DM (IBM, 2020)

En el presente trabajo se adopta esta metodología como marco estructural para la construcción del modelo predictivo.

2.2. Trabajos Relacionados

Predicción de la matrícula universitaria utilizando inteligencia computacional (2014)

Este trabajo, realizado por Ryan Stallings y Biswanath Samanta, se centró en la predicción de la matrícula universitaria mediante tres técnicas de inteligencia computacional (IC): redes neuronales artificiales (ANN), sistemas de inferencia neuro-difusos adaptables (ANFIS) y modelos de series temporales difusas agregadas. Cada técnica se evaluó como un problema de predicción de series temporales, utilizando datos de matrícula de la Universidad de Georgia Southern desde 1924 hasta 2012, y se probaron con datos de la Universidad de Alabama. Los resultados indicaron que cada técnica tiene distintos niveles de precisión: el modelo COPSO-SMN capturó adecuadamente la tendencia general de los datos, el modelo ANFIS mostró rendimiento excelente en entrenamiento pero menos preciso en pruebas, y el modelo de series temporales difusas agregadas obtuvo los mejores resultados con un error de predicción de solo 0.53 % para 2013. Este trabajo se relaciona directamente con nuestro proyecto debido a que al igual que ellos, buscamos realizar un modelo predictivo sobre el tema de la matrícula de estudiantes. Sin embargo, la diferencia principal radica en el enfoque del tipo de predicción: mientras ellos se centraron en predecir el número total de matrículas en un futuro periodo (pregunta numérica), nuestro proyecto se enfoca en determinar si un aspirante se matriculará o no

(pregunta binaria). Además, otra distinción importante es que nuestro modelo será desarrollado e integrado en Salesforce, aprovechando las herramientas de inteligencia artificial ya creadas en esta plataforma.

Implementación de Procesos de Matrícula en Salesforce (Desde 2016)

Desde la adopción de Salesforce en 2016, la universidad ha implementado diversos procesos clave en su gestión de matrícula y el ciclo de vida estudiantil, estableciendo un ecosistema robusto que facilita la integración de modelos de inteligencia artificial. Los procesos actuales incluyen:

- Atracción, Reclutamiento de Estudiantes y Admisiones.
- Gestión de Egresados y Donaciones.
- Créditos Financieros, Autoservicios Académicos y Financieros
- Sistema PQRFS (Peticiones, Quejas, Reclamos, Sugerencias y Felicitaciones).
- Unidades de Negocio Específicas.

Estudio sobre predicción de matrícula de estudiantes mediante minería de datos (2016)

Este estudio, realizado por Norhaidah A. Haris, Munaisyah Abdullah, Nurdatillah Hasim y Fauziah Abdul Rahman, explora cómo los métodos de minería de datos pueden mejorar la precisión de las predicciones de matrícula en instituciones de educación superior (HEI). Utilizando técnicas como regresión, clasificación, agrupamiento y reglas de asociación, y herramientas como WEKA, RapidMiner, KEEL, Orange, y Tanagra, se concluyó que estos métodos son efectivos para anticipar tendencias de inscripción, planificar recursos y tomar decisiones estratégicas. Los resultados destacaron la importancia de seleccionar técnicas adecuadas en función de los datos disponibles y objetivos específicos. Este estudio se relaciona con nuestro proyecto proporcionandonos información muy valiosa acerca del uso de modelos predictivos y procesamiento de grandes volúmenes de datos para predecir la matrícula de estudiantes utilizando las herramientas de Salesforce.

Modelos de análisis predictivo para la admisión y matrícula de estudiantes (2018)

Este estudio, conducido por Jared Cirelli, Andrea M. Konkol, Faisal Aqlan y Joshua C. Nwokeji, desarrolló modelos de análisis predictivo para mejorar la admisión y matrícula de estudiantes en programas universitarios, especialmente en ingeniería y ciencias de la computación. Se utilizaron modelos como regresión logística,

redes neuronales, redes bayesianas, árboles de decisión, CHAID y SVM, validados mediante la técnica de validación cruzada de k-pliegues. La precisión promedio de los modelos fue del 72 %, con la regresión logística y las máquinas de vectores de soporte (SVM) mostrando mayor precisión. Los resultados sugieren que la precisión podría mejorarse con más variables de entrada y datos históricos. Por otro lado, este proyecto es relevante para nuestro estudio ya que, aunque también busca la predicción de la matrícula y nos puede ayudar a encontrar algunas variables para entrenar el modelo predictivo, nuestro enfoque se limita a la matrícula utilizando datos del CRM de Salesforce.

Aplicación de modelos predictivos en la continuidad del proceso de formación (2022)

Este trabajo de grado desarrollado por Juan Felipe Mosquera García en la Pontificia Universidad Javeriana Cali propone un modelo para predecir la probabilidad de que un estudiante de pregrado continúe sus estudios en programas de posgrado dentro de la misma universidad.

A través del uso de técnicas de minería de datos y aprendizaje automático, y empleando la metodología CRISP-DM, se construyó un modelo predictivo basado en atributos académicos, demográficos y financieros de los estudiantes egresados. La investigación permitió identificar patrones relevantes en la continuidad académica que nos pueden servir y demuestra la viabilidad de aplicar la ciencia de datos en el contexto de la Universidad Javeriana Cali para mejorar la toma de decisiones estratégicas.

2.3. Diferenciadores del proyecto:

El presente trabajo se distingue de los estudios previamente revisados por varios factores relacionados con el enfoque del problema, las herramientas empleadas y el contexto institucional en el que se desarrolla.

- **Enfoque individual y tipo de predicción:** A diferencia de investigaciones como la de Stallings y Samanta (2014), que se centraron en la predicción del número total de matrículas mediante técnicas de series temporales, este proyecto aborda el problema desde una perspectiva individual, formulado como una tarea de determinar si un aspirante específico se matriculará o no. Esta diferencia en el enfoque permite orientar el modelo hacia decisiones operativas personalizadas dentro del proceso de admisión.
- **Plataforma tecnológica empleada:** El modelo propuesto se desarrolla directamente en la plataforma Salesforce, utilizando herramientas nativas como Prediction Builder, Prediction Builder y CRM Analytics. A diferencia de otros

estudios que emplearon plataformas genéricas como WEKA, RapidMiner o entornos de programación tradicional, este proyecto se apoya en soluciones no-code integradas en el CRM institucional.

En cuanto al trabajo de Mosquera (2022), se reconoce una cierta similitud, ya que ambos estudios están dirigidos a predecir comportamientos futuros de estudiantes de la Pontificia Universidad Javeriana Cali y comparten metodologia para el desarrollo del proyecto (CRISP-DM). Sin embargo, presentan diferencias en cuanto a los objetivos, mientras que el trabajo de Mosquera se orienta a la continuidad académica en programas de posgrado, el presente proyecto se enfoca en la etapa inicial del ciclo de vida estudiantil, especificamente en la facultad de Ingeniería y Ciencias de pregrado.

Capítulo 3

Aplicación de la Metodología CRISP-DM

Con el objetivo de implementar un modelo predictivo con Salesforce para estimar la probabilidad de matrícula de los aspirantes, se aplicó la metodología CRISP-DM para el desarrollo de este proyecto. A continuación, se describen las fases implementadas y las actividades específicas llevadas a cabo en cada una de ellas.

3.1. Entendimiento del negocio

Al contar previamente con experiencia en un proyecto desarrollado durante el periodo 2024-1 en el marco de la Escuela Salesforce, donde se comprendió de forma general el flujo de procesos entre áreas y los objetivos de cada una de ellas, se disponía ya de una base de conocimiento organizacional. Sin embargo, para este proyecto fue necesario enfocarse de manera particular en los procesos de captación de aspirantes y matrícula.

Entre las actividades realizadas en esta fase se encuentran:

- **Reunión con Kelly Yurani Vanegas, Coordinadora de Mercadeo Digital:** Se revisaron los parámetros utilizados en la creación y ejecución de campañas digitales, considerando las restricciones actuales sobre publicidad dirigida a menores de edad (vigentes desde 2024) y analizando aspectos como el rango de edades e intereses de los públicos objetivos.
- **Reunión con Claudia Ximena Gallo, Coordinadora de Servicio al Cliente (call center):** Se profundizó en el flujo del proceso de inscripción, los mecanismos de seguimiento aplicados a los aspirantes, y los factores críticos que pueden influir en la decisión de matrícula.

- **Reunión con Jerzy Andres Moncayo, profesional de Promoción de Programas:** Se compartieron informes y paneles de seguimiento de aspirantes almacenados en Salesforce, permitiendo una mejor comprensión de la estructura de los datos y la identificación preliminar de algunas variables relevantes para el análisis.

Tras estas reuniones de contextualización, se obtuvo un entendimiento más completo de las operaciones de mercadeo y promoción, lo cual permitió avanzar a la siguiente fase metodológica.

3.2. Entendimiento de los datos

En esta fase, se realizó la recolección inicial de los datos de aspirantes dentro del sistema Salesforce, los cuales sirvieron como base para el proceso de minería de datos. Se determinó la cantidad de registros disponibles, objetos de interes, sus variables y descripciones.

3.2.1. Fuente de los datos:

La información se obtuvo del CRM institucional Salesforce, donde los datos de los aspirantes se encuentran almacenados en diferentes objetos relacionados entre sí. En particular, los objetos de interés para este proyecto fueron `Cuenta`, `Campaña`, `Miembro de Campaña`, `Contacto`, `Contacto`, `Oportunidad` y `Aplicación`. Los cuales forman parte de la arquitectura EDA (Education Data Architecture), un modelo de datos estandarizado desarrollado por Salesforce para instituciones educativas, el cual organiza la información de aspirantes, estudiantes, cursos y relaciones institucionales de manera flexible y escalable [8].

A continuación, se describen brevemente los objetos tenidos en cuenta para la recolección de los datos:

- **Cuenta (Account):** Representa entidades institucionales o personas asociadas al proceso académico, tales como facultades, programas académicos o colegios de procedencia. Constituye un eje central en la arquitectura EDA, ya que múltiples objetos se relacionan con este.
- **Campaña (Campaign):** Permite agrupar esfuerzos de reclutamiento, eventos o estrategias de promoción, y realizar un seguimiento de su efectividad en la captación de aspirantes.
- **Miembro de Campaña (Campaign Member):** Objeto que asocia un contacto específico con una campaña determinada, permitiendo registrar su participación y estado dentro de dicha iniciativa.

- **Contacto (Contact):** Contiene la información personal del aspirante, incluyendo datos como nombre, correo, teléfono, identificación y otra información demográfica o de contacto relevante.
- **Tarea (Task):** Registra actividades o acciones realizadas dentro del sistema, como llamadas, seguimientos, correos enviados o tareas asignadas por los asesores de admisión a los aspirantes.
- **Oportunidad (Opportunity):** Representa el proceso de admisión o matrícula de un aspirante, incluyendo el estado de avance en dicho proceso. Es uno de los objetos clave para determinar el resultado del proceso, utilizado como base para definir la variable objetivo del modelo.
- **Aplicación (hed application c):** Centraliza la información académica, personal y administrativa de la postulación del aspirante, como el programa al que aplica, el periodo académico, y otros datos complementarios a la oportunidad.

En la Figura 3.1 se ilustra el esquema de relaciones entre los objetos utilizados.

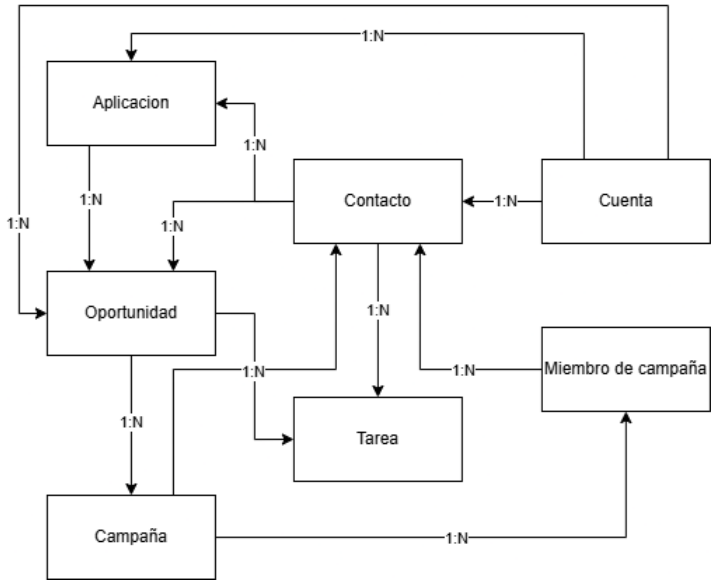

Figura 3.1: Esquema de relaciones entre los objetos de interés

Inicialmente, se contaba con acceso a un "sandbox", es decir, un entorno de pruebas proporcionado por la universidad en el proyecto Escuela Salesforce. Este entorno de pruebas contenía datos históricos desactualizados, los cuales eran útiles para la exploración de los objetos y sus variables, pero insuficientes para la construcción de un conjunto de datos preciso.

Por ello fue necesario gestionar los datos del entorno de producción institucional de Salesforce, el cual contiene información actualizada y sensible de aspirantes. Cabe destacar que este acceso no fue inmediato, se debió cumplir con protocolos internos

de la universidad, incluyendo la firma de un acuerdo de confidencialidad, incluido en el Anexo #3. Lo cual implicó una espera que retrasó ligeramente los tiempos proyectados, pero a su vez permitió comprender de forma práctica los procedimientos de gobernanza de datos en contextos reales.

3.2.2. Definición del conjunto de datos

Se extrajo la información relevante de los aspirantes mediante un proceso de filtrado en cuatro fases:

1. **Relevancia institucional:** Se filtraron los objetos para incluir únicamente aquellos utilizados de manera activa por la universidad.
2. **Utilidad para la investigación:** Se seleccionaron los objetos y variables que tenían un potencial valor analítico para el objetivo del estudio.
3. **Aceptabilidad de variables:** Se consideraron variables que, aunque no ideales, podían ser aprovechadas con ciertas precauciones metodológicas.
4. **Selección de variables clave:** Para cada objeto, se determinaron las variables más relevantes, priorizando aquellas con mayor completitud de datos.

A partir del análisis de los registros disponibles en Salesforce, se determinó que el período comprendido para la extracción de datos abarcó desde el 8 de septiembre de 2020 hasta el 26 de mayo de 2025. Dentro de este rango se obtuvieron **4.086 registros** de aplicaciones asociadas a aspirantes de pregrado de la facultad de Ingeniería y Ciencias, cada una vinculada con su respectiva oportunidad de admisión.

Además, se llevó a cabo una reunión con Alexander Valencia (jefe del CRM institucional), Carlos Augusto Gutierrez (Coordinador de Análisis de Datos) y Carlos Arturo Dominguez (Ingeniero de Proyectos de Desarrollo TI). Durante esta sesión, se discutió la selección de la variable objetivo (target) y se profundizó en su significado operativo.

Se definió que la variable objetivo sería `StageName` del objeto `Opportunity`, que modela el seguimiento del aspirante a través de las etapas del proceso de matrícula. Solo se consideraron las oportunidades cuya etapa estuviera cerrada, ya sea con éxito o sin éxito, correspondientes a las siguientes categorías:

- **Pagó:** El aspirante completó exitosamente el proceso de matrícula.
- **Cerrada perdida:** El aspirante no concretó la matrícula.
- **No admitido:** El aspirante no fue admitido tras el proceso de evaluación.

- **Desistió:** El aspirante inició el registro pero decidió no continuar.

A lo largo de este documento, esta variable será referida como `Estado_Oportunidad`, para mantener consistencia con la denominación utilizada durante el procesamiento de datos.

La Tabla 3.1 muestra la distribución de estas categorías. Como se puede observar, existe un desequilibrio notable entre ellas: la clase *Cerrada Perdida* representa más de la mitad de los casos (55.96 %), mientras que otras como *No Admitido* o *Desistió* no llegan ni al 5 %.

Valor	**Frecuencia**	**Porcentaje**
Cerrada Perdida	2,286	55.96 %
Pagó	1,563	38.26 %
Desistió	180	4.41 %
No admitido	57	1.40 %
Total	4,086	100 %

Tabla 3.1: Distribución de la variable *Estado_Oportunidad*

Este desequilibrio es particularmente relevante para el entrenamiento del modelo, ya que una distribución desbalanceada puede inducir al algoritmo a favorecer la predicción de la clase mayoritaria, reduciendo su capacidad de detectar correctamente casos menos frecuentes.

Para mitigar este problema y facilitar la interpretación, se optó por transformar la variable en una categoría binaria, agrupando las clases según el resultado del proceso de matrícula::

- **Ganada:**
  - Pagó
- **Pérdida:**
  - Cerrada perdida
  - No admitido
  - Desistió

Es decir, las oportunidades con etapa en estado “Pago” fueron clasificadas como aspirantes matriculados (Oportunidades GANADAS), mientras que las oportunidades en “Cerrada Perdida”, “No Admitido” o “Desistió” fueron clasificadas como aspirantes no matriculados (Oportunidades PÉRDIDAS).

Siguiendo el proceso de filtrado, se procedió a explorar los objetos previamente identificados, con el fin de extraer una lista preliminar de variables con valor predictivo para el modelo. Esta exploración comenzó en el entorno sandbox institucional, que si bien contenía datos desactualizados, permitió identificar la estructura de los objetos y las variables potencialmente útiles de cada uno de estos.

A continuación, en las tablas 3.2 y 3.3 se muestran variables preliminares identificadas de cada objeto:

Objeto	**Variables preliminares**
Cuenta	▪ Nombre_Facultad__c (En cuentas de Facultades) ▪ Name (En cuentas de Colegios y Programas Académicos)
Campaña	▪ Type (Tipo de campaña) ▪ NumberOfContacts (N.º de contactos) ▪ NumberOfOpportunities (N.º de oportunidades) ▪ NumberOfWonOpportunities (N.º de oportunidades ganadas)
Miembro de Campaña	▪ LeadSource (Origen del candidato)
Tarea	▪ WhatId (ID del registro relacionado) ▪ Subject (Breve descripción de la tarea) ▪ Tipo_de_llamada__c (Entrante/Saliente) ▪ Número_de_intentos__c (Intentos de llamada) ▪ Detalle_del_seguimiento_al_interés__c ▪ Resultado_de_gestión__c

Tabla 3.2: Variables preliminares por objeto (I)

Objeto	Variables preliminares
Aplicación y Oportunidad	▪ Interés_y_actitud__c ▪ Número_de_llamadas__c ▪ Ayuda_financiera__c ▪ Estrato__c ▪ Ingreso_en_SMLMV__c ▪ Present_SABER_11__c ▪ Stagename (Estado_Oportunidad) ▪ hed__Application_Status__c ▪ hed__Application_Type__c

Tabla 3.3: Variables preliminares por objeto (II)

Teniendo ya esta lista de variables preliminares, se siguió depurando de forma progresiva, aplicando los siguientes criterios:

- **Eliminación de variables con texto libre**, ya que no resultaban fácilmente categorizables ni compatibles con modelos supervisados.
- **Revisión metodológica**, descartando variables que, aunque disponibles, no ofrecían valor explicativo claro para el problema de predicción.

Otro factor determinante fue la necesidad de trasladar el conjunto de datos a un entorno QA, en el cual se consolidó la información relevante en un objeto personalizado construido específicamente para Prediction Builder. Este paso requería respetar las relaciones originales entre los objetos, lo que impuso ciertas restricciones técnicas.

En particular, los objetos Campaña, Miembro de Campaña y Tarea mantenían relaciones de tipo N:1 con el objeto Oportunidad, lo cual implicaba que una única oportunidad podía estar asociada a múltiples campañas, interacciones o tareas. Al intentar extraer esta información directamente, se generaban estructuras altamente redundantes: por cada relación múltiple, el resultado era una expansión horizontal en forma de docenas (o incluso cientos) de columnas adicionales por registro, como se muestra conceptualmente en la Figura 3.2.

Figura 3.2: Ejemplo de expansión de columnas por relaciones múltiples

Aunque se evaluó la posibilidad de usar agregaciones estadísticas (por ejemplo: número total de tareas, cantidad de campañas activas, promedio de intentos de contacto), se concluyó que dicho tratamiento requeriría una transformación adicional que no aportaba un valor significativo en ese momento del proyecto.

Afortunadamente, una vez se obtuvo acceso al entorno de producción actualizado, se observó que la mayoría de las variables de interés ya se encontraban directamente en los objetos Aplicación y Oportunidad. Esta observación fue clave, ya que permitió:

- **Simplificar el conjunto de datos**, concentrando las variables en dos objetos bien estructurados.
- **Evitar problemas de redundancia o relaciones complejas**, facilitando la busqueda con querys y creación del conjunto de datos.
- **Identificar nuevas variables** potencialmente predictivas directamente desde los objetos centrales.

En este contexto, uno de los objetos fundamentales para el modelo predictivo fue el objeto Aplicación, ya que concentra información sociodemográfica y académica clave de los aspirantes.

En la Tabla 3.4 se describen las variables seleccionadas del objeto Oportunidad, las cuales se relacionan principalmente con el seguimiento del proceso de matrícula y aspectos académicos. Por otro lado, en la Tabla 3.5 se presentan las variables extraídas del objeto Aplicación, donde predominan algunas de tipo sociodemográficas, económicas y de características personales del aspirante.

Tabla 3.4: Variables seleccionadas del objeto *Opportunity*

Variable	**Descripción**
Estado_Oportunidad	Estado de la oportunidad, utilizado como variable objetivo para determinar si el aspirante se matriculó o no.
Programa_Académico.Name	Nombre del programa académico al que aplica el aspirante.
Número_de_llamadas	Número de llamadas realizadas al aspirante como parte del proceso de admisión.
Número_de_Whatsapp	Número de interacciones por WhatsApp con el aspirante.

Tabla 3.5: Variables seleccionadas del objeto *hed__Application__c*

Variable	**Descripción**
Estrato	Nivel socioeconómico declarado por el aspirante.
Ayuda_financiera	Indica si el aspirante quiere solicitar algún tipo de ayuda financiera.
Método_de_financiación	Medio de financiación especificado por el aspirante, si aplica.
Sexo	Sexo biológico del aspirante.
Edad	Edad calculada automáticamente con base en la fecha de nacimiento.
Colegio	Institución educativa de procedencia del aspirante.
Puntaje_Global_Tx	Puntaje global del examen ICFES (Pruebas Saber 11).
generación_E_Aspirante	Indica si el aspirante desea participar en el programa Generación E.
Departamento_de_residencia	Departamento de residencia actúal del aspirante.
Madre_Labora	Indica si la madre del aspirante tiene empleo.
Padre_Labora	Indica si el padre del aspirante tiene empleo.
Cargo_madre	Ocupación o cargo laboral de la madre.
Cargo_padre	Ocupación o cargo laboral del padre.
Madre_fallecida_desconocida	Indica si la madre ha fallecido o no es conocida por el aspirante.
Padre_fallecido_desconocido	Indica si el padre ha fallecido o o no es conocido por el aspirante.
Nivel_formacion_de_la_madre	Ultimo nivel de formación académica alcanzado por la madre.
Nivel_formacion_del_padre	Ultimo nivel de formación académica alcanzado por el padre.

3.2.3. Descripción de los datos

Una vez realizada la selección de los datos relevantes para el análisis, el siguiente paso consistió en evaluar su estado inicial antes de aplicar cualquier técnica de preprocesamiento. Esta revisión preliminar permitió identificar posibles inconsistencias o deficiencias en la estructura del conjunto de datos, particularmente en cuanto al tipo de dato, distribución y a la proporción de valores faltantes o ausentes en las distintos variables.

En la Tabla 3.6 se presenta resaltando con asterisco (*) la variable objetivo, un resumen detallado que incluye el número y el porcentaje de valores no nulos por cada variable, así como el tipo de dato correspondiente. Lo que permitió identificar variables con una alta proporción de valores faltantes, como "Nivel de formación de madrez "Nivel de formación de padre", lo que podría afectar su utilidad en el modelo y obligar a considerar su exclusión o la imputación de datos.

También se observó que muchas variables son de tipo object, lo cual corresponde a variables categóricas representadas como texto en el contexto del presente proyecto. En modelos entrenados manualmente, este tipo de datos requiere transformaciones previas para ser interpretados por los algoritmos. No obstante, la herramienta de Prediction Builder maneja automáticamente estas variables mediante técnicas como one-hot encoding, que consiste en convertir cada categoría en una columna binaria independiente [9], facilitando su incorporación sin necesidad de codificación manual.

#	**Variable**	**Valores no nulos**	**Tipo de dato**
0	Estrato	3675 (89.94 %)	float64
1	Ayuda_financiera	2780 (68.04 %)	object
2	Metodo_de_Financiacion	1546 (37.84 %)	object
3	Edad	3776 (92.41 %)	float64
4	Genero	3770 (92.27 %)	object
5	generación_E_Aspirante	4086 (100.00 %)	object
6	Departamento_de_residencia	3738 (91.43 %)	object
7	Madre_Labora	2409 (58.96 %)	object
8	Padre_Labora	2114 (51.74 %)	object
9	Cargo_madre	1816 (44.44 %)	object
10	Cargo_padre	2074 (50.76 %)	object
11	Madre_fallecida_o_desconocida	2986 (73.08 %)	object
12	Padre fallecido_o_desconocido	3090 (75.62 %)	object
13	Puntaje_Global_Tx	2172 (53.16 %)	object
14	Nivel_formacion_de_la_madre	643 (15.74 %)	object
15	Nivel_formacion_del_padre	565 (13.83 %)	object
16	Estado_Oportunidad *	4086 (100.00 %)	object
17	Programa_Académico.Name	4086 (100.00 %)	object
18	Numero_de_llamadas	3970 (97.16 %)	float64
19	Numero_de_Whatsapp	2887 (70.66 %)	object
20	Colegio	2474 (60.55 %)	object

Tabla 3.6: Información general del conjunto de datos

En la Figura 3.3 se presenta la distribución de frecuencias de las dos clases correspondientes a la variable objetivo "Estado_Oportunidad": Ganada y Perdida. Se observa un leve desbalance, donde la clase Ganada, que representa las oportunidades que culminaron en matrícula efectiva corresponde al 38,3 % del total. Por su parte, la clase Perdida que agrupa los casos en los que no se concretó la matrícula, ya sea por no ser admitido, por desistimiento o por no formalizar el proceso representa el 61,7 %.

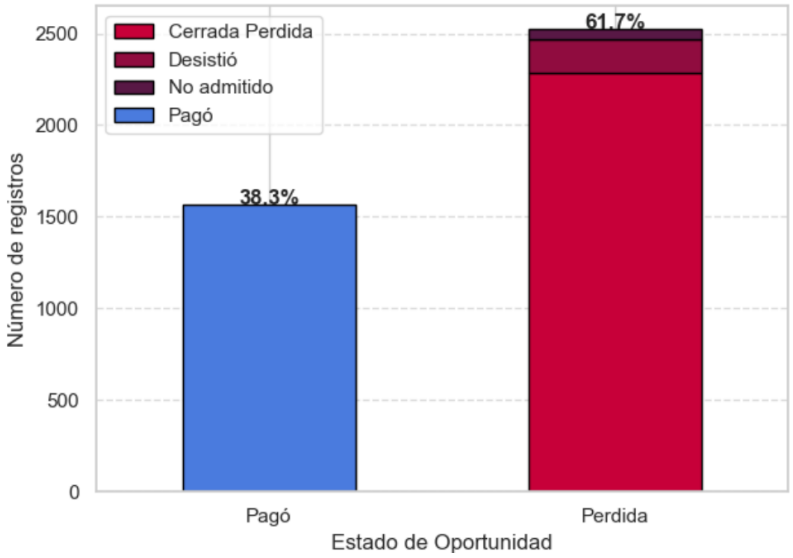

Figura 3.3: Distribución de frecuencia de clases

En la Figura 3.4 se muestra la distribución de edades de los aspirantes. Se evidencia una mayor concentración de datos entre los 15 y 20 años, lo cual era esperable dado el perfil típico de los aspirantes. No obstante, también se observan algunos valores atípicos a partir de los 30 años, que podrían corresponder a casos excepcionales o posibles errores de registro. Por otra parte, en la Figura 3.5 se presenta la distribución de frecuencias por género, donde se identifica una mayor representación del género masculino, con un 66,3 %, frente a un 33,7 % del género femenino.

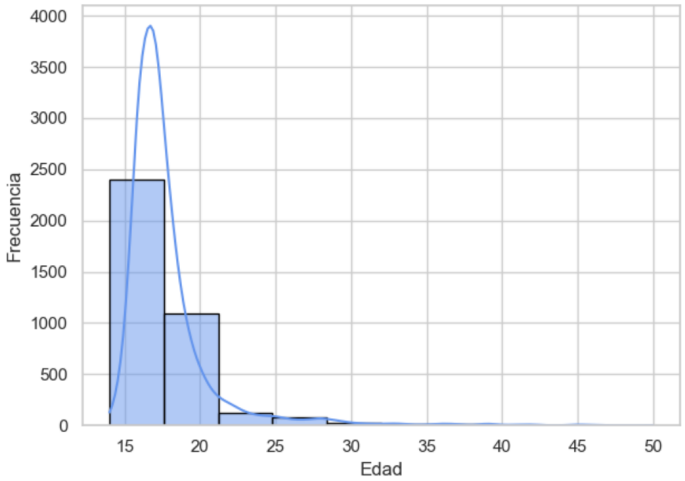

Figura 3.4: Distribución de frecuencia de la edad

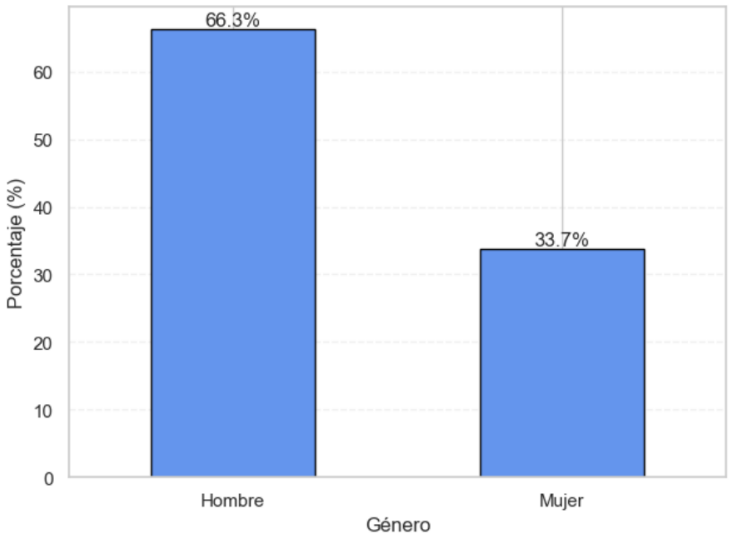

Figura 3.5: Distribución de frecuencia del genero

En la Figura 3.6 se muestra la distribución de frecuencias por programa académico. Se observa una clara predominancia del programa de Ingeniería de Sistemas y Computación, mientras que el programa con menor número de registros es Matemáticas Aplicadas.

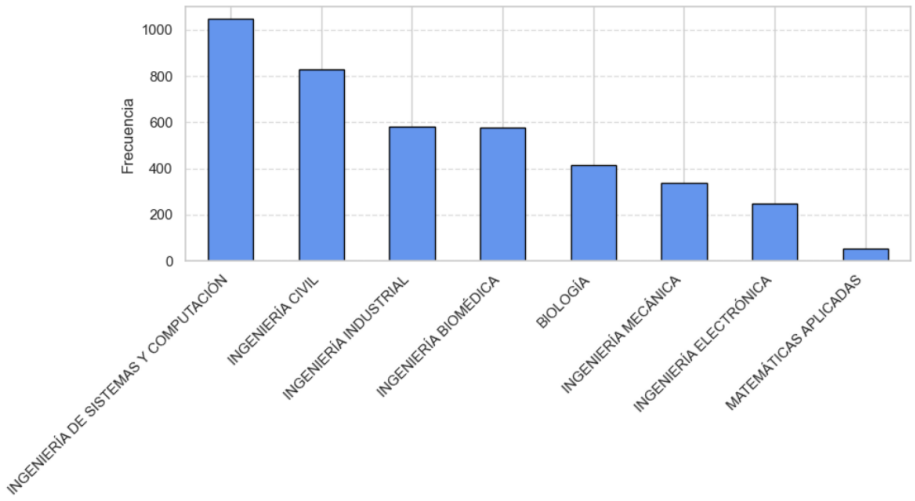

Figura 3.6: Distribución de frecuencia de Programas académicos

Finalmente, en la Figura 3.7 se presenta el gráfico de densidad del puntaje ICFES de los aspirantes. Los datos tienden a concentrarse principalmente en el rango de 300 a 350 puntos, lo cual es coherente con el rendimiento promedio esperado. Sin embargo, también se identifican algunos valores por encima del puntaje máximo permitido (500), lo que sugiere la presencia de datos erróneos o mal digitados que deben ser corregidos o excluidos en el preprocesamiento.

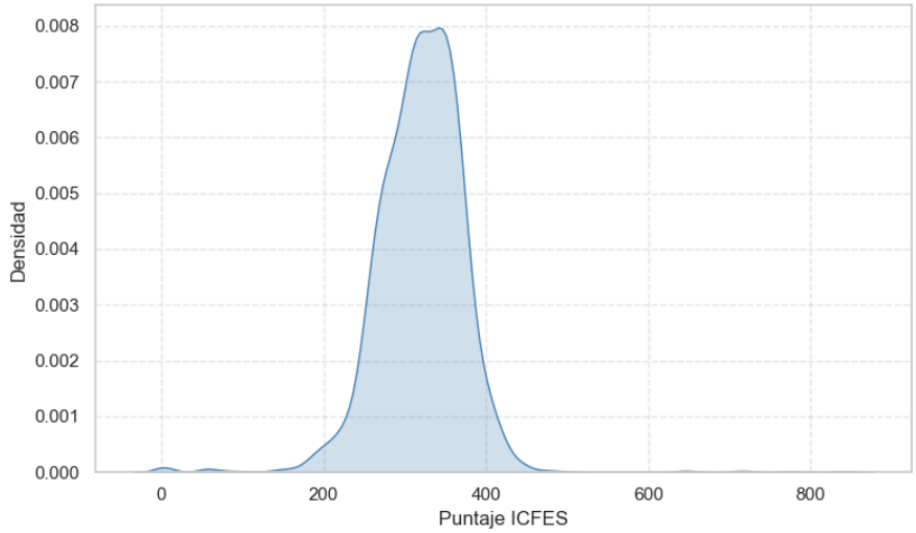

Figura 3.7: Gráfico de densidad del puntaje ICFES

3.3. Preparación de los datos

En esta etapa se llevó a cabo el procesamiento de los datos, finalizando el proceso de limpieza y abordando el tratamiento de los valores faltantes. Se analizaron distintas técnicas de imputación con el objetivo de mejorar la calidad de la información disponible, así como el formato más adecuado para preparar los datos de cara al entrenamiento del modelo. Una vez hecho esto, se importaron los datos al entorno de Salesforce y se configuró el espacio de trabajo necesario para utilizar la herramienta Prediction Builder.

3.3.1. Limpieza de los datos

Tras el análisis exploratorio inicial, se tomó la decisión de realizar una depuración de los datos, con el objetivo de eliminar registros que presentaran valores atípicos, errores evidentes de digitación o características que pudieran introducir ruido y afectar negativamente el desempeño del modelo.

En total, se identificaron y depuraron registros por las siguientes razones:

- **Registros con valor nulo en la variable "Estrato":** la mayoría de estos casos presentaban múltiples variables adicionales con valores nulos, lo que sugiere que se trataba de registros incompletos o inconsistentes, posiblemente basura.
- **Edad superior a 40 años:** se consideró que estos casos representaban valores atípicos dentro del contexto de aspirantes a programas de pregrado. Además,

algunos registros contenían edades irrealistas (superiores a los 100 años), indicando posibles errores de digitación. Cabe señalar que estos valores no se visualizaban en la gráfica de distribución de edad, ya que esta se acotó para resaltar el rango donde se concentra la mayoría de los datos.

- **Puntajes ICFES superiores al máximo permitido (500):** en estos casos no se eliminó el registro completo, sino únicamente el valor de la variable correspondiente, debido a que el resto de la información era válida y útil para el análisis.
- **Registros con múltiples variables de valor nulo o sin valor investigativo:** aquellos registros que presentaban una gran cantidad de datos faltantes o respuestas vacías poco informativas fueron considerados poco confiables y, por tanto, eliminados.

Como resultado de este proceso, se obtuvo un conjunto de datos final compuesto por **3.329 registros** válidos y aptos para continuar con el preprocesamiento y modelado.

Por otra parte, se realizó un análisis complementario de las variables Nivel de formación de madre y Nivel de formación de padre, con base en las Figuras 3.9 y 3.8, que muestran la distribución del porcentaje de aspirantes que completaron el proceso de matrícula (Pagó) según el nivel educativo de los padres. Este análisis se complementa con las Tablas 3.8 y 3.7, donde se presenta la frecuencia absoluta de cada categoría.

En ambos casos se evidencia una tendencia clara: a medida que aumenta el nivel educativo de los padres, también se incrementa la proporción de aspirantes que finalizan exitosamente el proceso de matrícula. Por ejemplo, entre los aspirantes cuyas madres tienen título de doctorado, el porcentaje de Pagó alcanza un 75.0 %, seguido por Magíster (66.7 %) y Especialista (59.0 %). Esta misma tendencia se observa en los padres: Doctor (71.4 %), Magíster (67.9 %) y Especialista (65.9 %). En contraste, los porcentajes más bajos de matrícula se presentan en los niveles más bajos de formación: “Ninguno”muestra apenas un 29.3 % para las madres y un 30.6 % para los padres. Lo cual puede estar asociado a una mayor valoración del sistema educativo dentro del entorno familiar

No obstante, al revisar las frecuencias absolutas (Tablas 3.8 y 3.7), se observa que los niveles académicos más altos (Doctorado y Magíster) tienen una representación mucho menor dentro del conjunto de datos con respecto a las demás categorias. Por ejemplo, solo se registran 4 casos de madres con formación doctoral y 14 registros en el caso de los padres. A esto se suma lo expuesto previamente en la sección de descripción de los datos, donde se indicó que ambas variables presentan un alto porcentaje de valores faltantes: 15,74 % de valores no nulos para Nivel de formación de madre y 13,83 % para Nivel de formación de padre. Esta escasez de datos válidos limita su aplicabilidad como variables predictoras y por ende, se decidió excluirlas del conjunto de entrenamiento del modelo.

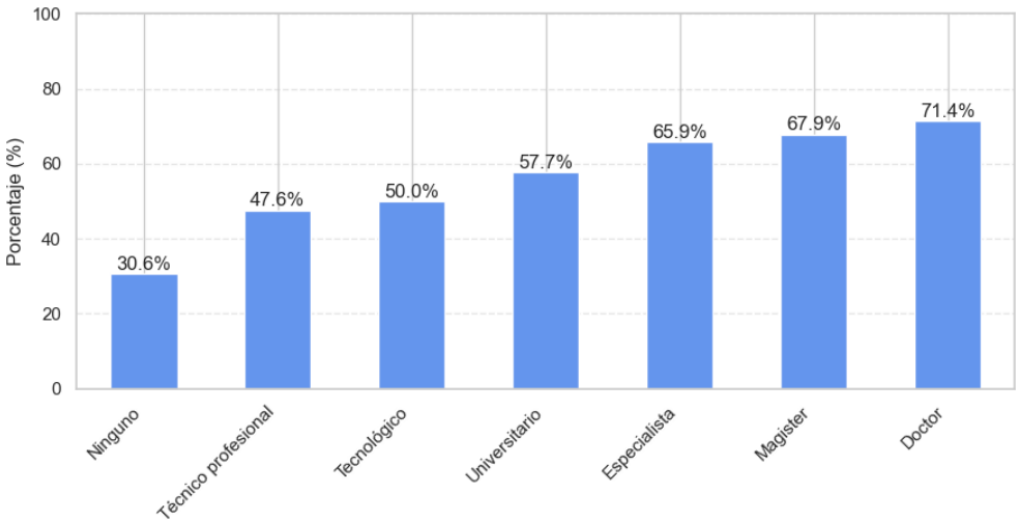

Figura 3.8: Distribución del porcentaje de "Pagó"según nivel de formación del padre

Nivel de formación del padre	**Cantidad**
Ninguno	134
Técnico profesional	63
Tecnológico	34
Universitario	182
Especialista	82
Magister	56
Doctor	14

Tabla 3.7: Distribución del nivel de formación del padre

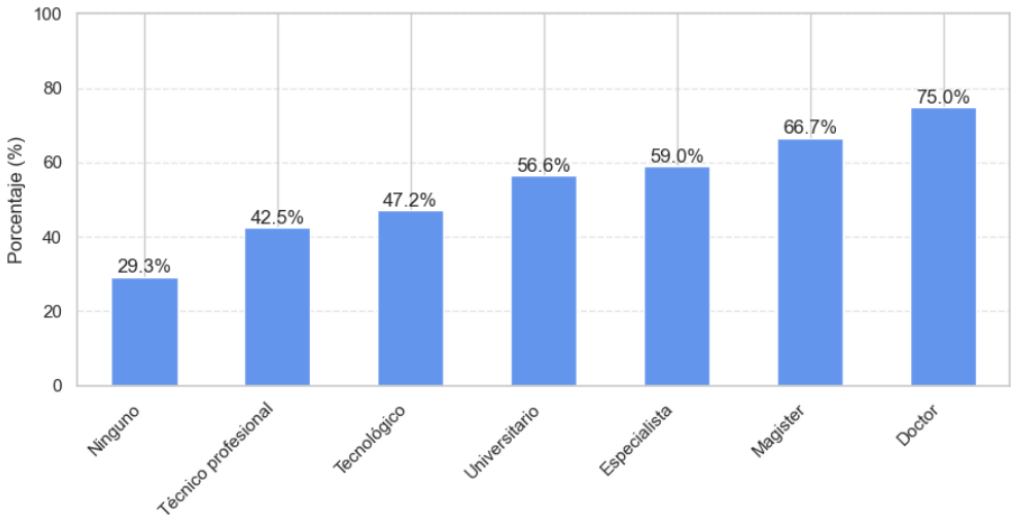

Figura 3.9: Distribución del porcentaje de "Pagó"según nivel de formación de madre

Nivel de formación de la madre	Cantidad
Ninguno	147
Técnico profesional	73
Tecnológico	36
Universitario	235
Especialista	100
Magister	48
Doctor	4

Tabla 3.8: Distribución del nivel de formación de la madre

3.3.2. Transformación de los datos

Después del proceso de limpieza, en esta etapa se tomaron decisiones clave para estructurar adecuadamente las variables categóricas y numéricas, así como para abordar los valores faltanes y minimizar posibles sesgos.

En primer lugar, se analizó la variable categórica Colegio, la cual presentó un comportamiento de alta cardinalidad: se identificaron 761 categorías distintas, de las cuales 445 (aproximadamente el 58 %) tenían una frecuencia de aparición igual a uno, como se resume en la Tabla 3.9.

Estadística	Valor
Total de registros	4086
Total de categorías distintas	761
Categorías con frecuencia 1	**445**
Categorías con frecuencia entre 2 y 5	208
Categoría más frecuente (moda)	COLEGIO BERCHMANS - CALI (VAL)
Frecuencia de la moda	1612
Porcentaje de la moda (%)	39.45 %
Valores nulos	1612

Tabla 3.9: Resumen estadístico de la variable *Colegio*

Debido a que una gran parte de las categorías tenían baja representación y que la variable también contenía valores nulos que no podían ser imputados sin riesgo de introducir sesgo, se optó por agrupar todas las categorías con frecuencia de uno en una categoría ya existente denominada 'OTRO'. Lo que permitió reducir la dimensionalidad de la variable sin perder la capacidad de discriminar entre casos representativos y casos atípicos.

En cuanto a la variable numérica Puntaje ICFES se realizó la conversión explícita del tipo de dato a float64 para asegurar la correcta aplicación de la tecnica de imputación de datos.

Luego, se abordó el tratamiento de los valores faltantes, ya que una parte significativa de las variables presentaba datos nulos. La estrategia se centró en aplicar técnicas de imputación adecuadas según el tipo y contexto de cada variable. En muchos casos, se aprovechó el hecho de que los modelos desarrollados con Salesforce Prediction Builder pueden manejar eficazmente variables categóricas, lo que permitió incorporar nuevas categorías que representaran explícitamente la ausencia de información.

A continuación, se describen las variables transformadas y la estrategia utilizada en cada caso:

- **Ayuda financiera**: variable categórica binaria. Se añadió una tercera categoría llamada `No responde` para representar la falta de respuesta. Esta decisión se basó en que un valor nulo no implica necesariamente una respuesta negativa, sino que puede corresponder a una omisión o falta de información.
- **Método de financiación**: variable categórica multiclase. Su valor depende directamente de la variable `Ayuda financiera`, ya que solo aplica cuando esta es igual a `Sí`. Por lo tanto, se creó la categoría `No aplica` y se asignó a los registros donde `Ayuda financiera` era distinta de `Sí`.
- **Madre labora** y **Padre labora**: ambas variables eran categóricas binarias. Se añadió una tercera categoría, `No responde`, para representar la ausencia de información sin asumir un valor arbitrario. Imputar con la moda podría distorsionar la distribución real de estas variables y afectar la interpretación del modelo.
- **Cargo madre** y **Cargo padre**: variables categóricas multiclase que solo se completan si la madre o el padre respectivamente se encuentran laboralmente activos. En estos casos se añadió la categoría `No aplica` para aquellos registros donde no se cumplía esta condición.
- **Madre fallecida o desconocida** y **Padre fallecido o desconocido**: estas variables ya incluían una categoría denominada `No aplica`, la cual se aprovechó para imputar los valores faltantes de forma coherente.
- **Puntaje Global Tx**: se aplicó una estrategia estructurada de imputación que permite conservar tanto la información faltante como los patrones particulares del conjunto de datos. Inicialmente, se generó una variable binaria auxiliar denominada `ICFES_faltante`, que identifica si el valor original estaba ausente (1) o presente (0). Este enfoque llamado *missing indicator method*, es útil cuando la ausencia del dato puede estar relacionada con patrones significativos o contener información relevante para el modelo predictivo [10].

  Posteriormente, se realizó la imputación del puntaje ICFES utilizando una estrategia jerárquica basada en subgrupos. En primer lugar, se imputaron los valores faltantes utilizando la mediana del colegio al que pertenecía cada registro, bajo la hipótesis de que el rendimiento promedio puede variar entre

instituciones educativas. Para los registros que no contaban con el nombre del colegio, se usó la mediana del `Estrato`, empleándolo como una variable proxy del contexto socioeconómico del aspirante.

3.3.3. Importación de datos al entorno QA

Una vez tenido, el conjunto de datos definitivo y ya tranformado correctamente, era necesario contar con un objeto personalizado que consolidara la información relevante para poder usar la prueba de la herramienta Prediction Builder dentro del entorno de Salesforce. Para esto, se importó de forma manual utilizando la herramienta de "Salesforce Inspector" los datos de distintos objetos de Salesforce hacia un nuevo entorno de pruebas (QA) creado específicamente para este proyecto, como se como se observa en la Figura 3.10. Esta tarea incluyó tanto las variables obligatorios de cada objeto como aquellas variables anteriormente mencionadas identificadas como relevantes.

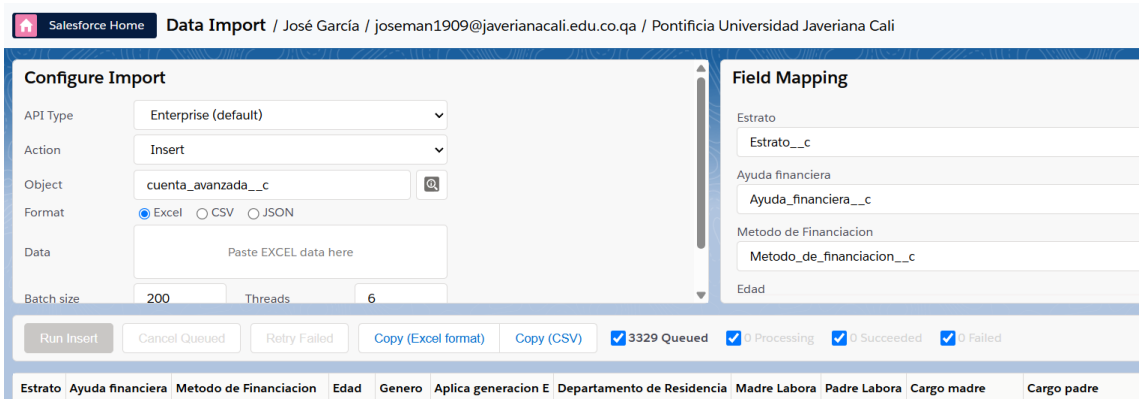

Figura 3.10: Proceso de importación de los datos a entorno QA

La importación de los datos exigió una planificación clara con el objetivo preservar las relaciones entre los objetos. El proceso se llevó a cabo en un orden jerárquico: primero se importaron los registros del objeto Cuenta, siguiendo un suborden interno que respetara las dependencias entre tipos de cuentas (facultades, programas académicos y colegios). Posteriormente, se cargaron los registros del objeto Oportunidad, y finalmente los del objeto Application (Aplicación), dado que cada aplicación está relacionada con una oportunidad, y esta a su vez con las cuentas previamente mencionadas.

Durante la importación surgieron diversos inconvenientes técnicos, por ejemplo:

- **Falta de variables necesarias:** Algunas variables requeridas por las relaciones entre objetos no existían en el entorno QA, lo cual requirió la creación manual de nuevas variables personalizadas.

- **Reglas de validación restrictivas:** Varias reglas de validación configuradas en los objetos impedían la carga de registros de prueba. Estas reglas tuvieron que ser modificadas o desactivadas temporalmente para permitir una importación sin errores.

- **Incompatibilidades en las características de las variables:** Se encontraron diferencias en los formatos, tipos de datos o restricciones entre los entornos de origen y destino. Fue necesario ajustar estas configuraciones para asegurar la integridad y consistencia de los datos importados.

3.3.4. Creación del objeto personalizado

Para facilitar el entrenamiento del modelo de inteligencia artificial en Salesforce y centralizar la información necesaria, se creó un objeto personalizado denominado **Cuenta Avanzada**. Este objeto consolida las variables relevantes provenientes de diferentes objetos del modelo EDA (como *Application*, *Opportunity*, entre otros), permitiendo una representación unificada de los registros de aspirantes.

Dado que los valores de dichas variables ya se encuentran estandarizados en sus respectivos objetos de origen, se optó por utilizar tipos de datos simples, principalmente cadenas de texto (`string`) en **Cuenta Avanzada**, lo cual simplifica tanto la importación como la manipulación de los datos dentro del entorno de pruebas (QA). Además, este diseño evita relaciones complejas y redundancias estructurales, asegurando la integridad de los datos y la compatibilidad con *Prediction Builder*.

Las variables seleccionadas en este objeto se detallan en la Tabla 3.10.

Tabla 3.10: Variables seleccionadas del objeto *Cuenta Avanzada*

Variable	**Descripción**
Área	Área académica o unidad administrativa relacionada con la postulación.
Ayuda Financiera	Indica si el aspirante desea solicitar apoyo financiero.
Año periodo académico	Año correspondiente al periodo de ingreso académico.
Cargo madre	Ocupación laboral actual o más reciente de la madre.
Cargo padre	Ocupación laboral actual o más reciente del padre.
Departamento de residencia	Departamento en el cual reside actualmente el aspirante.
Deseo aplicar generación E	Indica si el aspirante expresó interés en el programa Generación E.
Edad	Edad actual del aspirante.
Estrato	Nivel socioeconómico declarado por el aspirante.
Facultad	Facultad a la cual pertenece el programa académico solicitado.
Madre fallecida	Señala si la madre ha fallecido.
Madre labora	Indica si la madre del aspirante tiene empleo activo.
Padre fallecido	Señala si el padre ha fallecido.
Padre labora	Indica si el padre del aspirante tiene empleo activo.
Método de financiación	Medio de financiación previsto para cubrir los estudios.
Programa académico	Programa académico seleccionado por el aspirante.
Puntaje Global Tx	Puntaje total obtenido en las pruebas ICFES Saber 11.
Stage name	Estado final del proceso de matrícula (variable objetivo).
Último nivel formación del padre	Nivel educativo más alto alcanzado por el padre.
Último nivel formación de la madre	Nivel educativo más alto alcanzado por la madre.

3.4. Modelado

En esta fase del proyecto, se seleccionaron y aplicaron las técnicas de modelado para la construcción del modelo predictivo. A diferencia de un enfoque tradicional que implica la programación manual de algoritmos en lenguajes como Python, se optó por utilizar las capacidades de inteligencia artificial nativas de la plataforma

Salesforce, específicamente la herramienta Prediction Builder.

Esta decisión se fundamenta en varias ventajas estratégicas y técnicas. Primero, la integración directa con el CRM institucional garantiza un acceso fluido y en tiempo real a los datos, eliminando la necesidad de procesos complejos de exportación e importación y asegurando que el modelo opere sobre la información más reciente. Segundo, Prediction Builder funciona como una plataforma de Aprendizaje Automático Automatizado (AutoML).Lo que significa que la herramienta es capaz de evalúar múltiples algoritmos de clasificación para seleccionar automáticamente aquel que ofrece el mejor rendimiento predictivo, optimizando los hiperparámetros sin intervención manual. Este enfoque permite concentrar los esfuerzos en la calidad de los datos y en la interpretación de los resultados desde una perspectiva de negocio.

Aunque Salesforce Prediction Builder consolida el proceso en una única predicción optimizada, para este trabajo se analizará el rendimiento de tres configuraciones o modelos distintos. Este enfoque permitirá realizar un análisis comparativo robusto para seleccionar el que tenga mejor rendimiento, para eso se tomaran en cuenta ciertas mediciones clave.

3.4.1. Selección de Variables Predictoras

La construcción de un modelo predictivo robusto depende fundamentalmente de la selección de un conjunto de variables que aporten información relevante y eviten la introducción de ruido o sesgos. A partir de un conjunto inicial de más de 20 variables candidatas, identificadas en la fase de entendimiento de los datos, se llevó a cabo un proceso de selección que resultó en los 13 predictores finales utilizados para entrenar el modelo.

El criterio de selección se basó en dos pilares. En primer lugar, se priorizaron aquellas variables que, durante el Análisis Exploratorio de Datos (EDA), mostraron una mayor capacidad para discriminar entre las clases y revelaron patrones de interés. En segundo lugar, se tomaron en cuenta las recomendaciones y alertas automáticas generadas por la propia herramienta de Salesforce. Por ejemplo, la tabla que muestra la importancia de las diferentes variables frente a la variable objetivo 3.11, la cual permitió probar distintas combinaciones y determinar cuáles ofrecían una mayor precisión. Por otra parte, la plataforma tambien advirtió sobre una posible **multicolinealidad** entre las variables ‘Ayuda financiera‘ y ‘Método de financiación‘. La multicolinealidad ocurre cuando dos o más predictores están altamente correlacionados, lo que puede dificultar la interpretación del efecto individual de cada variable en el modelo. Atendiendo a esta recomendación, y dado que ‘Ayuda financiera‘ es la variable que condiciona la existencia de la segunda, se decidió excluir ‘Método de financiación‘ para preservar la simplicidad y robustez del modelo.

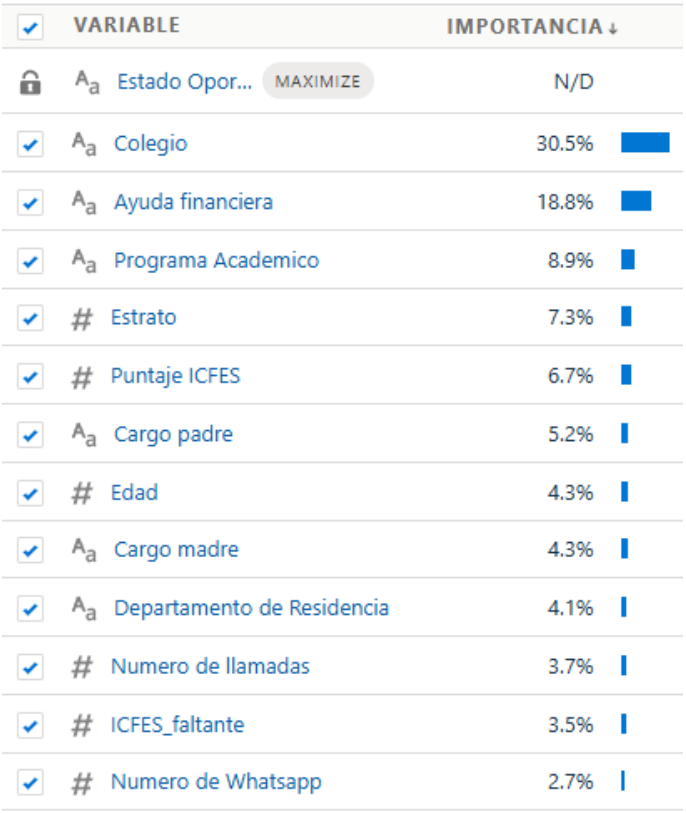

Figura 3.11: Tabla de importancia de variables.

A continuación, se describen las variables finales seleccionadas para el entrenamiento del modelo:

Estado Oportunidad (Variable Objetivo): Variable categórica binaria que toma los valores *"Ganada"* o *"Perdida"*. Representa el resultado final del proceso de admisión.

Colegio: Variable textual que identifica la institución de educación secundaria de la cual proviene el aspirante.

Ayuda financiera: Variable categórica que indica si el aspirante manifestó interés en solicitar algún tipo de apoyo financiero.

Programa Academico: Variable categórica que especifica el programa de pregrado al que el aspirante aplicó.

Puntaje ICFES: Variable numérica que representa la puntuación global obtenida por el aspirante en las pruebas de Estado Saber 11.

Estrato: Variable numérica ordinal que indica el nivel socioeconómico de la vivienda del aspirante.

Cargo padre: Variable categórica que describe la ocupación o cargo laboral del padre del aspirante.

Cargo madre: Variable categórica que describe la ocupación o cargo laboral de la madre del aspirante.

Edad: Variable numérica que representa la edad del aspirante al momento de la aplicación.

Departamento de Residencia: Variable categórica que indica la ubicación geográfica principal del aspirante.

Numero de llamadas: Variable numérica que cuantifica el total de llamadas de seguimiento realizadas al aspirante por parte del equipo de promoción.

Tiene_ICFES: Variable binaria (1 o 0) creada durante la preparación de los datos para indicar si el aspirante reportó o no un puntaje ICFES. Actúa como un *missing indicator*.

Numero de Whatsapp: Variable numérica que registra el total de interacciones mantenidas con el aspirante a través de este canal de comunicación.

3.4.2. Métricas de evaluación del rendimiento

Para evaluar adecuadamente el rendimiento de los modelos predictivos, es importante utilizar métricas que consideren el desbalance entre las 2 clases ("Ganada" y "Perdida") identificadas durante el análisis exploratorio. Para ello, se emplea un conjunto de métricas derivadas de la matriz de confusión, que permiten obtener una visión más completa del comportamiento del modelo.

La matriz de confusión es una tabla que resume el número de predicciones correctas e incorrectas [11], y para este proyecto, sus componentes se definen de la siguiente manera:

- **Verdadero Positivo (VP):** El modelo predice que un aspirante **se matriculará**, y en realidad **sí se matriculó**.
- **Verdadero Negativo (VN):** El modelo predice que un aspirante **no se matriculará**, y en realidad **no se matriculó**.
- **Falso Positivo (FP):** El modelo predice que un aspirante **se matriculará**, pero en realidad **no se matriculó** (Error Tipo I).
- **Falso Negativo (FN):** El modelo predice que un aspirante **no se matriculará**, pero en realidad **sí se matriculó** (Error Tipo II).

Cabe destacar que en el caso del proyecto, el Error Tipo II es un fallo importante porque se omite un caso de matrícula, clase minoritaria en el dataset.

A partir de estos componentes, se definen las siguientes métricas de evaluación clave utilizadas en este trabajo:

- **AUC (Area Under the Curve):** El Área bajo la Curva ROC (Receiver Operating Characteristic) es una métrica que evalúa qué tan bien un modelo es capaz de distinguir entre clases [12]. La curva ROC grafica la Tasa de Verdaderos Positivos (Sensibilidad) frente a la Tasa de Falsos Positivos para diferentes umbrales de clasificación. Un valor de AUC de 1.0 representa un modelo perfecto, mientras que 0.5 indica un rendimiento no mejor que el azar. Es una medida robusta del poder de discriminación general del modelo.

- **Coeficiente de Gini:** se calcula como la proporción del área entre la línea de perfecta igualdad y la curva de Lorenz, dividida por el área total bajo la línea de igualdad [13]. En términos más simples, es una medida de desigualdad, donde 0 representa la igualdad perfecta y 1 representa la desigualdad total.

$$\text{Gini} = \frac{A}{(A+B)}$$

- **Puntuación F1 (F1-Score):** Es la media armónica de la Precisión y la Sensibilidad [14]. Es una de las métricas más importantes para problemas con clases desbalanceadas, ya que busca un equilibrio entre no generar falsos positivos y no omitir verdaderos positivos. Un valor alto indica un modelo robusto.

$$\text{Puntuación F1} = 2 \times \frac{\text{Precisión} \times \text{Sensibilidad}}{\text{Precisión} + \text{Sensibilidad}}$$

- **Coeficiente de Correlación de Matthews (MCC):** Es una medida de la calidad de las clasificaciones binarias que resulta especialmente robusta, ya que tiene en cuenta las cuatro categorías de la matriz de confusión. Su valor varía entre -1 (predicción totalmente errónea), 0 (predicción aleatoria) y +1 (predicción perfecta) [15]. Es recomendable para conjuntos de datos desbalanceados.

$$\text{MCC} = \frac{(VP \times VN) - (FP \times FN)}{\sqrt{(VP+FP)(VP+FN)(VN+FP)(VN+FN)}}$$

3.4.3. Seleccion de modelos

Para la construcción del modelo predictivo, la herramienta de Salesforce evaluó diferentes algoritmos de aprendizaje automático. Los tres modelos analizados en este trabajo —Random Forest, GBM y XGBoost— son técnicas de **ensamblaje (*ensemble learning*)**. Estos métodos combinan las predicciones de múltiples modelos

más simples (generalmente árboles de decisión) para obtener un resultado final más preciso y robusto que el que podría ofrecer un único modelo.

A continuación, se describe brevemente cada uno.

Random Forest (Bosque Aleatorio)

El modelo **Random Forest** es un método de aprendizaje conjunto (ensemble) que opera mediante la construcción de una multitud de árboles de decisión durante la fase de entrenamiento [16]. Su robustez se fundamenta en la introducción de aleatoriedad a través de dos mecanismos clave. Primero, cada árbol se entrena sobre una muestra de datos distinta, obtenida mediante la técnica de *bootstrap aggregating* o **bagging**. Esto implica que para un conjunto de datos de tamaño N, se extraen N muestras con reemplazo, resultando en subconjuntos que contienen aproximadamente el 63.2 % de los datos originales, con algunas instancias repetidas y otras omitidas. Segundo, al construir cada nodo de un árbol, no se evalúan todas las variables disponibles, sino un **subconjunto aleatorio de características** (*max_features*).

Esta doble aleatoriedad asegura que los árboles individuales sean decorrelacionados entre sí. Para realizar una predicción, el modelo agrega los resultados de todos los árboles: en problemas de clasificación, se elige la clase con la **mayoría de votos**, mientras que en regresión se promedia el resultado [17]. Este modelo tiene una buena resistencia al sobreajuste (*overfitting*), ya que el promediado de las predicciones de árboles no correlacionados reduce significativamente la varianza del modelo final sin un aumento sustancial en el sesgo. En este estudio, sirvió como el modelo base para comparar el rendimiento.

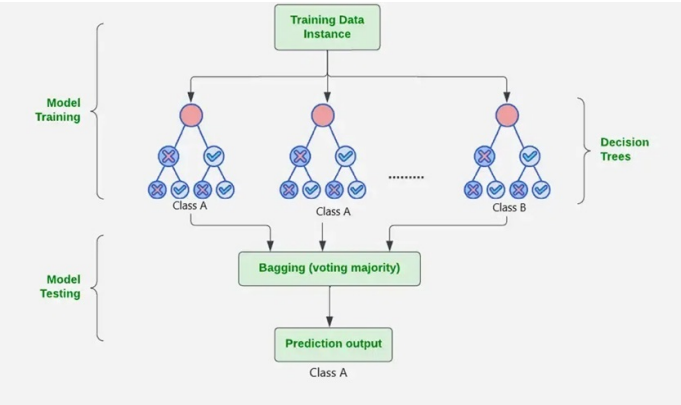

Figura 3.12: Ejemplo del funcionamiento de Random Forest. Tomado de [17].

GBM (Gradient Boosting Machine)

A diferencia del enfoque paralelo de Random Forest, el **Gradient Boosting Machine (GBM)** es un método de *boosting* que construye los árboles de decisión de manera **secuencial y aditiva**. El objetivo del algoritmo es minimizar una función de pérdida (loss function) predefinida mediante un procedimiento análogo al descenso de gradiente. Cada nuevo árbol se entrena para corregir los errores residuales del conjunto de árboles anteriores [18].

Más específicamente, el primer árbol se ajusta para predecir el objetivo. Luego, en cada iteración, se calculan los **pseudo-residuales**, que son el gradiente negativo de la función de pérdida con respecto a las predicciones del ensamble actual. Un nuevo árbol de decisión se ajusta para predecir estos pseudo-residuales. La contribución de este nuevo árbol al ensamble es ponderada por un **factor de aprendizaje** (*learning rate* o *eta*), un hiperparámetro crucial que reduce el impacto de cada árbol y previene el sobreajuste. De esta forma, el modelo se enfoca iterativamente en los casos donde su error es mayor, mejorando su rendimiento de manera gradual. Como se evidenció en los resultados, GBM representó una mejora significativa en las métricas operativas en comparación con el modelo base.

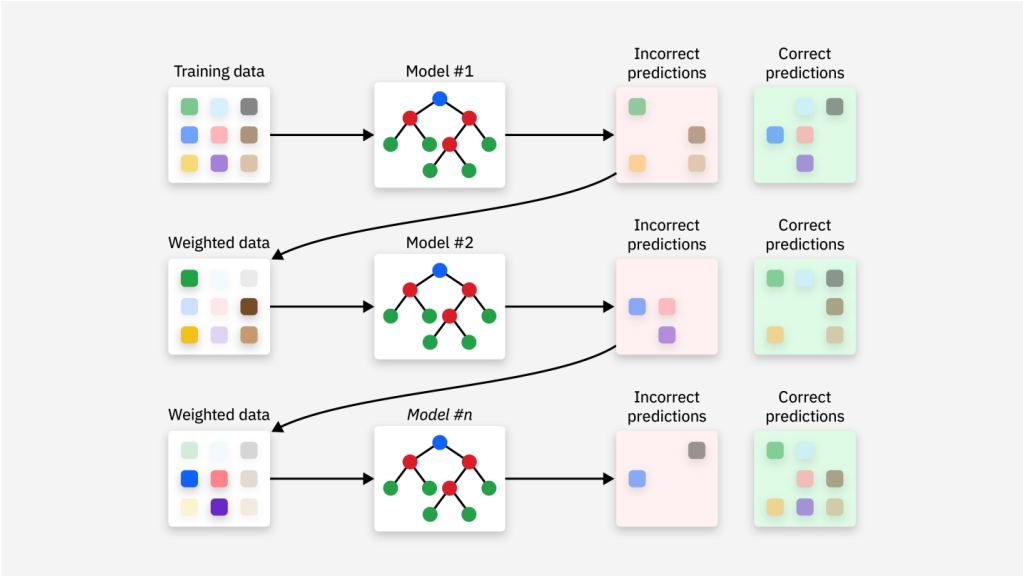

Figura 3.13: Ejemplo del funcionamiento del modelo GBM. Tomado de [18]

XGBoost (Extreme Gradient Boosting)

El modelo **XGBoost** es una implementación avanzada y altamente optimizada del algoritmo de Gradient Boosting, reconocida por su eficiencia computacional y su rendimiento predictivo superior. XGBoost refina el GBM tradicional en varias áreas clave:

1. **Regularización explícita**: A diferencia del GBM estándar, que controla la complejidad implícitamente a través del *learning rate* y la profundidad de los árboles, XGBoost incorpora términos de **regularización L1 (Lasso) y L2 (Ridge)** en su función objetivo [19]. Estos términos penalizan la complejidad del modelo (tanto el número de hojas como la magnitud de sus pesos), lo que combate eficazmente el sobreajuste.

2. **Optimización del rendimiento**: XGBoost está diseñado para ser computacionalmente eficiente. Permite el **procesamiento en paralelo** durante la fase de construcción de los árboles; aunque los árboles se generan secuencialmente, el algoritmo puede paralelizar el proceso de búsqueda del mejor punto de corte (*split*) dentro de cada nodo. Además, utiliza estructuras de datos optimizadas (bloques comprimidos y ordenados por columnas) y algoritmos *cache-aware* para un uso eficiente del hardware.

3. **Manejo avanzado de datos**: El algoritmo puede gestionar **valores faltantes** de manera nativa. Durante el entrenamiento, aprende una dirección por defecto en cada nodo para las instancias con valores ausentes, optimizando la ganancia de la división.

Gracias a estas mejoras, que también incluyen una aproximación de Taylor de segundo orden en la función de pérdida para una convergencia más rápida y precisa, XGBoost demostró en este proyecto un rendimiento superior y más equilibrado en las métricas clave, consolidándose como la opción final para el modelo predictivo.

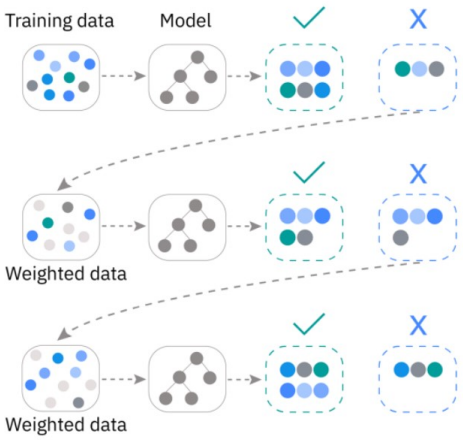

Figura 3.14: Ejemplo del funcionamiento del modelo XGBoost. Tomado de [20]

3.4.4. Resultados y Análisis Comparativo de Modelos

Una vez finalizado el proceso de entrenamiento y optimización con las variables seleccionadas, se procedió a evaluar el rendimiento de tres configuraciones de modelos distintas para realizar una comparación exhaustiva; se evaluaron los modelos con 2

clases, teniendo en esta última en la clase cerrada pérdidas, valores de No admitido y Desistió.

Si bien se analizaron múltiples métricas, se puso especial énfasis en la **Puntuación F1** y el **Coeficiente de Correlación de Matthews (MCC)** como indicadores clave para la selección del modelo final. Estas métricas son particularmente adecuadas para este caso de uso, dado el desbalance de clases existente, ya que ofrecen una medida más equilibrada del rendimiento que la exactitud por sí sola. Los resultados obtenidos para cada modelo se resumen en la Tabla 3.11.

Tabla 3.11: Comparativa del Desempeño entre Modelos Predictivos para clase Pagó

Métricas	**Random Forest**	**GBM**	**XG Boost**
AUC	0.8331	0.9190	**0.9369**
Gini	0.6662	0.8381	**0.8737**
Puntuación F1	0.718	0.806	**0.833**
Coeficiente de Correlación de Matthews	0.5086	0.667	**0.7114**

Se evidencia un patrón claro del análisis comparativo: el modelo **XGBoost** supera consistentemente a **Random Forest** y **GBM** en todas las métricas evaluadas, como AUC, Gini, puntuación F1 y coeficiente de correlación de Matthews (MCC), demostrando una mayor capacidad para generar predicciones precisas y confiables. Esta diferencia se acentúa frente a Random Forest, cuyas métricas reflejan un aprendizaje deficiente. En contraste, XGBoost destaca por su robustez y por manejar de buena forma el desbalance de clases, lo cual es fundamental para una aplicación práctica que requiere decisiones concretas. En consecuencia, se concluye que XGBoost es el modelo más confiable y equilibrado entre los evaluados, con una mejor capacidad para identificar correctamente tanto a los aspirantes mayoritarios como minoritarios.

3.4.5. Análisis de Resultados por Modelo

Tras la evaluación de los tres modelos de ensamblaje (*ensemble*) seleccionados, se consolidaron sus métricas de rendimiento para una comparación directa. El objetivo era identificar no solo el modelo con la mayor precisión general, sino aquel que ofreciera el mejor equilibrio para la tarea específica de predecir la matrícula de aspirantes. Los resultados se presentan en las tablas del inciso anterior.

A continuación, se realiza un análisis individual de cada modelo.

Random Forest (Modelo Base): Este modelo, aunque conceptualmente robusto, sirvió como línea base. Como se observa en cada una de las tablas, sus métricas de rendimiento en un punto de corte de clasificación específico (Puntuación F1 de 0.718) son las más modestas del grupo. Curiosamente, a su vez obtuvo el valor

más bajo en la métrica AUC, lo que indica una capacidad de discriminación entre clases menor a través de todos los posibles umbrales. Sin embargo, para la aplicación práctica que requiere tomar decisiones con un umbral fijo, su rendimiento es superado por los otros modelos.

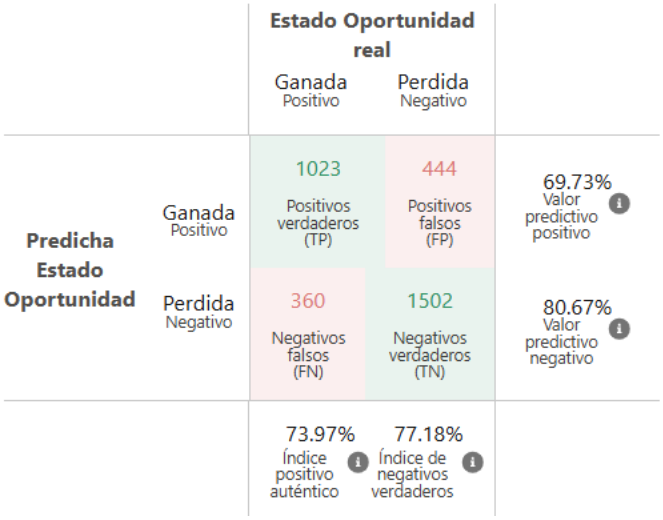

Figura 3.15: Matriz de Confusión para el modelo Random Forest.

GBM - Gradient Boosting Machine (Modelo Intermedio): El modelo GBM representó una mejora significativa sobre Random Forest en casi todas las métricas operativas. Con una Puntuación F1 de 0.806, demuestra ser un modelo más equilibrado y preciso en el punto de decisión. Su enfoque de corregir secuencialmente los errores le permite construir un predictor más fuerte, aunque su capacidad de discriminación general (AUC) fue inferior a la del modelo XG Boost.

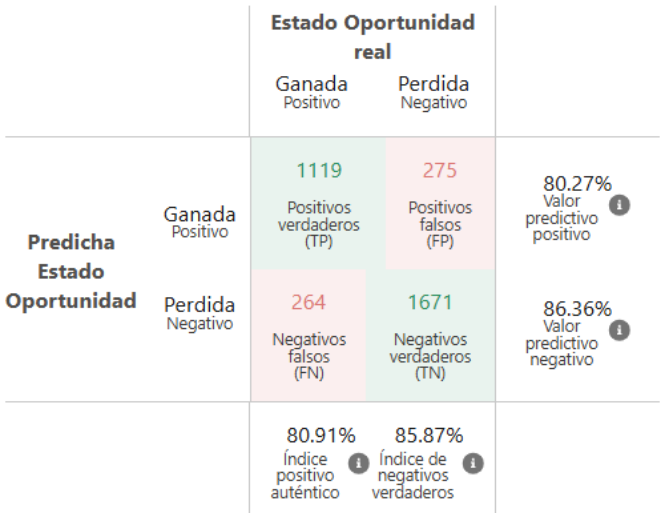

Figura 3.16: Matriz de Confusión para el modelo GBM.

XGBoost - Extreme Gradient Boosting (Modelo Final y Seleccionado): Finalmente, el modelo XGBoost se consolidó como la opción superior para este proyecto. Logró el mejor rendimiento en todas las demás métricas, que son cruciales para la toma de decisiones: Precisión máxima, Puntuación F1 , MCC y Precisión Media por Clase . Esto significa que XGBoost es el modelo más confiable para identificar correctamente a los aspirantes que sí se matricularán (alta precisión), sin dejar de capturar a la mayoría de ellos (alta sensibilidad, reflejada en una alta F1). Por su rendimiento superior y equilibrado, se selecciona como el modelo final para este trabajo de grado.

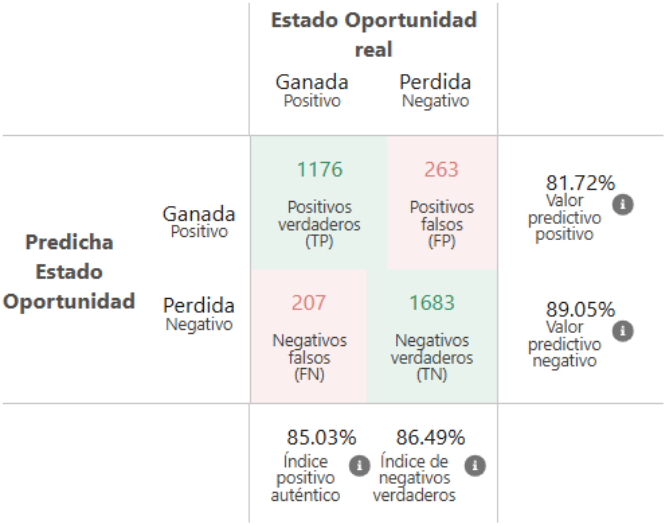

Figura 3.17: Matriz de Confusión para el modelo XGBoost.

3.4.6. Hallazgos Clave a partir del Modelo

Más allá de la capacidad predictiva del modelo, su entrenamiento y análisis revelaron una serie de patrones y *insights* estratégicos sobre el comportamiento de los aspirantes. Estos hallazgos permiten a la universidad comprender con mayor profundidad los factores que influyen en la decisión de matrícula. A continuación, se detallan los más relevantes:

- **Impacto de la Ayuda Financiera:** Se identificó una fuerte tendencia en los aspirantes que disponen de medios propios para financiar sus estudios. El modelo mostró que aproximadamente el 65 % de los aspirantes que indicaron no necesitar ayuda financiera completaron exitosamente su proceso de matrícula.
- **Influencia del Colegio de Procedencia:** El análisis destacó la importancia de la ubicación y el perfil del colegio. Se observó que la probabilidad de matrícula de los aspirantes provenientes de colegios con cercanía geográfica o convenios con la universidad, como el Colegio Berchmans o el Philadelphia Internacional que superan el 70 %.

- **Relevancia del Seguimiento y Contacto:** La interacción proactiva por parte del equipo de admisiones es un factor crítico. El modelo determinó que la probabilidad de que una oportunidad se considere como "Perdida" asciende al 40.2 % cuando el número de interacciones por WhatsApp es mínimo (0 a 1) y se han realizado menos de dos llamadas de seguimiento.
- **Correlación con el Estrato Socioeconómico:** Se confirmó que el estrato es un predictor significativo. A partir del estrato 3, se evidencia una correlación positiva directa: a mayor estrato socioeconómico, mayor es la probabilidad de que el aspirante se matricule en la universidad.

Estos hallazgos son fundamentales, pues permiten a los equipos de promoción y mercadeo segmentar a los aspirantes de manera más efectiva, personalizar las estrategias de comunicación y optimizar la asignación de recursos hacia los perfiles con mayor probabilidad de conversión.

3.4.7. Validación del Modelo Seleccionado con Datos Futuros

Con el fin de evaluar la capacidad de generalización del modelo **XGBoost** seleccionado, se realizó una prueba de validación adicional utilizando un conjunto de datos completamente nuevo y no visto durante el entrenamiento. Este conjunto correspondía a **296 registros de aspirantes del periodo académico 2025-2**, lo que permitió simular el comportamiento del modelo en un escenario operativo real.

Para esta prueba, se aplicó el umbral de clasificación recomendado por la propia herramienta, fijado en 52.2, el cual define el punto a partir del cual una predicción se considera como "Ganada". Al procesar los 296 registros con el modelo predictivo, se obtuvieron los siguientes resultados:

- **Predicciones Correctas:** 203 registros.
- **Predicciones Incorrectas:** 93 registros.

Estos resultados equivalen a una **precisión general (accuracy) del 68.58 %** sobre datos futuros. Este desempeño obtenido sugiere que el modelo posee un nivel razonable de generalización, siendo capaz de identificar patrones útiles en escenarios distintos al entrenamiento inicial. Estos resultados respaldan su potencial para ser incorporado en procesos institucionales de análisis de aspirantes, aunque podrían mejorarse con futuras iteraciones y un mayor enriquecimiento de los datos disponibles.

Capítulo 4

Conclusiones y Proyección a Futuro

4.1. Conclusiones generales

El presente proyecto logró desarrollar un modelo predictivo, utilizando las herramientas de inteligencia artificial de Salesforce, para estimar la probabilidad de que un aspirante se matricule en la Pontificia Universidad Javeriana Cali. La implementación se guió por la metodología CRISP-DM, que proporcionó un marco estructurado para abordar el problema desde la comprensión del negocio y los datos hasta la evaluación y el análisis de los resultados del modelo. El resultado es una herramienta analítica que permite a la universidad anticipar decisiones de matrícula y obtener hallazgos significativos, transformando datos históricos en conocimiento estratégico.

El análisis comparativo entre modelos de clasificación evidenció que la configuración binaria (Ganada y Perdida) con el algoritmo **XGBoost** ofreció el mejor desempeño, superando a Random Forest y GBM en métricas clave para contextos con desbalance de clases, como la puntuación F1 (0.833) y el coeficiente de correlación de Matthews (0.7114).

Este modelo demostró también una buena capacidad de generalización al ser validado con datos reales de aspirantes del periodo académico 2025-2, alcanzando una **precisión general del 68.58 %**. Lo anterior sugiere su viabilidad para ser utilizado como herramienta de apoyo en procesos de matrícula. No obstante, se detectaron oportunidades de mejora asociadas a la calidad y completitud de los datos, especialmente en variables con alto impacto como el nivel educativo del padre o la madre, que presentaban vacíos significativos. Abordar estas deficiencias podría potenciar aún más la efectividad del modelo.

La contribución de este trabajo para la Pontificia Universidad Javeriana Cali es

de alto valor estratégico, potenciado por el uso de herramientas nativas de Salesforce, lo cual permite una integración del modelo en los procesos de negocio, acceso en tiempo real a los datos y decisiones más ágiles, sin necesidad de desarrollos externos complejos. Esta propuesta fue bien recibida por directivos como Alberto Arias (Director de Relacionamiento), Luis Eduardo Rojas (Jefe de Desarrollo e Innovación TI), Alexander Valencia (Jefe del CRM), Carlos Augusto Gutiérrez (Coordinador de Análisis de Datos) y Carlos Arturo Dominguez (Ingeniero de proyectos de Desarrollo TI) quienes destacaron el impacto organizacional del proyecto. Reconocieron que los hallazgos del modelo respaldan preocupaciones previamente identificadas por los equipos funcionales y subrayan la importancia de ciertas variables clave. Además, valoraron el potencial del proyecto para democratizar el análisis de datos, facilitando su uso en áreas no técnicas y visibilizando capacidades de Salesforce que, como señaló Carlos Augusto, “hasta ahora se encontraban subutilizadas”.

4.2. Lecciones Aprendidas

El desarrollo de este proyecto permitió una aproximación práctica a un problema real con datos sensibles y de gran relevancia para la universidad, como es el caso de la matrícula de aspirantes. Esta experiencia proporcionó aprendizajes tanto desde lo técnico como desde lo organizacional, especialmente al abordar desafíos propios de un entorno institucional complejo y reglamentado.

- **Valor de una prueba de concepto aplicada:** La implementación de un modelo predictivo funcional demostró que es posible aportar soluciones concretas a problemáticas relevantes para la universidad, como la optimización del proceso de captación de estudiantes. Esta experiencia refuerza la importancia de desarrollar proyectos con un enfoque práctico y orientado al impacto.
- **Manejo de datos reales y sensibles:** El trabajo con información de aspirantes implicó una responsabilidad ética y técnica, dada la naturaleza sensible de estos datos. Esto requirió tener un cuidado y uso responsable de la información, elementos fundamentales en cualquier proyecto de analítica en entornos institucionales.
- **Limitaciones operativas por protocolos institucionales:** Algunas etapas del desarrollo debieron ser ajustadas en sus tiempos debido a protocolos internos de seguridad y aprobaciones requeridas para el acceso a ciertos entornos de trabajo o información. Esto resaltó la necesidad de planear con flexibilidad en contextos reales de implementación.
- **Desafíos en la calidad de los datos:** A diferencia de la mayoria de ejercicios académicos, los datos reales presentaron problemas de completitud, consistencia y formato, lo cual demandó un análisis exploratorio riguroso y un proceso de preprocesamiento cuidadoso. Esta etapa fue crítica para garantizar un entrenamiento válido del modelo y subraya la importancia de la ingeniería de datos en cualquier proyecto de analítica.
- **Interdisciplinariedad y comunicación con usuarios clave:** La comprensión del problema y la validación de resultados exigieron una interacción constante con actores institucionales. Esto evidenció que el éxito de una solución de IA no depende solo del desempeño técnico del modelo, sino también de su alineación con las necesidades y restricciones del entorno de aplicación.

4.3. Implicaciones Éticas

El modelo predictivo desarrollado en este proyecto involucra consideraciones éticas relevantes, ya que su uso puede impactar directamente en la gestión de admisiones y en las decisiones estratégicas de la Universidad:

Privacidad y protección de datos: El modelo utiliza información personal, académica y socioeconómica de los aspirantes. Este tratamiento de datos ya cumple con la Ley 1581 de 2012 sobre protección de datos personales, dado que todo el proceso se realiza dentro de un entorno controlado y seguro proporcionado por la Universidad, utilizando Salesforce como plataforma de gestión. Esto asegura que la información se maneje bajo estrictos protocolos de confidencialidad y acceso restringido.

Sesgos algorítmicos: Algunas de las variables utilizadas, como el estrato socioeconómico, el departamento de residencia o el puntaje de las pruebas ICFES, pueden estar asociadas a desigualdades preexistentes. Esto implica el riesgo de que el modelo reproduzca dichos sesgos. Por ello, es fundamental realizar evaluaciones periódicas que permitan detectar y corregir cualquier tendencia discriminatoria, asegurando que las predicciones se mantengan justas e inclusivas para todos los aspirantes.

Uso responsable y trato justo a los aspirantes: Las predicciones del modelo deben emplearse como una herramienta de apoyo al equipo de admisiones, complementadas siempre con la valoración humana y factores cualitativos. Un uso inadecuado podría derivar en estrategias de comunicación invasivas o en la exclusión de aspirantes con bajas probabilidades estimadas. Por ello, las acciones basadas en los resultados deben ser éticamente justificadas, garantizando un trato equitativo y orientado a mejorar la experiencia de todos los aspirantes.

En resumen, esta solución constituye una herramienta estratégica para mejorar procesos, pero demanda un ejercicio ético sostenido en el tiempo.

4.4. Trabajo Futuro

Este proyecto sienta las bases para el uso de herramientas de inteligencia artificial en la predicción de matrícula de aspirantes, integrando datos institucionales dentro del entorno de Salesforce. Sin embargo, su alcance también evidencia oportunidades de mejora y expansión, tanto en el plano técnico como en su aplicabilidad institucional. A continuación, se presentan algunas líneas de extensión que podrían fortalecer y ampliar los resultados obtenidos:

4.4.1. Implementación del modelo en entorno de producción

Como trabajo futuro, se propone llevar el modelo predictivo a un entorno de producción dentro del ecosistema de Salesforce. Para ello, el primer requisito es contar con la licencia CRM Analytics Plus, cuyo valor es de aproximadamente $165 USD mensuales. Una vez adquirida, se podría implementar el modelo siguiendo un flujo sencillo y reproducible:

1. Definir las variables clave a utilizar (las cuales ya fueron identificadas durante este proyecto).
2. Exportar dichas variables desde Salesforce en formato Excel, utilizando la consulta SOQL incluida en el Anexo #1.
3. Procesar los datos mediante el script de preprocesamiento en Python desarrollado, que limpia, transforma e imputa los valores necesarios. Este script, junto con su explicación detallada, se encuentra en el Anexo #2.
4. Importar el archivo CSV resultante en Analytics Studio.
5. Entrenar y evaluar el modelo predictivo con la herramienta Prediction Builder, seleccionando la variable objetivo correspondiente al estado de la oportunidad.

Este flujo constituye una solución funcional que puede ponerse en marcha a corto plazo, ya que una vez se cuente con la licencia, no requiere desarrollos adicionales y utiliza herramientas y procesos ya probados durante el proyecto.

A mediano plazo, se podría avanzar hacia una mayor automatización e integración dentro del entorno Salesforce. Para ello, se recomienda:

- Reemplazar el procesamiento externo en Python por el uso de Recipes (Recetas) en Analytics Studio, herramienta que permite transformar, filtrar, imputar y preparar datasets dentro de Salesforce sin necesidad de escribir código.

- Utilizar Salesforce Flows para activar procesos automáticamente cuando se creen o actualicen registros relevantes, lo cual permitiría alimentar el modelo predictivo de forma continua.
- Configurar una ingesta directa desde objetos estándar o personalizados (como `Lead` u `Opportunity`) hacia Analytics Studio, eliminando la necesidad de exportaciones manuales.

En el proyecto actual no fue posible utilizar Recipes ni Flows debido a las limitaciones de la versión gratuita, particularmente en el tratamiento de variables como el Puntaje Global Tx (ICFES), que requerían técnicas más avanzadas, como la agrupación condicional en subgrupos con base en otras variables. Estas transformaciones no pudieron ser implementadas directamente desde las herramientas integradas de Salesforce.

Sin embargo, con una mejora progresiva en la calidad de los datos y el acceso a funcionalidades completas como flows personalizados, se espera que en el futuro el preprocesamiento pueda realizarse íntegramente dentro del CRM, eliminando la necesidad de herramientas externas.

4.4.2. Ampliación del modelo predictivo hacia etapas posteriores del proceso de admisión

Se propone extender el alcance del modelo más allá de la decisión de matrícula, hacia aspectos como la permanencia estudiantil o el rendimiento académico, lo cual permitiría un acompañamiento más completo del ciclo de vida del estudiante.

4.4.3. Integración operativa del modelo en las estrategias de admisión

Una posible línea de trabajo consiste en desarrollar un proceso que permita exponer los resultados del modelo predictivo directamente en el entorno operativo de Salesforce, a través de objetos personalizados o campos calculados que estén disponibles para los asesores durante el seguimiento de aspirantes. Por ejemplo, se podría crear un campo en el objeto de Oportunidad que almacene la probabilidad de matrícula generada por el modelo, acompañado de una etiqueta categórica (como “Alta”, “Media”, “Baja”) que facilite su interpretación.

Asimismo, se podrían configurar flujos de trabajo (*Flows*) o reglas de automatización (*Process Builder*) para generar alertas o tareas cuando la probabilidad sea baja, permitiendo así priorizar acciones de acompañamiento. Esta integración requeriría articular los resultados del modelo con los objetos existentes en Salesforce y

garantizar que la información se actualice con la frecuencia necesaria para mantener su utilidad operativa.

4.4.4. Mejora en la Calidad y Completitud de los Datos

El análisis exploratorio de los datos reveló que varias de las variables con potencial predictivo presentaban un porcentaje considerable de valores ausentes. Campos como el *Puntaje ICFES*, la solicitud de *Ayuda financiera* y la información laboral de los padres, entre otros, requirieron la aplicación de técnicas de imputación para poder ser utilizados en el entrenamiento del modelo.

Por lo tanto, se recomienda como una línea de trabajo fundamental la implementación de estrategias para mejorar la calidad y completitud de los datos desde su origen. Esto podría incluir la configuración de campos obligatorios en los formularios de Salesforce, la creación de reglas de validación para garantizar la coherencia de la información y la capacitación de los equipos de admisión para asegurar un diligenciamiento más riguroso durante el contacto con los aspirantes.

Bibliografia

[1] Salesforce, "¿Qué es Salesforce?" Salesforce, 2025.

[2] F. Buttle y S. Maklan, *Customer Relationship Management: Concepts and Technologies. Routledge*, 2019.

[3] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, Cengage Learning, 2020.

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

[5] J. Han, M. Kamber, & J. Pei *Data Mining: Concepts and Techniques*, Morgan Kaufmann, 2011.

[6] G. Shmueli & O. R. Koppius, *Predictive Analytics in Information Systems Research*, MIS Quarterly, sep. 2011

[7] P. Chapman et al., *CRISP-DM 1.0: Step-by-step data mining guide*, The CRISP-DM Consortium, 2000.

[8] Salesforce.org, *Education Data Architecture (EDA)*, Salesforce.org, 2024.

[9] W. McKinney, *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and Jupyter*, O'Reilly Media, 2022.

[10] Y. Bengio and Y. Grandvalet, "No unbiased estimator of the variance of k-fold cross-validation," Journal of Machine Learning Research, vol. 5, pp. 1089–1105, 2004.

[11] J. M. Ph.D and E. Kavlakoglu, "Matriz de confusión," Ibm.com, Jan. 19, 2024. https://www.ibm.com/es-es/think/topics/confusion-matrix.

[12]V. Chugh, "AUC y Curva ROC en Aprendizaje Automático," Datacamp.com, Oct. 2024. https://www.datacamp.com/es/tutorial/auc.

[13] Vive UNIR, "¿Qué es el coeficiente de Gini y cómo calcularlo?," UNIR, Aug. 04, 2023. https://www.unir.net/revista/ciencias-sociales/coeficiente-gini.

[14] R. Kundu, “F1 Score in Machine Learning: Intro & Calculation,” V7, Dec. 16, 2022. https://www.v7labs.com/blog/f1-score-guide.

[15] “What is Matthews Correlation Coefficient (MCC),” www.activeloop.ai. https://www.activeloop.ai/resources/glossary/matthews-correlation-coefficient-mcc.

[16] Ibm, “Random Forest,” IBM, Feb. 27, 2025. https://www.ibm.com/mx-es/think/topics/random-forest.

[17] GeeksforGeeks, “Random Forest Algorithm in Machine Learning,” [Online]. Available: https://www.geeksforgeeks.org/machine-learning/random-forest-algorithm-in-machine-learning/. [Accessed: Aug. 13, 2025].

[18] B. Clark and F. Lee, “What is Gradient Boosting?” IBM, Apr. 7, 2025. https://www.ibm.com/think/topics/gradient-boosting

[19] İ. Kılıç, “Gradient Boosting Machines (GBM) with Python Example,” Medium, Sep. 23, 2023.

[20] E. Kavlakoglu and E. Russi, “What is XGBoost?” IBM, May. 9, 2024. [Online]. https://www.ibm.com/think/topics/xgboost

Capítulo 5

Anexo

Anexo #1. Consulta SOQL para la extracción de variables

A continuación se presenta la consulta utilizada para extraer las variables relevantes relacionando los objetos Aplicación, Oportunidad y Cuenta:

```
SELECT
    Estrato__c,
    Ayuda_financiera__c,
    M_todo_de_financiacion__c,
    Edad__c,
    Sexo__c,
    Colegio__r.Name,
    Deseo_aplicar_a_generaci_n_E_Aspirante__c,
    Departamento_de_residencia__c,
    Madre_Labora__c,
    Padre_Labora__c,
    Cargo_madre__c,
    Cargo_padre__c,
    Madre_fallecida_desconocida__c,
    Padre_fallecido_desconocido__c,
    Puntaje_Global_Tx__c,
    ltimo_nivel_de_formaci_n_de_la_madre__c,
    ltimo_nivel_de_formaci_n_del_padre__c,
    Oportunidad_asociada__r.StageName,
    Oportunidad_asociada__r.A_o_per_odo_acad_mico__c,
    Oportunidad_asociada__r.Programa_Acad_mico__r.Name,
    Oportunidad_asociada__r.N_mero_de_llamadas__c,
    Oportunidad_asociada__r.N_mero_de_Whatsapp__c
FROM
```

```
    hed__Application__c

WHERE Oportunidad_asociada__r.Programa_Acad_mico__r.Nombre_Facultad__c =
'INGENIERÍA Y CIENCIAS'
AND (Oportunidad_asociada__r.StageName = 'No admitido'
OR Oportunidad_asociada__r.StageName = 'Cerrada Perdida'
OR Oportunidad_asociada__r.StageName = 'Desistió'
OR Oportunidad_asociada__r.StageName = 'Pagó')
```

Anexo #2. Script de preprocesamiento de datos

Este anexo contiene el código en Python desarrollado para realizar el preprocesamiento del conjunto de datos de aspirantes. El script incluye limpieza, imputación y transformación de variables con el fin de preparar los datos para el modelado.

Lectura del archivo CSV

```
def read_csv(file_name: str) -> pd.DataFrame:
    try:
        return pd.read_csv(file_name, sep=';', encoding='utf-8')
    except Exception as e:
        print(f"Error reading file: {e}")
        return None
```

Esta función permite cargar los datos desde un archivo CSV utilizando el delimitador ; y codificación UTF-8. En caso de error, se captura la excepción y se devuelve `None`.

Limpieza inicial de columnas y registros

```
def remove_unnecessary_columns(df: pd.DataFrame) -> pd.DataFrame:
    col_elim = [
        '_',
        'Colegio__r',
        'Oportunidad_asociada__r',
        'Oportunidad_asociada__r.A_o_per_odo_acad_mico__c',
        'Oportunidad_asociada__r.Programa_Acad_mico__r'
    ]
    col_found = [col for col in col_elim if col in df.columns]
    df.drop(columns=col_found, inplace=True)
    print(f"Columnas eliminadas.")
    return df
```

Se eliminan columnas irrelevantes para el análisis. Esto reduce la dimensionalidad del dataset y previene ruido en el modelo.

```
def filter_stratum(df: pd.DataFrame) -> pd.DataFrame:

    antes = df.shape[0]
    df = df[(df['Estrato__c'].notnull()) & (df['Estrato__c'] != 0)]
    print(f"Filtro por estrato completado: {antes - df.shape[0]}")
    return df
```

Esta función elimina registros cuyo estrato socioeconómico está vacío o es cero.

Tratamiento del puntaje ICFES

```
def clean_icfes_score(df: pd.DataFrame) -> pd.DataFrame:

    df['Puntaje_Global_Tx__c'] =
    pd.to_numeric(df['Puntaje_Global_Tx__c'], errors='coerce')
    fuera_rango = (df['Puntaje_Global_Tx__c'] < 1) |
                  (df['Puntaje_Global_Tx__c'] > 500)
    df.loc[fuera_rango, 'Puntaje_Global_Tx__c'] = pd.NA
    print(f"Limpieza del puntaje icfes completada.")
    return df
```

Se convierte el puntaje a valores numéricos y se marcan como nulos aquellos que no están en el rango permitido (1 a 500).

```
def impute_icfes_score(df: pd.DataFrame) -> pd.DataFrame:

    df["ICFES_faltante"] = df["Puntaje_Global_Tx__c"].isnull().astype(int)
    df["Puntaje_ICFES"] = df["Puntaje_Global_Tx__c"]

    # Imputar por colegio
    medianas_colegio = df.groupby("Colegio__r.Name")
    ["Puntaje_Global_Tx__c"].median()
    df["Puntaje_ICFES"] = df.apply(
        lambda row: medianas_colegio[row["Colegio__r.Name"]]
        if pd.isnull(row["Puntaje_ICFES"]) and row["Colegio__r.Name"]
        in medianas_colegio
        else row["Puntaje_ICFES"], axis=1
    )

    # Imputar por estrato
```

```
    medianas_estrato = df.groupby("Estrato__c")
    ["Puntaje_Global_Tx__c"].median()
    df["Puntaje_ICFES"] = df.apply(
        lambda row: medianas_estrato[row["Estrato__c"]]
        if pd.isnull(row["Puntaje_ICFES"]) and row["Estrato__c"]
        in medianas_estrato
        else row["Puntaje_ICFES"], axis=1
    )

    # Eliminación de columna original
    if 'Puntaje_Global_Tx__c' in df.columns:
        df.drop(columns=['Puntaje_Global_Tx__c'], inplace=True)

    print("Imputación de Puntaje ICFES completada.")
    return df
```

Los valores faltantes se imputan usando la mediana por colegio y, si sigue faltando, la mediana por estrato. Se genera además una columna binaria que indica si el dato original estaba ausente, tecnica conocida como *missing indicator*.

Manejo de valores nulos

```
def replace_null_values(df: pd.DataFrame) -> pd.DataFrame:

    no_responde = ['Ayuda_financiera__c', 'Madre_Labora__c',
                   'Padre_Labora__c',
                   'ltimo_nivel_de_formaci_n_de_la_madre__c',
                   'ltimo_nivel_de_formaci_n_del_padre__c']

    no_aplica = ['M_todo_de_financiaci_n__c', 'Cargo_madre__c',
                 'Cargo_padre__c', 'Madre_fallecida_desconocida__c',
                 'Padre_fallecido_desconocido__c']

    numericas = ['Oportunidad_asociada__r.N_mero_de_llamadas__c',
                 'Oportunidad_asociada__r.N_mero_de_Whatsapp__c']

    for col in [c for c in no_responde if c in df.columns]:
        df[col].fillna("No responde", inplace=True)

    for col in [c for c in no_aplica if c in df.columns]:
        df[col].fillna("No aplica", inplace=True)

    for col in [c for c in numericas if c in df.columns]:
        df[col].fillna(0, inplace=True)

    print("Reemplazo de Valores nulos completado.")
```

```
    return df
```

Rellena los valores nulos de variables categóricas con ‘‘No responde’’ o ‘‘No aplica’’ según el contexto, y los valores numéricos con cero.

Homogeneización de variables categóricas

```
def replace_opp_stagename(df: pd.DataFrame) -> pd.DataFrame:

    col = 'Oportunidad_asociada__r.StageName'
    if col in df.columns:
        df[col] = df[col].replace({
            'Pagó': 'Ganada',
            'Desistió': 'Perdida',
            'Cerrada Perdida': 'Perdida',
            'No admitido': 'Perdida'
        })
        print(f"Reemplazo de valores en '{col}' completado.")
    else:
        print(f"La columna '{col}' no existe.")
    return df
```

Simplifica los valores de la columna `StageName` a sólo dos categorías: `Ganada` y `Perdida`.

```
def group_unique_schools(df: pd.DataFrame) -> pd.DataFrame:

    if 'Colegio__r.Name' in df.columns:
        frecuencia = df['Colegio__r.Name'].value_counts()
        únicos = frecuencia[frecuencia == 1].index
        df['Colegio__r.Name'] =
        df['Colegio__r.Name'].apply(lambda x: 'OTRO' if x in únicos else x)
        print("Agrupación de Colegios como 'OTRO' completada.")
    return df
```

Los colegios que aparecen sólo una vez en el dataset se etiquetan como `OTRO`, lo cual ayuda a reducir la cardinalidad y evitar el sobreajuste.

Función principal y guardado

```
def csv_save(df: pd.DataFrame, nombre_salida: str):

    df.to_csv(nombre_salida, index=False, sep=';', encoding='utf-8-sig')
```

```
    print(f"Archivo guardado como: {nombre_salida}")


def data_proccesing(nombre_archivo: str,
                    nombre_salida: str = 'preprocessed_data2.csv'):

    df = csv_read(nombre_archivo)
    if df is None:
        return

    df = remove_unnecessary_columns(df)
    df = filter_stratum(df)
    df = clean_icfes_score(df)
    df = replace_null_values(df)
    df = replace_opp_stagename(df)
    df = group_unique_schools(df)
    df = impute_icfes_score(df)
    csv_save(df, nombre_salida)
```

csv_save() guarda el resultado del procesamiento en un nuevo archivo CSV codificado en UTF-8 para asegurar compatibilidad con Excel y otros programas.

Por otro lado, *data_proccesing()* es la función principal que ejecuta todos los pasos del preprocesamiento de forma secuencial y recibe el archivo que se va a procesar.

Ejecución del script

```
if __name__ == "__main__":
    data_proccesing('originalData.csv')
```

Permite ejecutar el script directamente desde la terminal, procesando automáticamente el archivo de entrada y generando el de salida.

Anexo #3. Acuerdo de confidencialidad

A continuación, se presenta el acuerdo de confidencialidad firmado que rige el uso de la información de los aspirantes y demás datos sensibles empleados durante el desarrollo del proyecto:

ACUERDO DE CONFIDENCIALIDAD Y RESERVA EN EL USO DE LA INFORMACIÓN

Entre los suscritos a saber: **VICENTE DURAN CASAS, S.J.,** mayor de edad, vecino de Cali e identificado con la cedula de ciudadanía No. 3.227.972 de Usaquen, quien obra en nombre y representación de la **PONTIFICIA UNIVERSIDAD JAVERIANA**, **SECCIONAL CALI**, institución de educación superior privada, de utilidad común, sin ánimo de lucro, con personería jurídica reconocida por medio de la Resolución N° 73 de 1933, emanada del Ministerio de Gobierno, reconocida como Universidad mediante el Decreto 1297 del 30 de mayo de 1964, sometida a la vigilancia del Ministerio de Educación Nacional, en virtud del poder general que le fue conferido por medio de la escritura pública 497 otorgada el 02 de abril de 2014 en la Notaría 26 del Círculo de Bogotá, quien para efectos de este documento se denominará **LA UNIVERSIDAD** por una parte, y por la otra **MIGUEL ÁNGEL SÁNCHEZ PÁEZ** Y **JOSÉ MANUEL GARCÍA LÓPEZ,** identificado (a) como aparece al pie de su firma y que para todos los efectos se llamarán **LOS ESTUDIANTES,** se celebra el presente acuerdo de confidencialidad y reserva en el uso de la Información, previo a:

CONSIDERACIONES

1. Como parte del desarrollo académico y formativo de los estudiantes de pregrado en Ingeniería de Sistemas y Computación, se establece como requisito para la obtención del título la realización de un proyecto de grado.
2. En este contexto, los estudiantes desarrollaron el anteproyecto "*Uso de IA en Salesforce para el desarrollo de un modelo predictivo de matrícula de aspirantes*" bajo la supervisión de la profesora Luisa Rincón, aplicando y fortaleciendo los conocimientos adquiridos a lo largo de su formación para abordar problemáticas y desafíos en el ámbito de la ingeniería de sistemas y computación.
3. Actualmente, los estudiantes se encuentran en la fase de elaboración de su trabajo de grado, para lo cual requieren acceso a información confidencial almacenada en nuestro CRM. Dicha información es fundamental para el desarrollo de su investigación y análisis.

Teniendo en cuenta lo anterior, las partes acuerdan celebrar el presente acuerdo de confidencialidad, el cual se regirá por las siguientes disposiciones:

CLÁUSULAS

PRIMERA.- Objeto: En virtud del presente acuerdo **LOS ESTUDIANTES** asumen la obligación de no revelar, divulgar, o exhibir información financiera, técnica, comercial o científica relacionada con los proyectos en los que participa o tenga conocimiento y otros relacionados o derivados de estos a cargo del Centro de Consultoría y Educación Continua de **LA UNIVERSIDAD**, a persona alguna natural o jurídica, (incluyendo parientes en cualquier grado de consanguinidad o afinidad), ni utilizar o emplear dicha información en su favor o en el de terceros, ya sea de manera directa o indirecta y en perjuicio o no de las anteriores.

SEGUNDA.- Autorización previa: Mediante la firma del presente documento, **LOS ESTUDIANTES** aceptan que sin la previa autorización escrita de **LA UNIVERSIDAD** u orden de autoridad competente en ejercicio de funciones legales y en desarrollo de actuación administrativa o judicial, no podrá revelar, divulgar, exhibir y, en general, dar a conocer a ningún tercero, información

o documento alguno que haya recibido o esté por recibir de parte de **LA UNIVERSIDAD**, en razón de sus funciones, relativo a la información legal, financiera, técnica o comercial relacionada con el giro ordinario de su quehacer o de los proyectos a desarrollar.

En consecuencia, **LOS ESTUDIANTES** reconocen y aceptan que **LA UNIVERSIDAD** y/o **LA EMPRESA** son titulares exclusivos de cualquier derecho sobre dicha información, y que no tendrá derecho alguno sobre la misma, obligándose a no copiarla, duplicarla, sustraerla o comunicarla, para sí o para terceros.

CUARTA.- Cláusula penal: En adición a lo establecido en la cláusula precedente, el incumplimiento por parte de **LOS ESTUDIANTES** les hará incurrir automáticamente en una pena a favor de **LA UNIVERSIDAD** por una suma equivalente cien (100) salarios mínimos legales mensuales vigentes al momento en que se dé el incumplimiento del presente acuerdo, sin perjuicio de que **LA UNIVERSIDAD** pueda adelantar las acciones legales que tiendan al resarcimiento total de los perjuicios causados. El pago de la suma estipulada como pena, no releva a **LOS ESTUDIANTES** del cumplimiento de la obligación de confidencialidad contraído mediante este acuerdo; en consecuencia, deberá pagar esta penalidad a **LA UNIVERSIDAD** cada vez que se presente un incumplimiento de esta índole. En los términos del inciso segundo del artículo 1595 del Código Civil, **LOS ESTUDIANTES** reconocen expresamente que incurrirá en la pena establecida en la presente cláusula a partir del momento en que ejecute cualquiera de los hechos o conductas que se ha obligado a abstenerse.

Parágrafo. Sin perjuicio de la pena pactada en la presente cláusula, **LOS ESTUDIANTES** se harán responsable en forma integral por todos los perjuicios que llegare a causar por el incumplimiento de las obligaciones contraídas dentro de los proyectos en los que intervenga, y por el mal manejo de la información confidencial que le sea confiada o a la que llegare a tener acceso cualquiera que sea su causa.

QUINTA.- Permanencia de las obligaciones: La terminación del contrato de prestación de servicios entre **LA UNIVERSIDAD** y **LOS ESTUDIANTES** no lo exonera del cumplimiento de las obligaciones previstas en este acuerdo, las cuales subsisten durante el tiempo de su permanencia en la universidad. En caso de incumplimiento de alguna de estas obligaciones dentro de dicho período, **LA UNIVERSIDAD** se reserva el derecho de reclamar, judicial o extrajudicialmente, la pena y la indemnización plena de los perjuicios causados.

SEXTA.- Vigencia: El presente acuerdo regirá durante la vigencia de los proyectos y cinco (5) años más contados a partir de su terminación, cualquiera que sea la causa que dé lugar a ella.

SÉPTIMA.- Mérito Ejecutivo: El presente documento junto con la prueba sumaria del incumplimiento por parte de **LOS ESTUDIANTES** prestará mérito ejecutivo para el cobro de las sanciones pecuniarias que en virtud de él se establecen. Bastará también para proceder al cobro ejecutivo, el original del presente acuerdo acompañado de copia auténtica de la carta donde la parte cumplida manifieste a la parte incumplida las causas y motivos de su incumplimiento.

OCTAVA.- Diferencias: Toda controversia o diferencia que surja entre las partes con ocasión de la interpretación, ejecución, modificación, suspensión, terminación o incumplimiento de este acuerdo se resolverá mediante conciliación. De no llegarse a un acuerdo conciliatorio, las diferencias se solucionarán por vía judicial.

La conciliación estará sujeta a las siguientes reglas:

a) La conciliación se llevará a cabo en el Centro de Arbitraje y Conciliación de la Cámara de Comercio de Cali.
b) El conciliador será escogido de común acuerdo de la lista de conciliadores del Centro de Arbitraje y Conciliación de la Cámara de Comercio de Cali.
c) La conciliación deberá intentarse dentro de los tres meses siguientes a la presentación en debida forma de la solicitud de conciliación.
d) Los gastos generados a partir del procedimiento de conciliación serán asumidos por partes iguales entre **LOS ESTUDIANTES** y **LA UNIVERSIDAD**.

NOVENA.- Notificaciones: Para los efectos a que haya lugar en el desarrollo del presente contrato, las partes recibirán notificaciones en las siguientes direcciones:

- **LA UNIVERSIDAD** en la calle 18 N.º 118-250, vía Pance de la ciudad de Cali (Valle).
- **LOS ESTUDIANTES**, en la:

Estudiante:	**Dirección:**	**Correo:**
Miguel Ángel Sánchez Páez	Calle 10 #22A-700	misancio@javerianacali.edu.co
José Manuel García López	Calle 28 #113-57	joseman1909@javerianacali.edu.co

En constancia de todo lo anterior, se firma en dos originales, en la ciudad de Santiago de Cali, el _____.

LA UNIVERSIDAD

_____
VICENTE DURAN CASAS, S.J.
Rector Seccional Cali

EL ESTUDIANTE

_____
Nombre: Miguel Ángel Sánchez Páez
C.C. 1117019086

EL ESTUDIANTE

_____
Nombre: José Manuel García López
C.C. 1006331152