MODELO PREDICTIVO DE RIESGOS DE DESERCIÓN ESTUDIANTIL PARA LA INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

DANIEL JAIME FLÓREZ AGUIRRE

Trabajo de grado para optar al título de Ingeniero de Software

Asesor:
PhD. Juan Carlos Briñez de León
MsC. Yudy Andrea Quintero Tangarife

INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO
FACULTAD DE INGENIERÍA
INGENIERÍA DE SOFTWARE
MEDELLÍN
2025

Tabla de Contenido

Contenido

Tabla de ilustraciones

Lista de Tablas

Resumen

MODELO PREDICTIVO DE RIESGOS DE DESERCIÓN ESTUDIANTIL PARA LA INSTITUCIÓN UNIVERSITARIA PASCUAL BRAVO

DANIEL JAIME FLOREZ AGUIRRE

La deserción universitaria es un problema complejo con repercusiones económicas, sociales y personales significativas. Los altos índices de abandono escolar representan una pérdida de recursos financieros y oportunidades de movilidad social, perpetuando ciclos de pobreza y desigualdad. Identificar tempranamente a los estudiantes en riesgo de abandonar permite implementar programas de apoyo específicos, optimizar recursos, y mejorar las políticas educativas, lo que fortalece el compromiso estudiantil y mejora las tasas de retención. La intervención temprana es crucial para abordar este fenómeno de manera efectiva, beneficiando tanto a los individuos como a la sociedad en su conjunto.

Este proyecto propone el desarrollo de un modelo predictivo de deserción académica, utilizando técnicas avanzadas de modelado de datos. El objetivo es generar información relevante que permita la creación de alertas preventivas y tempranas para mitigar este fenómeno. Se espera que este modelo contribuya a la identificación de los factores clave que influyen en la deserción, ofreciendo a la institución herramientas más eficaces para intervenir antes de que los estudiantes abandonen sus estudios. Este enfoque no solo busca reducir las tasas de deserción, sino también mejorar la eficiencia en la utilización de recursos públicos destinados a la educación superior.

Utilizando técnicas avanzadas de análisis de datos y modelado estadístico, se buscará generar alertas tempranas que ayuden a las instituciones educativas a implementar estrategias de prevención más efectivas. Este estudio se estructurará en cuatro fases principales: la recopilación y limpieza de datos, el desarrollo y ajuste del modelo, y la validación de este con datos históricos de estudiantes. Este trabajo se inscribe en el marco del semillero de investigación SAMDATA, donde participa el investigador externo Juan Carlos Briñez de León.

La solución planteada se apoya en el análisis de datos históricos de la institución, integrando variables académicas, socioeconómicas, demográficas y comportamentales para construir un modelo de riesgo. Como producto final, se espera generar una API que integre este modelo con los sistemas institucionales, permitiendo el monitoreo automático del riesgo de deserción y facilitando la toma de decisiones.

Palabras claves: Deserción universitaria, Modelado predictivo, Técnicas de análisis de datos, Alertas tempranas, Estrategias de prevención, Eficiencia de recursos públicos, Educación superior.

Abstract

University dropout is a complex issue with significant economic, social, and personal repercussions. High dropout rates represent a loss of financial resources and opportunities for social mobility, perpetuating cycles of poverty and inequality. Early identification of students at risk of dropping out makes it possible to implement targeted support programs, optimize resource allocation, and improve educational policies—thereby strengthening student engagement and improving retention rates. Early intervention is crucial to effectively address this phenomenon, benefiting both individuals and society as a whole.

This project proposes the development of a predictive model for academic dropout, using advanced data modeling techniques. The objective is to generate relevant information that enables the creation of early and preventive alerts to mitigate this issue. The model is expected to contribute to identifying the key factors that influence dropout, providing the institution with more effective tools to intervene before students abandon their studies. This approach aims not only to reduce dropout rates but also to improve the efficiency in the use of public resources allocated to higher education.

Using advanced data analysis and statistical modeling techniques, the project aims to generate early warnings that help educational institutions implement more effective prevention strategies. This study will be structured in four main phases: data collection and cleaning, model development and tuning, and validation using historical student data. This work is part of the research group SAMDATA, which includes the participation of external researcher Juan Carlos Briñez de León.

The proposed solution is based on the analysis of the institution's historical data, integrating academic, socioeconomic, demographic, and behavioral variables to build a risk model. As a final product, the project aims to develop an API that integrates the model with institutional systems, enabling automatic monitoring of dropout risk and supporting informed decision-making.

Keywords:University dropout, Predictive modeling, Data analysis techniques, Early warning alerts, Prevention strategies, Public resource efficiency, Higher education.

Glosario

Aprendizaje Automático (Machine Learning):
Subcampo de la inteligencia artificial que se enfoca en el desarrollo de algoritmos y modelos que permiten a los sistemas aprender automáticamente a partir de datos, sin necesidad de ser programados explícitamente. A través del aprendizaje automático, un sistema puede identificar patrones, realizar predicciones y mejorar su rendimiento con el tiempo a medida que se expone a nuevos datos. En este proyecto, se emplearon algoritmos de clasificación supervisada para predecir la deserción estudiantil.

API (Application Programming Interface):
Interfaz de programación de aplicaciones que permite la comunicación e intercambio de datos entre distintos sistemas o aplicaciones. Una API define métodos estandarizados para enviar y recibir información, lo cual facilita la integración de modelos predictivos con plataformas web o sistemas institucionales. En este trabajo se desarrolló una API para consultar el modelo predictivo desde una interfaz web.

Deserción Estudiantil:
Fenómeno por el cual un estudiante interrumpe su trayectoria académica y abandona el programa de estudios sin haberlo finalizado. Las causas pueden ser múltiples: económicas, personales, académicas o institucionales. La predicción de este evento permite a las instituciones implementar estrategias preventivas para mejorar la permanencia estudiantil.

Imputación de Datos:
Técnica utilizada en el procesamiento de datos que consiste en reemplazar valores faltantes por estimaciones calculadas a partir de la información disponible. En este proyecto se utilizó la imputación mediante el algoritmo K-Nearest Neighbors (KNN), lo que permitió reconstruir un conjunto de datos más completo y adecuado para el entrenamiento de modelos de aprendizaje automático.

Modelo Predictivo:

Algoritmo o sistema computacional que, a partir de datos históricos, estima o anticipa la ocurrencia de eventos futuros. En este caso, se entrenaron modelos para predecir la probabilidad de que un aspirante o estudiante deserte del programa. Los modelos generan salidas como probabilidades o clases (deserta/no deserta) en función de los datos de entrada.

Retención Estudiantil:

Indicador que refleja la capacidad de una institución educativa para mantener a sus estudiantes matriculados hasta la culminación de sus programas académicos. Mejorar la retención implica comprender y mitigar los factores que conducen a la deserción. Los sistemas predictivos pueden contribuir significativamente a este objetivo, al identificar estudiantes en riesgo y permitir intervenciones oportunas.

Voto de Clasificadores (Voting Classifier):

Técnica de ensamblado en aprendizaje automático que combina múltiples modelos base para generar una única predicción más robusta. Puede aplicarse mediante votación dura (por mayoría de clases predichas) o votación blanda (promedio de probabilidades). Esta estrategia permite mejorar la estabilidad y el rendimiento general del sistema predictivo.

Joblib:

Librería de Python utilizada para la serialización y almacenamiento eficiente de objetos grandes, como modelos entrenados de machine learning. Permite guardar y cargar modelos para reutilizarlos en producción sin necesidad de volver a entrenarlos, lo cual es clave para el despliegue eficiente de soluciones basadas en ML.

Frontend Web:

Parte visible de una aplicación o sistema web con la que interactúan los usuarios finales. En este proyecto, se desarrolló un frontend que permite a los usuarios acceder a la herramienta predictiva a través de una interfaz gráfica sencilla e intuitiva, conectada al modelo mediante una API.

XGBoost (Extreme Gradient Boosting):

Algoritmo de aprendizaje automático basado en árboles de decisión optimizados mediante técnicas de boosting. Es ampliamente utilizado por su alta precisión, capacidad de manejo de datos faltantes y eficiencia en tareas de clasificación y regresión. Fue uno de los modelos con mejor desempeño en este trabajo.

Random Forest:

Algoritmo de aprendizaje automático que utiliza conjuntos de árboles de decisión entrenados con diferentes subconjuntos de datos y características. Su naturaleza de conjunto (bagging) lo hace resistente al sobreajuste y altamente efectivo en problemas de clasificación como el abordado en este proyecto.

Capítulo 1: Introducción

1.1 Contexto: La deserción académica universitaria: relevancia, impacto y nuevas perspectivas analíticas

La deserción universitaria es uno de los fenómenos más complejos, persistentes y preocupantes que enfrenta actualmente la educación superior en América Latina y, particularmente, en Colombia. Este fenómeno no solo compromete la eficiencia del sistema educativo, sino que representa un fracaso en la realización del derecho a la educación como herramienta para la equidad, el desarrollo humano y la transformación social. Se trata de un proceso multifactorial que implica el abandono parcial o definitivo por parte de un estudiante de su trayectoria educativa sin haber obtenido el título correspondiente, y que puede manifestarse en cualquier etapa de la formación profesional, aunque es más frecuente durante los primeros semestres académicos.

La deserción universitaria no puede ser entendida simplemente como una decisión individual desligada de su contexto. Por el contrario, constituye la expresión de un entramado de factores estructurales, institucionales y personales que interactúan de manera compleja: condiciones económicas precarias, falta de orientación vocacional, bajo rendimiento académico, dificultades psicosociales, debilidades en los modelos pedagógicos, y contextos familiares y socioculturales desfavorables. Estas dimensiones, lejos de actuar de forma aislada, se superponen y refuerzan mutuamente, generando un escenario de vulnerabilidad educativa para miles de jóvenes.

Es un problema reconocido globalmente, con tasas de deserción significativas reportadas para diferentes regiones a nivel global; Un ejemplo de esto es el reporte de la OCDE con tasas promedio cercanas al 20% en 2019, mientras que en Estados Unidos la deserción para programas de pregrado alcanzó el 40% para el 2020 (Liu et al., 2025).

En Colombia, pese a los avances en cobertura, las cifras de abandono siguen siendo preocupantes. El Ministerio de Educación Nacional (MEN) ha reportado históricamente tasas

cercanas al 50% por cohorte, lo que indica que la mitad de los estudiantes que ingresan a la educación superior no culminan sus estudios dentro del tiempo previsto ni fuera de él. Esto no solo genera pérdidas económicas para los estudiantes y sus familias, sino que también afecta la planeación y sostenibilidad de las instituciones educativas, así como el desarrollo social y económico del país, al disminuir el número de profesionales capacitados que pueden aportar al crecimiento de los sectores productivos y sociales.

La deserción universitaria también compromete la legitimidad y el cumplimiento del mandato social de las instituciones de educación superior, que deben garantizar no solo el acceso sino también la permanencia y el egreso exitoso de sus estudiantes. Frente a esta problemática, surge la necesidad de adoptar enfoques analíticos más robustos e integrales que permitan entender mejor las causas del abandono y predecir con mayor precisión los perfiles de riesgo. En este sentido, las nuevas tendencias de análisis basadas en datos, incluyendo técnicas de minería de datos, inteligencia artificial y aprendizaje automático, están ofreciendo herramientas innovadoras para enfrentar el problema de manera más efectiva, anticipando comportamientos, identificando patrones y diseñando intervenciones preventivas.

Así, la comprensión y abordaje de la deserción académica universitaria exige una mirada interdisciplinaria, que articule la investigación educativa, la política pública y las capacidades tecnológicas emergentes para avanzar hacia un sistema de educación superior más inclusivo, equitativo y resiliente.

1.2 Importancia y consecuencias

El abandono universitario representa un obstáculo importante en el logro de la cobertura, calidad y equidad en la educación superior. Las cifras reportadas en el contexto colombiano, por ejemplo, evidencian que casi el 50% de los estudiantes desertan antes de finalizar sus estudios, lo que indica no solo una pérdida de recursos invertidos, sino también una reducción del capital humano cualificado en el país (Vélez, 2020; Castro-Montoya et al., 2021). Esta situación tiene efectos acumulativos sobre el desarrollo económico y social del país, pues debilita la capacidad de innovación, reduce la productividad laboral futura y limita las posibilidades de movilidad social ascendente para los jóvenes más vulnerables.

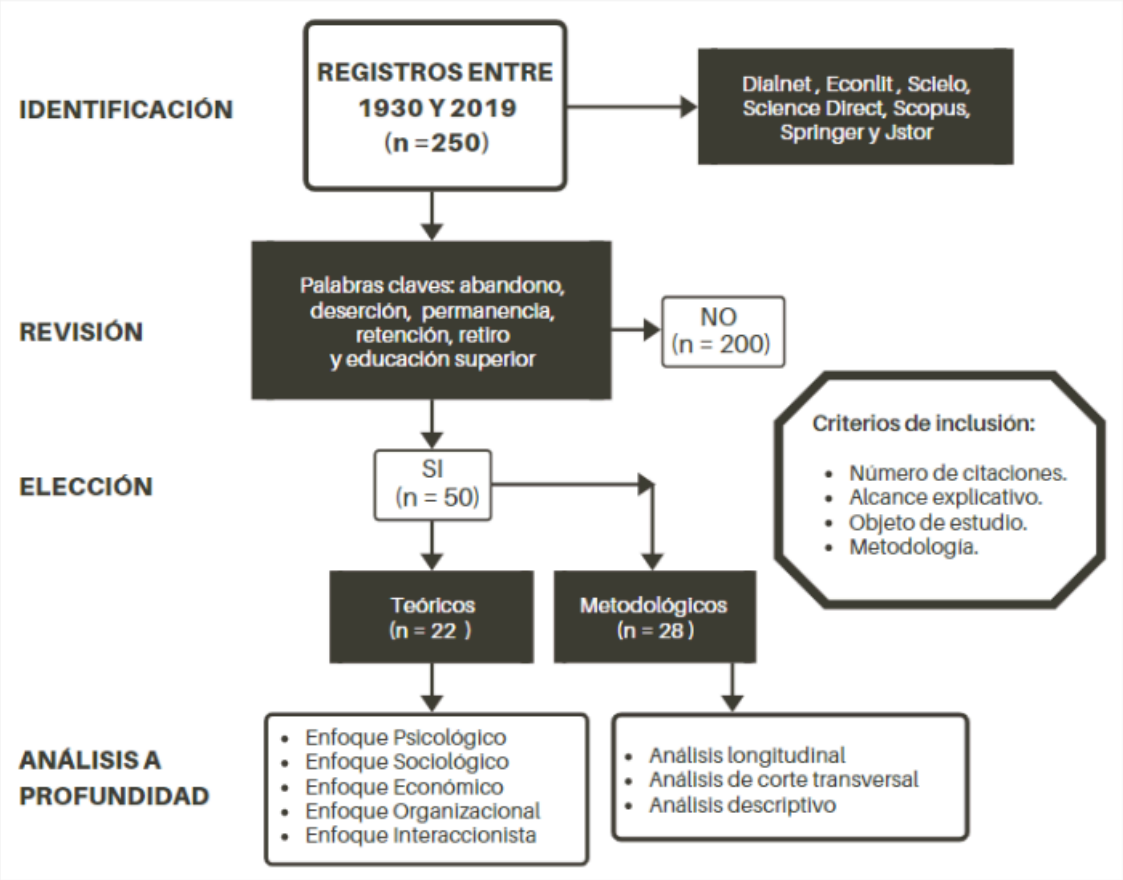

Ilustración 1 - ***Gasto Total del Sector Público en Educación Superior, tomado de (Vélez, 2020).***

A nivel institucional, la deserción afecta la planeación financiera, la eficiencia de los programas y la capacidad de las IES de cumplir con su misión formativa. Además, genera una pérdida en el retorno de la inversión realizada en procesos de admisión, formación inicial y programas de bienestar estudiantil. En muchos casos, el abandono también repercute negativamente en indicadores de calidad y acreditación institucional, lo que impacta la reputación y competitividad de las universidades.

Desde la perspectiva individual, los efectos son igual de significativos. El abandono puede generar frustración, desmotivación y perpetuar ciclos de pobreza al impedir el acceso a mejores oportunidades laborales (Moreno, 2022; Erazo & Rosero, 2021). Muchos estudiantes que abandonan sus estudios lo hacen sin haber desarrollado competencias certificables o sin adquirir una formación profesional que les permita acceder a empleos de mayor calidad. Esto los expone

a condiciones laborales precarias, subempleo o informalidad, y contribuye a la segmentación del mercado laboral. También representa una inversión fallida para las familias, muchas veces realizada con grandes sacrificios económicos.

Por otro lado, la deserción universitaria puede tener efectos emocionales y psicológicos duraderos. Los estudiantes que abandonan sus estudios a menudo experimentan sentimientos de fracaso, baja autoestima, presión social y ansiedad respecto a su futuro. Estas consecuencias afectan su bienestar personal y pueden disminuir su motivación para retomar estudios en el futuro o para participar activamente en procesos formativos continuos. La comprensión profunda de estas consecuencias resulta clave para el diseño de políticas de intervención que promuevan no solo la permanencia, sino también el bienestar integral del estudiante universitario.

1.2 Factores determinantes

El fenómeno ha sido abordado desde múltiples enfoques teóricos: psicológicos, sociológicos, económicos y organizacionales. En términos generales, los factores que influyen en la deserción pueden clasificarse como académicos, institucionales, personales, socioeconómicos y contextuales (Castro-Montoya et al., 2021).

Desde la perspectiva académica, el bajo rendimiento, la sobrecarga académica, la falta de acompañamiento pedagógico y la desconexión entre los contenidos del programa y los intereses del estudiante aparecen como causas recurrentes de abandono. Estos factores suelen intensificarse cuando los estudiantes no cuentan con una base sólida de competencias adquiridas en la educación media o no desarrollan habilidades de autorregulación del aprendizaje en la universidad.

A nivel institucional, la ausencia de políticas claras de permanencia, la escasa articulación entre bienestar y docencia, la rigidez curricular y la insuficiente oferta de apoyos financieros, inciden negativamente en la trayectoria estudiantil. Asimismo, las estrategias de admisión y nivelación académica juegan un rol crucial en la capacidad de las IES para retener a sus estudiantes más vulnerables.

En el plano personal y familiar, variables como la falta de motivación intrínseca, los problemas de salud mental, las responsabilidades familiares y laborales, así como las expectativas poco

realistas sobre la vida universitaria, son determinantes importantes. Estas condiciones tienden a afectar de forma más significativa a los estudiantes de primera generación en educación superior.

El entorno socioeconómico, por su parte, impacta directamente en la permanencia. El nivel de ingreso del hogar, el estrato socioeconómico, el acceso a recursos tecnológicos y la ubicación geográfica son factores que amplifican o mitigan el riesgo de abandono. En zonas rurales o marginadas, por ejemplo, la falta de infraestructura adecuada y las brechas digitales actúan como barreras adicionales para la continuidad académica.

Estudios recientes, como los realizados en el contexto de Medellín, han identificado estos determinantes mediante modelos estadísticos como el análisis de supervivencia y componentes principales, revelando patrones significativos entre el perfil del estudiante y su probabilidad de desertar (Vélez, 2020).

1.3 Estado del arte: una perspectiva integral

La literatura revisada en el periodo 2015-2022 evidencia una evolución en los enfoques de análisis del abandono universitario. Mientras que estudios previos se enfocaban en descripciones y categorizaciones, las investigaciones más recientes incorporan métodos cuantitativos y modelamientos predictivos para una comprensión más profunda del fenómeno (Moreno, 2022; Gutiérrez & López, 2021).

Una de las tendencias más destacadas ha sido el uso de modelos de análisis multivariado, análisis de cohortes, análisis de supervivencia y minería de datos educativa para identificar perfiles de riesgo y predecir la deserción. Estos enfoques han permitido no solo confirmar factores previamente identificados, sino también descubrir nuevas correlaciones, como la relación entre la carga académica semestral y el nivel de satisfacción institucional.

Además, ha surgido un interés creciente por incluir variables cualitativas, como la percepción de apoyo institucional, el sentido de pertenencia, la autopercepción del desempeño académico y la motivación vocacional. Esto ha enriquecido los modelos explicativos y ha permitido diseñar estrategias de intervención más contextualizadas y sensibles a las realidades del estudiante.

En este sentido, se ha hecho visible la necesidad de diferenciar entre tipos de deserciones (voluntaria, forzada, temprana, tardía), así como entre factores modificables y estructurales. Esta

distinción resulta clave para priorizar recursos y diseñar intervenciones oportunas que impacten efectivamente los niveles de retención estudiantil.

Además, la inclusión del enfoque de permanencia estudiantil ha permitido no solo estudiar las causas del abandono, sino también proponer acciones preventivas orientadas a fortalecer la trayectoria académica desde el ingreso, a través de programas de inducción, mentoría, acompañamiento académico, psicológico y financiero (Castro-Montoya et al., 2021; Erazo & Rosero, 2021).

La literatura revisada en el periodo 2015-2022 evidencia una evolución en los enfoques de análisis del abandono universitario. Mientras que estudios previos se enfocan en descripciones y categorizaciones, las investigaciones más recientes incorporan métodos cuantitativos y modelamientos predictivos para una comprensión más profunda del fenómeno (Moreno, 2022; Gutiérrez & López, 2021).

En este sentido, se ha hecho visible la necesidad de diferenciar entre tipos de deserciones (voluntaria, forzada, temprana, tardía), así como entre factores modificables y estructurales. Además, la inclusión del enfoque de permanencia estudiantil ha permitido no solo estudiar las causas del abandono, sino también proponer intervenciones efectivas para su prevención (Castro-Montoya et al., 2021; Erazo & Rosero, 2021).

1.4 Nuevas tendencias: análisis basados en datos

Con el auge de las tecnologías de la información, el análisis de datos ha emergido como una herramienta clave para anticipar comportamientos de abandono y generar sistemas de alerta temprana. Algoritmos de machine learning como árboles de decisión, regresión logística y k-nearest neighbor han mostrado altos niveles de precisión en la predicción de estudiantes en riesgo (Valero et al., 2022; Castro et al., 2023).

Estos modelos permiten no solo identificar a tiempo a los estudiantes con mayor probabilidad de desertar, sino también entender qué variables tienen mayor peso en su decisión. Esto facilita a las IES diseñar estrategias personalizadas de acompañamiento, tutoría, orientación vocacional y apoyo económico, maximizando así la eficiencia en el uso de recursos institucionales.

Un ejemplo del uso de estos modelos predictivos fue una investigación llevada a cabo por (Valero et al., 2022), dónde consiguieron entrenar un modelo basado en KNN con una precisión del 91% para pronóstico temprano de deserción universitaria.

```
model=KNeighborsClassifier(n_neighbors=3) #this examines 3 neighbours for puttin
model.fit(train_X,train_y)
prediction=model.predict(test_X)
ejemplo=model.predict([[1,0,0,1,0.5,10.0,0.4,2]])
print (ejemplo)
print('The accuracy of the KNN is',metrics.accuracy_score(prediction,test_y))
```

```
['retirado']
The accuracy of the KNN is 0.90625
```

Ilustración 2 - Ejemplo de Implementación de Modelo Predictivo basado en KNN Tomado de (Valero et al., 2022).

Una de las ventajas principales de estos enfoques es su capacidad de procesar grandes volúmenes de datos de forma automática, detectando relaciones que podrían pasar desapercibidas mediante análisis tradicionales. Por ejemplo, mediante árboles de decisión o modelos de redes neuronales, es posible identificar combinaciones de variables (como edad, modalidad de estudio, promedio acumulado y número de créditos inscritos) que se asocian consistentemente con riesgo de abandono.

También existen varios modelos estadísticos, como los estimadores Kaplan-Meier, que estiman una función de supervivencia, en este caso dónde la supervivencia es definida como la no deserción de los programas académicos, tal como lo muestra (Vélez, 2020) en la

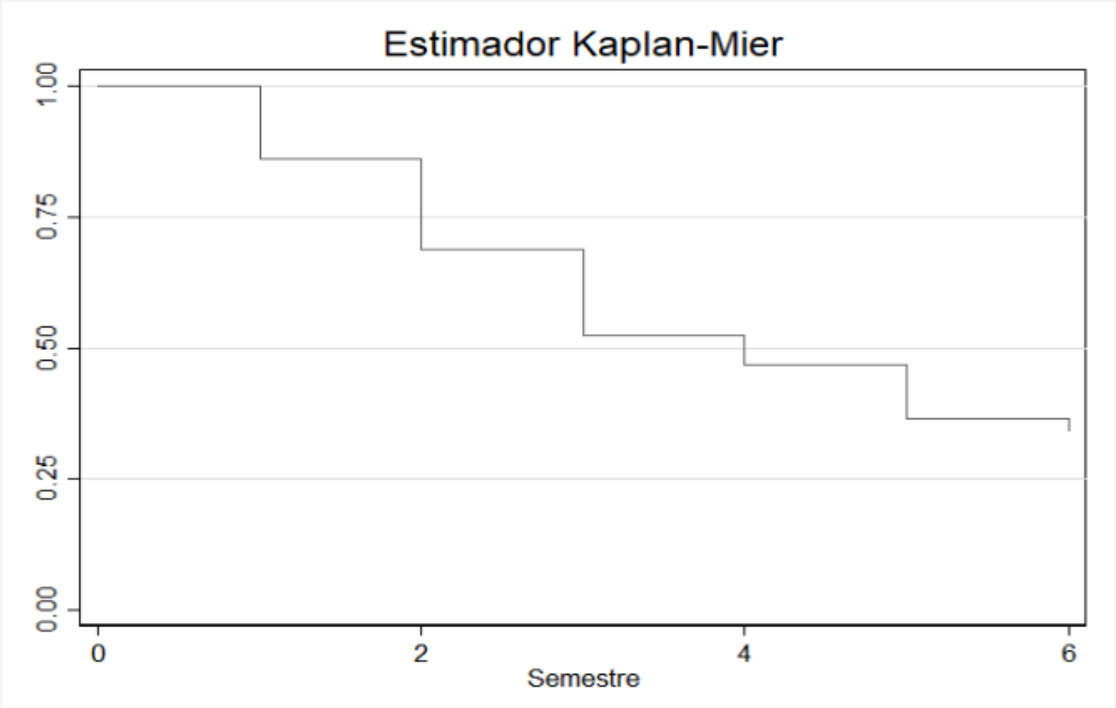

Ilustración 3 - Estimador Kaplan-Mier de la función de supervivencia de una cohorte de un programa tecnológico. Tomado de (Vélez, 2020)

Asimismo, enfoques como la minería de datos educativa (EDM) permiten extraer patrones ocultos en grandes volúmenes de datos históricos, mejorando la toma de decisiones institucionales basadas en evidencia (Castro et al., 2023). Estas metodologías no solo representan una innovación tecnológica, sino una oportunidad para avanzar hacia una educación superior más equitativa, eficaz y centrada en el estudiante.

En paralelo, el desarrollo de sistemas de información académica integrados y plataformas analíticas institucionales está favoreciendo la creación de entornos predictivos que funcionan en tiempo real. Estos sistemas pueden emitir alertas a coordinadores, docentes y áreas de bienestar cuando detectan indicadores de riesgo en la trayectoria de un estudiante, lo cual permite intervenciones proactivas en lugar de reactivas.

En este contexto, el uso de dashboards analíticos y tableros de control inteligentes permite a las universidades monitorear tendencias agregadas y tomar decisiones informadas respecto a sus políticas de retención, incluso segmentadas por programas, facultades, cohorte o perfil socioeconómico. Así, el análisis basado en datos se posiciona como una herramienta estratégica

indispensable para la gestión educativa contemporánea, brindando a las IES una ventaja competitiva en términos de eficiencia institucional y responsabilidad social.

1.5 Síntesis y planteamiento del problema

En síntesis, la deserción universitaria constituye un fenómeno complejo que implica múltiples dimensiones del entorno del estudiante. Desde las condiciones socioeconómicas hasta los aspectos emocionales, institucionales y pedagógicos, esta problemática debe ser abordada con una perspectiva sistémica y multidisciplinar. Si bien existen múltiples estudios y esfuerzos institucionales orientados a su mitigación, las tasas de abandono continúan siendo elevadas, lo cual pone en evidencia las limitaciones de las estrategias actuales.

Uno de los desafíos más importantes radica en la integración efectiva de los datos institucionales disponibles y en la capacidad para transformar esa información en conocimiento útil para la toma de decisiones. En muchas universidades, los procesos de admisión, caracterización, seguimiento y evaluación académica generan una gran cantidad de datos que, en su mayoría, no se analizan de forma sistemática para identificar patrones de riesgo de deserción. A esto se suma la falta de coordinación entre las áreas académicas, administrativas y de bienestar, lo que fragmenta las intervenciones y dificulta el abordaje integral del problema.

En este contexto, se hace indispensable que las instituciones de educación superior desarrollen marcos analíticos propios que les permitan comprender cómo se manifiesta la deserción dentro de su realidad específica. Esto implica considerar las particularidades de sus sistemas de admisión, el perfil de sus estudiantes, la modalidad de sus programas (presenciales, virtuales, híbridos), las rutas de acompañamiento disponibles y las condiciones del entorno socioeconómico en el que operan.

Así, el análisis de la deserción universitaria debe trascender los modelos generales y centrarse en la exploración detallada de los datos internos de cada institución. Solo a partir de esta comprensión contextualizada será posible diseñar estrategias de intervención pertinentes, que articulen acciones de prevención, seguimiento y recuperación en función de los factores que afectan a su propia comunidad estudiantil.

Planteamiento del problema:

¿Cuáles son los factores críticos que influyen en la deserción universitaria y cómo pueden ser identificados a tiempo mediante herramientas analíticas basadas en datos que permitan implementar intervenciones eficaces para la retención estudiantil en una institución específica?

1.6 Nuevos horizontes de análisis: hacia el aprendizaje automático

A medida que los desafíos asociados a la deserción universitaria se vuelven más sofisticados, también lo deben ser las herramientas para enfrentarlos. En este contexto, el machine learning y la inteligencia artificial representan una nueva frontera para la predicción y prevención del abandono. Estas tecnologías permiten modelar comportamientos complejos mediante algoritmos que aprenden de los datos, ajustan patrones y mejoran su rendimiento predictivo con el tiempo.

La implementación de modelos como redes neuronales, máquinas de soporte vectorial y bosques aleatorios abre posibilidades para personalizar la intervención educativa con base en el perfil individual del estudiante. Además, el aprendizaje automático permite una actualización constante de los modelos en función de nuevos datos, lo que mejora la precisión y relevancia de las predicciones.

El potencial transformador de estas herramientas radica en su capacidad para convertir grandes volúmenes de datos institucionales en conocimiento útil para la toma de decisiones. Esto posibilita a las IES desarrollar sistemas de gestión del riesgo de deserción más eficientes, que integren factores académicos, psicológicos, sociales y económicos, y que fomenten una cultura institucional de retención basada en evidencia.

Por tanto, es indispensable avanzar hacia un ecosistema educativo que reconozca el valor estratégico de los datos y adopte tecnologías de análisis avanzado como parte integral de sus procesos de gestión, planeación y acompañamiento estudiantil.

1.7 Nuestra propuesta

Se plantea la necesidad de desarrollar un modelo predictivo avanzado que utilice técnicas de análisis de datos modernas y comprehensivas, capaces de identificar patrones y señales de alerta

temprana. Este modelo buscará superar las limitaciones de las herramientas tradicionales y ofrecer a las instituciones educativas un medio más efectivo para implementar estrategias de prevención de la deserción.

1.7.1 Objetivos de la propuesta de trabajo de grado

Objetivo general:

Desarrollar un modelo predictivo de riesgos de deserción estudiantil en la institución universitaria Pascual Bravo, utilizando modelos de machine learning sobre datos locales de la institución.

Objetivos específicos:

- Identificar y recopilar datos históricos de los estudiantes, incluyendo variables académicas, socioeconómicas, demográficas y de comportamiento, que puedan influir en la deserción estudiantil.
- Utilizar técnicas de análisis de datos y aprendizaje automático para desarrollar un modelo predictivo que estime la probabilidad de deserción de cada estudiante, y validar su precisión mediante la comparación con datos históricos y la realización de pruebas de desempeño.
- Desplegar el modelo generado para ser consumido por los sistemas de información de la institución, utilizando la construcción de APIs.

1.7.2 Metodología Propuesta para el trabajo de grado

La metodología para este trabajo se propone tres etapas, como se indica a continuación:

Etapa 1: Identificación y Recopilación de Datos

Actividades:

- Caracterizar los datos de interés en la plataforma de datos universitarios.
- Colección de datos asociados a la deserción escolar.
- Informe de comportamientos de deserción escolar.
- Selección de variables de interés.
- Conversión de la colección de datos a un formato compatible con Python.
- Procesamiento de los datos en términos de limpieza y normalización.

- Etiquetado de los datos con respecto a los riesgos de deserción.

Productos:

- Datos históricos recopilados.
- Variables de interés identificadas y seleccionadas.
- Datos procesados y normalizados listos para el análisis.

Etapa 2: Desarrollo y Ajuste del Modelo Predictivo

Actividades:

- Implementación de una estrategia de balanceo de los datos.
- Identificación de usuarios con patrones similares a partir de modelos de aprendizaje no supervisado.
- Construcción de algoritmos predictivos a partir de modelos de aprendizaje supervisado.
- Evaluación del desempeño de los modelos mediante métricas de calidad.

Productos:

- Algoritmo predictivo entrenado y evaluado.
- Desempeño del modelo evaluado y validado con datos históricos.

Etapa 3: Despliegue y Validación del Modelo

Actividades:

- Generación de una API que tome como entrada los datos de un estudiante, utilice el modelo entrenado y devuelva el riesgo de deserción.
- Integración de la API a una interfaz gráfica de usuario.
- Validación del desempeño del modelo mediante la API y datos de usuarios nuevos.
- Generación de un sistema de alertas basado en los resultados del modelo.

Productos:

- Modelo desplegado y disponible para su uso.
- API funcional integrada con sistemas de información institucionales.

Capítulo 2:

Análisis exploratorio de los datos de deserción

2.1 Estructura de los datos de deserción.

La **Institución Universitaria Pascual Bravo (IUPB)**, en el marco de su proceso de admisión y seguimiento a los estudiantes admitidos, ha implementado diversas estrategias para la recolección sistemática de información relevante. Una de las principales herramientas empleadas para este propósito es el **formulario de inscripción**, el cual debe ser diligenciado por los aspirantes y contiene información relacionada con su procedencia geográfica, resultados en las pruebas Saber (ICFES), así como otros datos de carácter demográfico, socioeconómico y educativo.

Adicionalmente, una vez el aspirante es admitido y se formaliza su ingreso, la IUPB cuenta con una **plataforma de gestión académica e institucional**, en la que se consolida información clave sobre los estudiantes. Esta plataforma integra datos vinculados al proceso de matrícula, asignación de cursos, asistencia a clases, desempeño académico, historial de inscripción, entre otros aspectos que permiten hacer un seguimiento integral al ciclo de vida del estudiante dentro de la institución.

En el contexto de este trabajo, se tomaron como base dos fuentes de información:

A. **El formulario de inscripción**, diligenciado por los aspirantes.

B. **La plataforma institucional de gestión de estudiantes**.

A partir de estas fuentes, se seleccionó un subconjunto representativo compuesto por **40 variables**, las cuales han sido recopiladas principalmente durante el proceso de inscripción a un programa académico. Estas variables fueron clasificadas en categorías con el fin de facilitar su análisis y permitir una comprensión más clara de su naturaleza y utilidad. Las categorías

definidas corresponden a: **componentes personales, socioeconómicos, académicos, información del proceso de inscripción e información derivada o transformada**.

A continuación, se presenta una lista de las variables principales utilizadas en el análisis, acompañada de una breve descripción de cada una, agrupadas por tipo:

2.1.1 Información Personal

Estas variables describen características básicas del aspirante:

- **ciudad_de_residencia**: Ciudad en la que reside el aspirante al momento de la inscripción.
  - **direccion**: Dirección exacta del domicilio del aspirante.
  - **fecha_nacimiento / data_fecha_nacimiento**: Fecha de nacimiento del aspirante. Puede usarse para calcular edad.
  - **municipio**: Municipio (subdivisión territorial) correspondiente al lugar de residencia.
  - **etnia / data_etnia**: Pertenencia étnica del aspirante. Por ejemplo: indígena, afrodescendiente, blanco, mestizo, etc.
  - **genero / data_genero**: Identidad de género del aspirante (por lo general, masculino, femenino u otra identidad).
  - **desplazado**: Indica si el aspirante es víctima de desplazamiento forzado (sí/no), en el contexto del conflicto armado colombiano.

2.1.2 Información Académica

Relacionada con el proceso educativo del aspirante dentro de la institución:

- **programa / data_programa**: Programa académico al cual se postula (por ejemplo: Ingeniería Civil, Derecho, etc.).
  - **metodologia / data_metodologia**: Tipo de modalidad de estudio: presencial, virtual o a distancia.
  - **facultad / Facultad**: Facultad o unidad académica a la que pertenece el programa (puede haber duplicación de nombres con diferente capitalización).
  - **jornada / data_jornada**: Jornada de estudio: diurna, nocturna, fines de semana, etc.
  - **codigo_snies**: Código único asignado por el **Sistema Nacional de Información de la**

Educación Superior (SNIES) al programa académico.

- **sede / data_sede**: Sede o campus de la institución donde se ofrece el programa.
- **estado / data_estado**: Estado actual del aspirante o estudiante en el proceso (activo, retirado, admitido, cancelado, etc.).
- **codigo_de_estudiante**: Código interno o identificador único del estudiante en la base de datos institucional.
- **nivel**: Nivel académico del programa (por ejemplo, técnico, tecnológico, profesional, posgrado).
- **fecha_de_matricula**: Fecha en que se formalizó la matrícula.
- **fecha_iam**: Podría referirse a la fecha de inicio de actividades académicas o matrícula interna (IAM puede ser sigla institucional).
- **semestre**: Semestre académico en el que se encuentra el estudiante (por ejemplo, 1, 2, 3, etc.).
- **file_grupos**: Posiblemente una asignación de grupo académico, cohorte o clase del estudiante.
- **data_plan_de_estudio**: Hace referencia al plan curricular específico del programa al que se matricula el estudiante.
- **data_fecha_de_admision**: Fecha formal de admisión al programa académico.

2.1.3 Información del Proceso de Inscripción

Relacionada con el tipo de aspirante y su proceso de ingreso:

- **tipo_de_aspirante**: Categoría del aspirante, por ejemplo: nuevo, reingreso, transferencia, cambio interno, entre otros.
- **fecha_de_inscripcion.y**: Fecha en la que se registró formalmente la inscripción del aspirante (el sufijo ".y" sugiere que puede haber varias versiones de la misma variable en diferentes tablas fusionadas).

2.1.4 Información Socioeconómica

- **forma_de_pago / data_forma_de_pago / forma_de_pago_agrupada**: Método con el cual el estudiante financia sus estudios (puede incluir pago de contado, crédito ICETEX, financiación con entidad bancaria, subsidio gubernamental, entre otros). La versión "agrupada" puede ser una categorización más general.
  - **data_tipo_de_vivienda**: Tipo de vivienda en la que reside el aspirante, por ejemplo: propia, arrendada, familiar, compartida.

2.1.5 Variables Derivadas o de Control de Calidad

- **data_ciudad_de_residencia**: Versión normalizada o transformada de la ciudad de residencia.
  - **data_direccion**: Versión transformada de la dirección (posiblemente corregida o con formato estandarizado).

De la información recolectada en cada uno de los campos, la distribución de tipos de datos se vió mas influenciada por los textos, ellos debido a los altos componentes categóricos del instrumento

aplicado, así como resume la siguiente ilustración:

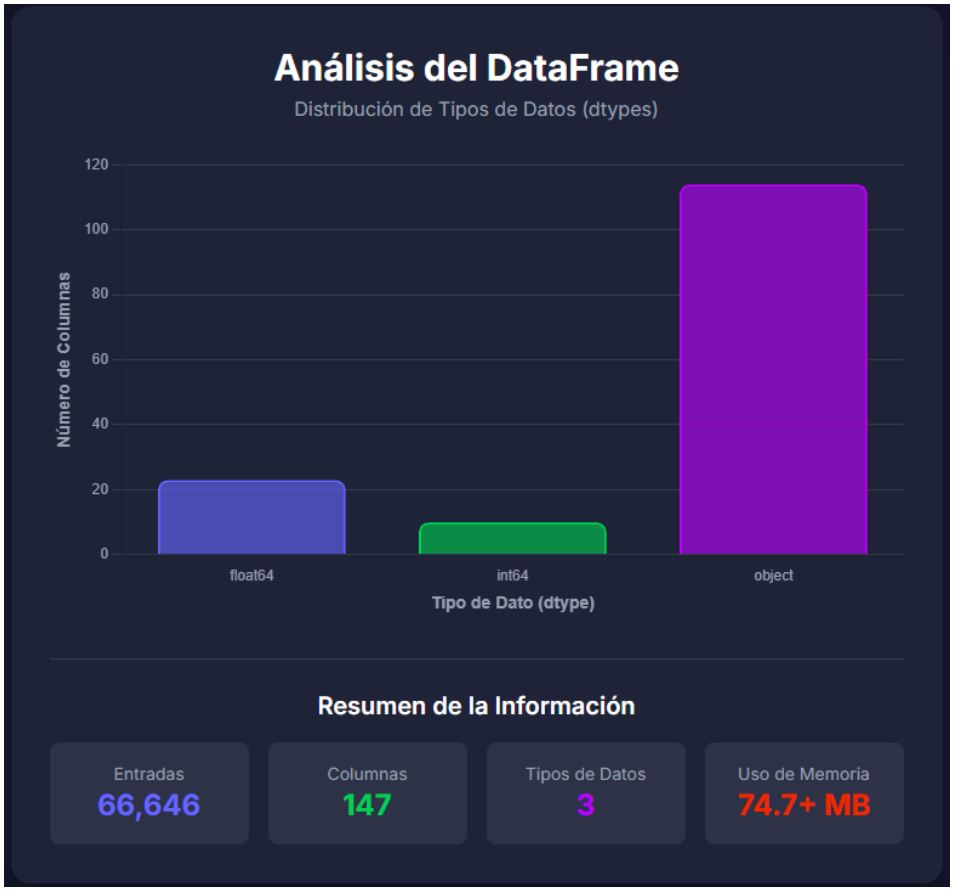

Ilustración 4-Distribución tipos de datos

2.2 Limpieza y preprocesamiento.

Para entrar en detalle sobre la etapa de procesamiento de los datos, es pertinente aclarar que la información utilizada en este trabajo corresponde a registros internos de carácter privado de la Institución Universitaria Politécnico de Bogotá (IUPB). En cumplimiento de las disposiciones legales sobre protección de datos personales y en virtud de un acuerdo de confidencialidad, se establece que los datos brutos no serán divulgados públicamente. Únicamente se presentarán resultados agregados en forma de gráficos estadísticos y análisis generales, sin comprometer la identidad de los individuos.

En este sentido, el trabajo parte de un proceso riguroso de anonimización de los datos, mediante el cual se eliminaron o transformaron todos los campos que pudieran permitir la identificación directa o indirecta de los aspirantes y estudiantes. Entre otras medidas, se excluyeron identificadores personales como nombres, documentos de identidad, códigos

estudiantiles y direcciones exactas; además, se aplicaron técnicas de generalización o codificación en variables sensibles como la fecha de nacimiento y la dirección de residencia.

Este tratamiento garantiza que los análisis posteriores se realicen sobre una base de datos protegida, alineada con los principios de seguridad, integridad y confidencialidad establecidos por la Ley 1581 de 2012 y demás normativas vigentes en Colombia sobre protección de datos personales. Así, se asegura el uso ético y responsable de la información para fines académicos y de investigación institucional.

2.2.1 Evaluación Inicial de la Calidad de los Datos

La fase de preprocesamiento constituyó una etapa fundamental para garantizar la integridad y confiabilidad de los datos utilizados en el análisis . La evaluación inicial reveló múltiples problemas de calidad que son característicos de los conjuntos de datos institucionales, incluyendo valores faltantes, duplicados, inconsistencias de formato y valores atípicos.

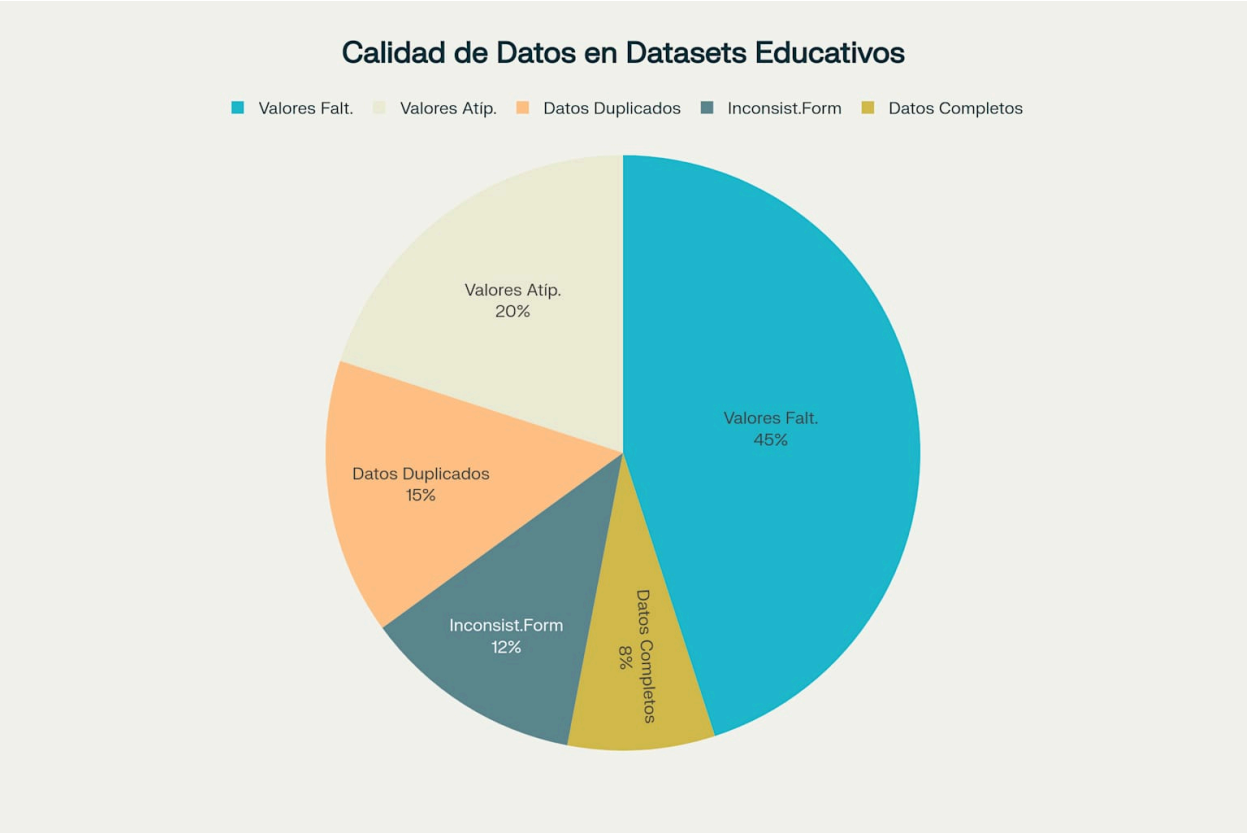

Ilustración 5. Distribución de problemas de calidad de datos encontrados en datasets educativos antes del preprocesamiento

Con respecto al conjunto de datos que se obtuvo, en la fase de Análisis Exploratorio de Datos (EDA) se desarrolló siguiendo la metodología KDD (Knowledge Database Discovery) o Descubrimiento de Conocimiento en Bases de datos con ciertas variaciones; La metodología KDD contempla las siguientes etapas: Entendimiento del Problema, Preparación de los datos, Minería de datos, Interpretación de los resultados y Evaluación y presentación del conocimiento. Se desarrollo un EDA para diferentes bases de datos institucionales, siendo la principal “Desercion.csv”, que consta de 66646 registros y 147 índices.

Posterior al cargue de los datos, se realizó una limpieza del dataset para eliminar columnas redundantes. En este caso, En el proceso de depuración del conjunto de datos, eliminamos una serie de columnas redundantes o innecesarias para facilitar el análisis y evitar duplicación de información. Según la imagen, las columnas eliminadas incluyen identificadores personales o sensibles como "id", "direccion", "codigo_snies", "codigo_de_programa", "fecha_iam", "file.x", "file.y", "eps", "colegio" y "codigo_de_estudiante", que podrían no aportar valor analítico o repetir datos en otras variables. También se descartaron campos redundantes bajo el comentario "Columnas duplicadas o redundantes", entre ellas: "data_fecha_nacimiento", "data_estado_civil", "data_rh", "data_ciudad_de_residencia", "data_comuna_de_residencia", "data_colegio", "data_direccion", "data_tipo_de_vivienda", "data_puntaje_sisben", "data_tipo_de_desplazamiento", "data_edad", "data_metodologia", "data_codigo_de_programa", "data_programa", "data_plan_de_estudio", "data_convenio", "data_jornada", "data_programa_segunda_opcion", "data_jornada_segunda_opcion", "data_barrio", "data_estrato", "data_etnia", "data_genero", "data_periodo", "data_sede", "data_estado", "data_forma_de_pago" y "data_discapacidades". Muchas de estas columnas duplicaban la información contenida en otras variables o correspondían a atributos poco relevantes para el objetivo del análisis, como datos sensibles o identificadores personales. Al eliminar estas columnas, se busca simplificar el dataset, reducir ruido, y enfocarse en las variables realmente significativas. De este proceso se obtuvo un dataset con 66646 registros convertidos, como resume la siguiente ilustracion.

Ilustración 6. registros intervenidos mediante la eliminación de información redundante.

Posteriormente, se realizó un proceso de normalización del dataset con el objetivo de unificar los valores ausentes o inconsistentes que pudieran dificultar el análisis. En particular, se identificaron diversas representaciones de datos faltantes tales como valores vacíos (""), la cadena "NA", paréntesis vacíos ("()") y valores nulos reconocidos por Pandas como NaN. Estas formas de datos faltantes, aunque distintas en su forma, representan una misma condición de ausencia de información, por lo que se decidió estandarizarlas e imputarlas bajo una categoría común llamada "Desconocido".

El propósito de esta transformación fue evitar errores en el procesamiento posterior, como fallos en la codificación de variables categóricas o distorsión en los análisis estadísticos, ya que la presencia de múltiples formatos de datos nulos podría llevar a una interpretación incorrecta de la información. Además, esta imputación facilita la trazabilidad y el análisis de los casos con información incompleta, sin necesidad de eliminar registros completos que podrían ser valiosos en otros aspectos.

Este proceso de limpieza y normalización fue acompañado por la generación de un reporte de imputación, que se presenta en la siguiente ilustracion, donde se muestra el impacto de esta transformación sobre el dataset, indicando la cantidad de datos que fueron modificados, las

columnas afectadas y la proporción de valores sustituidos por la etiqueta "Desconocido". Esto permite tener una visión clara de la calidad del dato y del alcance de la intervención sobre los valores ausentes. Para esto se observó que más del 90% de los registros requieren atención. En ellos, se identificaron las variables que requieren de intervención, y las que están mayormente gestionadas.

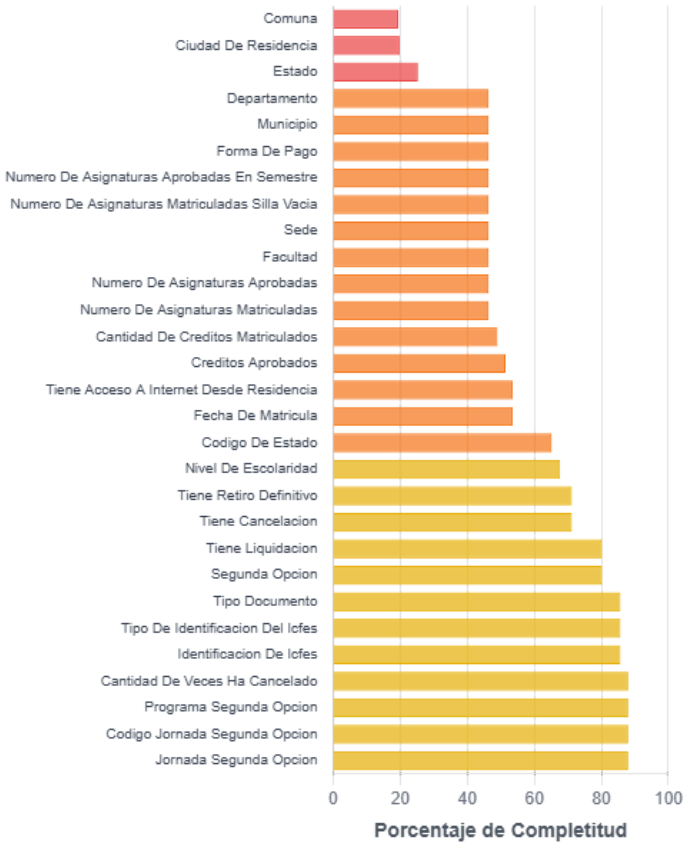

Ilustración 7. Variables con baja y alta Completitud de datos img1.

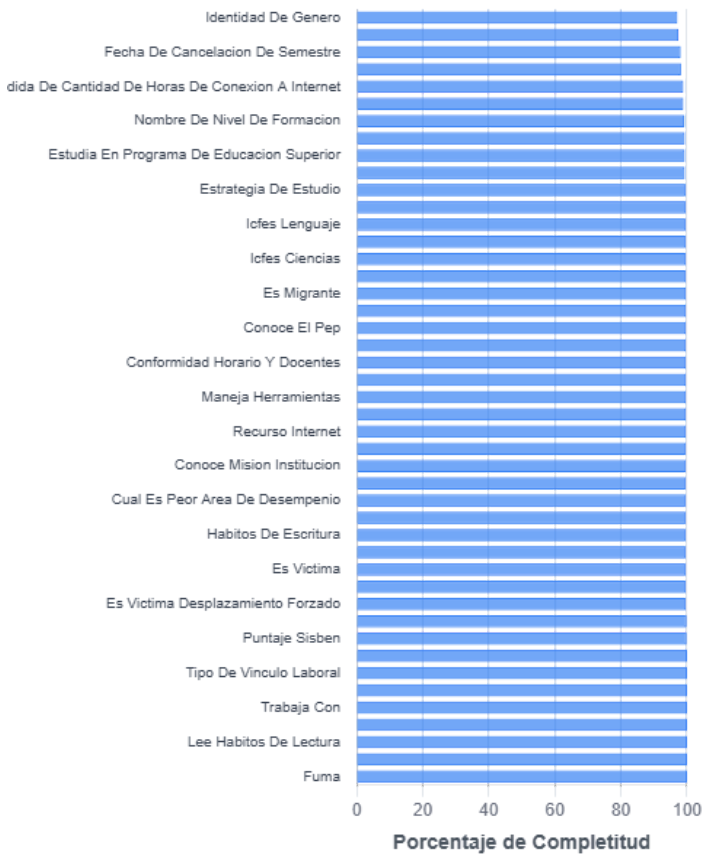

Ilustración 8. Variables con baja y alta Completitud de datos img2

Con lo anterior, y teniendo en cuenta las limitaciones observadas en los datos originales —tales como la presencia de valores faltantes, información duplicada o variables poco informativas—, se tomó la decisión de complementar el dataset con información adicional proveniente de fuentes institucionales internas previamente almacenadas. Esta estrategia permite enriquecer la base de datos original, mejorar la calidad de la información y aportar nuevas variables que puedan resultar relevantes para el análisis posterior, especialmente en estudios de permanencia, desempeño o caracterización estudiantil.

En esta fase, se partió de un análisis exploratorio preliminar sobre la data disponible, enfocado principalmente en una variable clave: si los estudiantes tienen o no un retiro definitivo del programa académico. Esta variable, denominada tiene_retiro_definitivo, resulta fundamental

para el objetivo del estudio, ya que permite establecer una primera segmentación de los casos y orientar el análisis hacia la detección de patrones asociados con la deserción o la continuidad en el sistema educativo.

El análisis superficial consistió en verificar la disponibilidad y consistencia de esta variable en la nueva fuente de datos, así como su relación con los registros del dataset original. A partir de esta evaluación, se diseñó una estrategia de integración para incorporar dicha variable al conjunto de datos ya depurado, asegurando su correcta correspondencia mediante claves compartidas como el identificador del estudiante o el código del programa. Esta integración marca el inicio de una fase de enriquecimiento de la información que permitirá profundizar en el estudio de trayectorias estudiantiles y mejorar la precisión de los modelos o análisis que se desarrollen a continuación. La siguiente tabla resume el resultado del proceso.

Tabla 1 Valores Si/No

SÍ	NO
56	17718

De la misma manera, se analizó la variable “Estado”, como resume la siguiente tabla:

Tabla 2. Estado Variables

MATRICULADO	21957
ADMITIDO	13101
RETIRO ACADÉMICO	11724
PENDIENTE DE PAGO	2427
CANCELACIÓN REGLAMENTARIA	1291
APROBADO	1206
NO ADMITIDO	510
EN REVISIÓN	389

EGRESADO	329
GRADUADO	193
RETIRO DEFINITIVO	123
INSCRITO	82
MATRICULA ESPECIAL	78
NO APROBADO	59

En la tabla podemos observar un problema de facto: el desbalance de clases, se tienen 56 registros de retiro definitivo, pero 17718 de no, lo que a simple vista podría indicar que se tiene una excelente retención de los estudiantes y una baja deserción, sin embargo, existen varias complicaciones con esto: Sí tomamos la variable “tiene_retiro_definitivo” como variable objetivo tenemos un problema de balance, o sí tomamos como aumentación de la data de la tabla estado y catalogamos como “Desertores” los que están presentes en las variables de Retiro Académico, Cancelación Reglamentaria y Retiro Definitivo, y como “No Desertores” los que existen dentro de las variables “Graduados”, “Egresados” y “Aprobado”, sería ahora de 13138 estudiantes Desertores y 1728 estudiantes No desertores. No se tomaron en cuenta los estudiantes “Matriculados” ni “Admitidos” ya que los matriculados no nos dan información relevante de sí el estudiante desertó en algún punto durante el semestre, y los estudiantes admitidos aún no se encuentran en calidad de estudiantes activos, y son susceptibles a no empezar su semestre académico por diversas razones.

El reto sigue presente, se tienen 13138 estudiantes Desertores vs 1728 estudiantes No Desertores.

Se realizó una segunda iteración de EDA sobre el dataset objetivo “Deserciones” pero con un enfoque ligeramente diferente que permitió identificar la estructura del dataset, la distribución de las variables y la presencia de valores faltantes y atípicos.

Se observó una cantidad considerable de columnas con tipos de datos heterogéneos, particularmente en variables como 'eps', 'colegio', 'nivel_de_escolaridad', 'sector_educativo', 'estrato', 'etnia', 'genero', 'estado' y 'forma_de_pago'. La imputación de valores faltantes se realizó utilizando la mediana para variables numéricas y la moda para las categóricas. Se identificaron y manejaron valores atípicos en las columnas numéricas, por ejemplo, en la variable 'edad', donde

se encontraron valores máximos de hasta 124 años antes de la winsorización, técnica estadística utilizada para limitar los valores extremos en un conjunto de datos, reemplazando los valores más pequeños o grandes con valores menos extremo, los cuales fueron tratados mediante la técnica de capping o control de frecuencias, limitando los valores extremos al percentil 99. Adicionalmente, se eliminaron 0 filas duplicadas detectadas en el proceso.

En el ámbito de la ingeniería de características, se crearon nuevas variables con el propósito de capturar posibles interacciones y relaciones no lineales que pudieran influir en la deserción. Se incluyó un término de interacción entre la edad y la cantidad de créditos matriculados ('age_x_credits'), y un término cuadrático para la edad ('edad_squared'). Estas transformaciones se basan en la premisa de que la combinación de factores demográficos y académicos puede tener un efecto diferenciado en la propensión a desertar.

Como un análisis complementario hacia el entendimiento de la información, la visualización de datos, a través de histogramas, diagramas de caja y mapas de calor, proporcionó una comprensión visual de las distribuciones de variables clave como la edad, la cantidad de créditos matriculados y, de manera exploratoria, las calificaciones (representadas por el "índice", aunque este requirió manejo específico debido a posibles inconsistencias en el nombre de la columna). Los diagramas de caja, por ejemplo, mostraron la distribución de la edad a través de diferentes programas académicos, mientras que los histogramas mostraron las distribuciones de las variables transformadas.

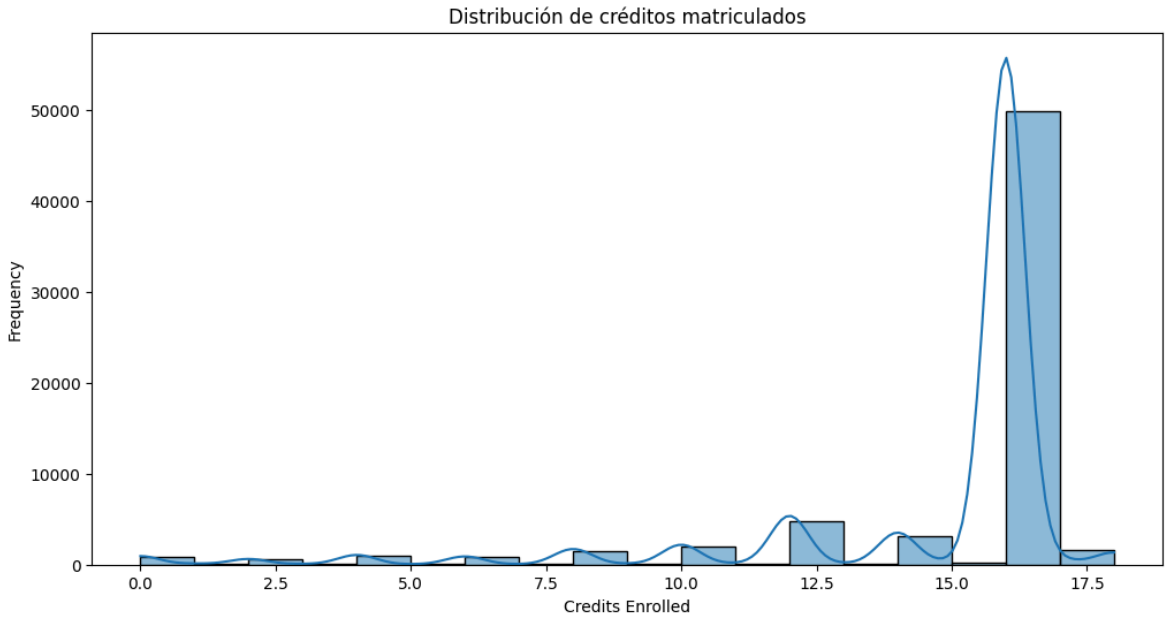

Ilustración 9. Análisis de la distribución de créditos matriculados.

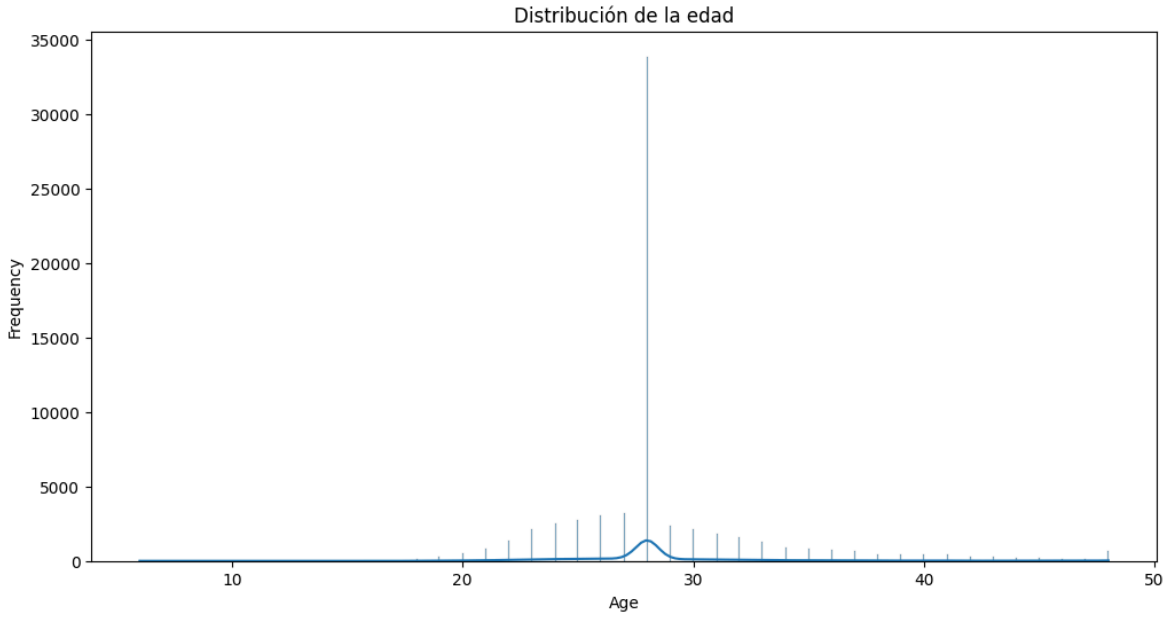

Ilustración 10. Análisis de la distribución de la edad.

A partir de los datos procesados, se realizó un análisis de relaciones entre las columnas numéricas presentes en el conjunto de datos. En consecuencia, una limitación crítica para el desarrollo de un modelo de regresión predictivo de deserción, a partir del dataset actual, radica en la ausencia de una variable objetivo explícita que indique si un estudiante ha desertado o no. Un modelo de regresión para la predicción de deserción (que comúnmente se aborda como un problema de clasificación, utilizando modelos como regresión logística, máquinas de soporte vectorial, o árboles de decisión) requiere una variable dependiente dicotómica (por ejemplo, 0 para no desertor y 1 para desertor) que sirva como la etiqueta de clase a predecir. Sin esta variable, no es posible entrenar un modelo supervisado que aprenda a discriminar entre estudiantes que perseveran y aquellos que abandonan sus estudios basándose en las características disponibles en el conjunto de datos.

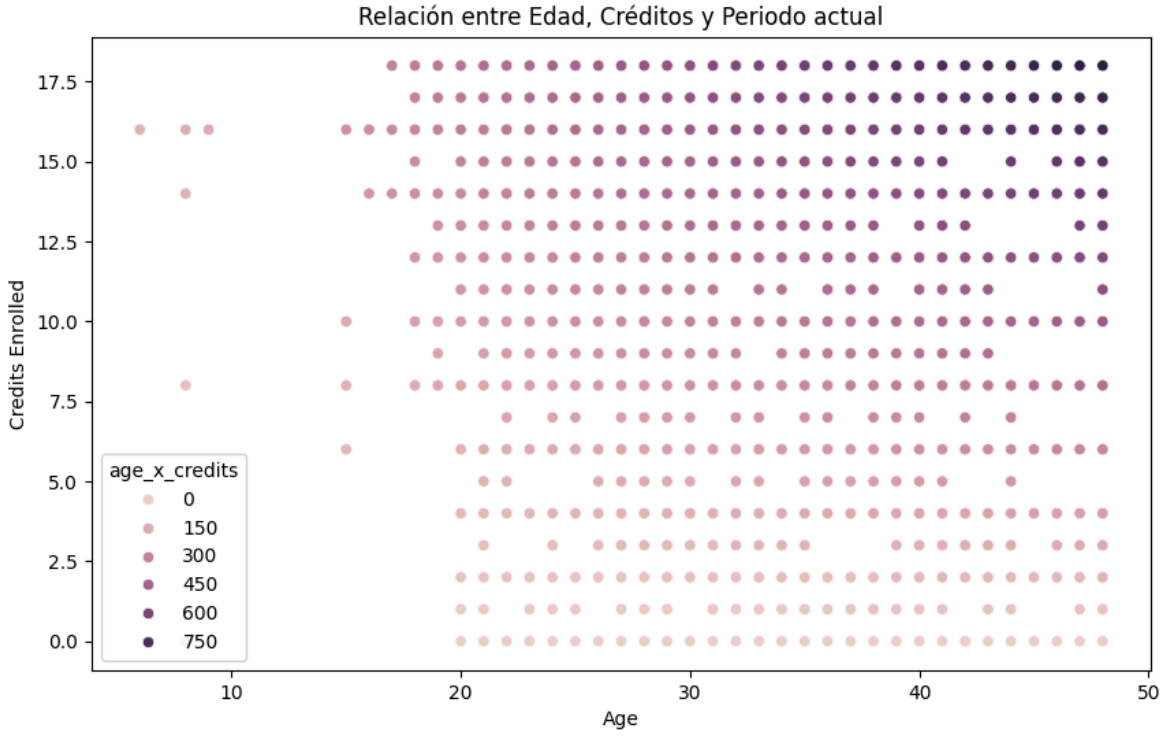

Ilustración 11. Relación de variables Edad,Creditos y Periodo

Aunque más adelante se presenta un análisis de correlaciones entre las variables, los análisis estadísticos comparativos, como el ANOVA realizado para examinar las diferencias en la edad entre distintos estratos socioeconómicos, revelaron un valor F de 70.2675 y un p-value de

1.7531e-59. Dado que el p-value fue menor que el umbral de significancia (comúnmente 0.05), se concluye que existe una diferencia estadísticamente significativa en las distribuciones de edad entre los diferentes estratos socioeconómicos. Esto sugiere que 'estrato' podría ser potencialmente un predictor de student dropout, pero no puede confirmar una relación causal o predictiva directa en ausencia de la variable de resultado.

Para este conjunto de datos que ha sido sometido a un proceso inicial de limpieza, exploración y transformación, revelando patrones y posibles relaciones entre variables que podrían ser predictivas de la deserción. Los datos mostraron un tamaño considerable de 99,493 entradas y 146 columnas, con la presencia de valores faltantes y atípicos que fueron abordados. Se observaron diferencias estadísticamente significativas en la distribución de edad entre los estratos socioeconómicos. Sin embargo, la imposibilidad de construir un modelo predictivo de deserción con el conjunto de datos actual se debe directamente a la falta de una variable que defina el evento de deserción. Para avanzar en la construcción de un modelo predictivo robusto, es indispensable la incorporación de una columna que identifique de manera precisa el estado de deserción de cada estudiante en el conjunto de datos o tomar un enfoque no supervisado para este problema.

Finalmente, se realizó una concatenación de dataset con la distinción de sí está activo o es desertor, esta concatenación se hizo con la eliminación de registros duplicados ya que algunos estudiantes se encontraban varias veces dentro de los datasets revisados inicialmente, las dimensiones del dataset resultante y con los labels de anotación fueron:

Tabla 3. Etiquetas

Label	Descripción	Cantidad
0	No Desertor	44651
1	Desertor	17140

El total de los datos en el dataset en conjunto es de 61791, se encontró que las dificultades siguen presentes para la generación de algún modelo de Machine Learning regresivo o de Deep Learning: Desbalance de Clases y exceso de datos desconocidos:

Ilustración 12. Etiquetas por clase.

2.2.2 Estrategias de Imputación de Valores Faltantes

Se implementaron múltiples técnicas de imputación de valores faltantes, adaptadas según el tipo de variable y la proporción de datos ausentes . Para variables numéricas con distribuciones aproximadamente normales, se aplicó la imputación por mediana, método que demostró mayor robustez ante la presencia de valores atípicos comparado con la imputación por media . En el caso de variables categóricas, se utilizó la imputación por moda, complementada con la creación de una categoría específica "Desconocido" para casos donde la ausencia de información podría ser informativa per se .

Las técnicas avanzadas de imputación, incluyendo métodos basados en K-Nearest Neighbors (KNN) y algoritmos de aprendizaje automático como Random Forest, mostraron mejores resultados en términos de preservación de la estructura de correlación original de los datos . Específicamente, la imputación por KNN demostró un rendimiento superior con valores de R^2 que superaron en 0.195 puntos a métodos tradicionales en datos con frecuencia horaria .

2.2.3 Detección y Tratamiento de Valores Atípicos

La identificación de valores atípicos se realizó mediante un enfoque multimétodo que combinó técnicas estadísticas univariadas y multivariadas . Se aplicó la técnica de winsorización

al percentil 99 para variables como la edad, donde se detectaron valores extremos de hasta 124 años que claramente representaban errores de captura o codificación . Esta técnica permitió preservar el tamaño de la muestra mientras se limitaba la influencia de observaciones extremas .

La identificación de valores atípicos se realizó mediante un enfoque multimétodo que combinó técnicas estadísticas univariadas y multivariadas . Se aplicó la técnica de Winsorización al percentil 99 para variables como la edad, donde se detectaron valores extremos de hasta 124 años que claramente representan errores de captura o codificación . Esta técnica permitió preservar el tamaño de la muestra mientras se limitaba la influencia de observaciones extremas.

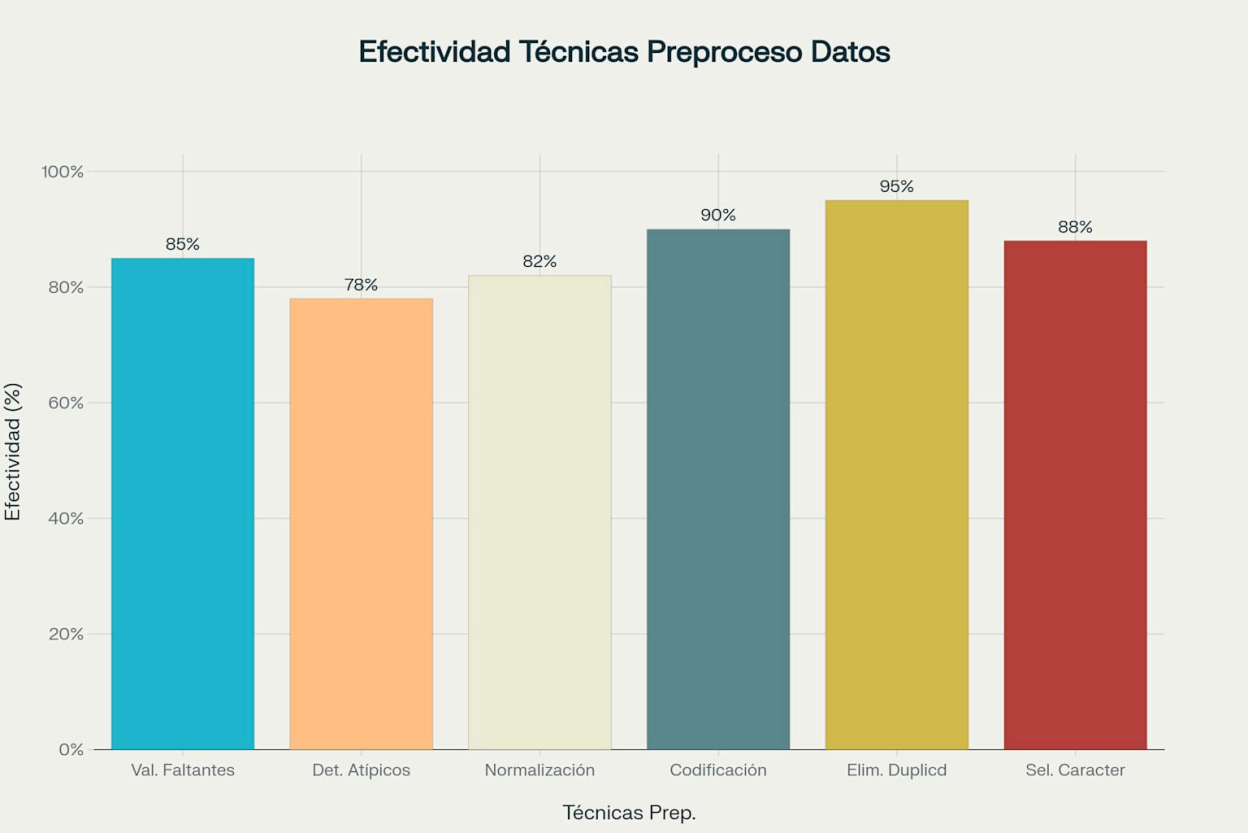

Ilustración 13.Efectividad de diferentes técnicas de preprocesamiento de datos en contextos educativos

Para la detección multivariada de valores atípicos, se implementaron algoritmos como Isolation Forest, Local Outlier Factor (LOF) y One-Class SVM, que demostraron eficacia superior a métodos tradicionales basados en distancia euclidiana . El método One-Class SVM, en

particular, mostró mejoras en el coeficiente de determinación (R^2) de hasta 0.215 puntos cuando se combinó con técnicas de imputación basadas en Random Forest.

2.3 Análisis de los Datos y Características Relevantes

2.3.1 Análisis de Correlación y Multicolinealidad

El análisis de correlación reveló patrones significativos entre las variables predictoras que son fundamentales para la comprensión del fenómeno de deserción estudiantil . Se identificaron correlaciones fuertes entre variables académicas como el número de asignaturas aprobadas y los créditos acumulados ($r > 0.78$), así como relaciones moderadas entre variables socioeconómicas y rendimiento académico .

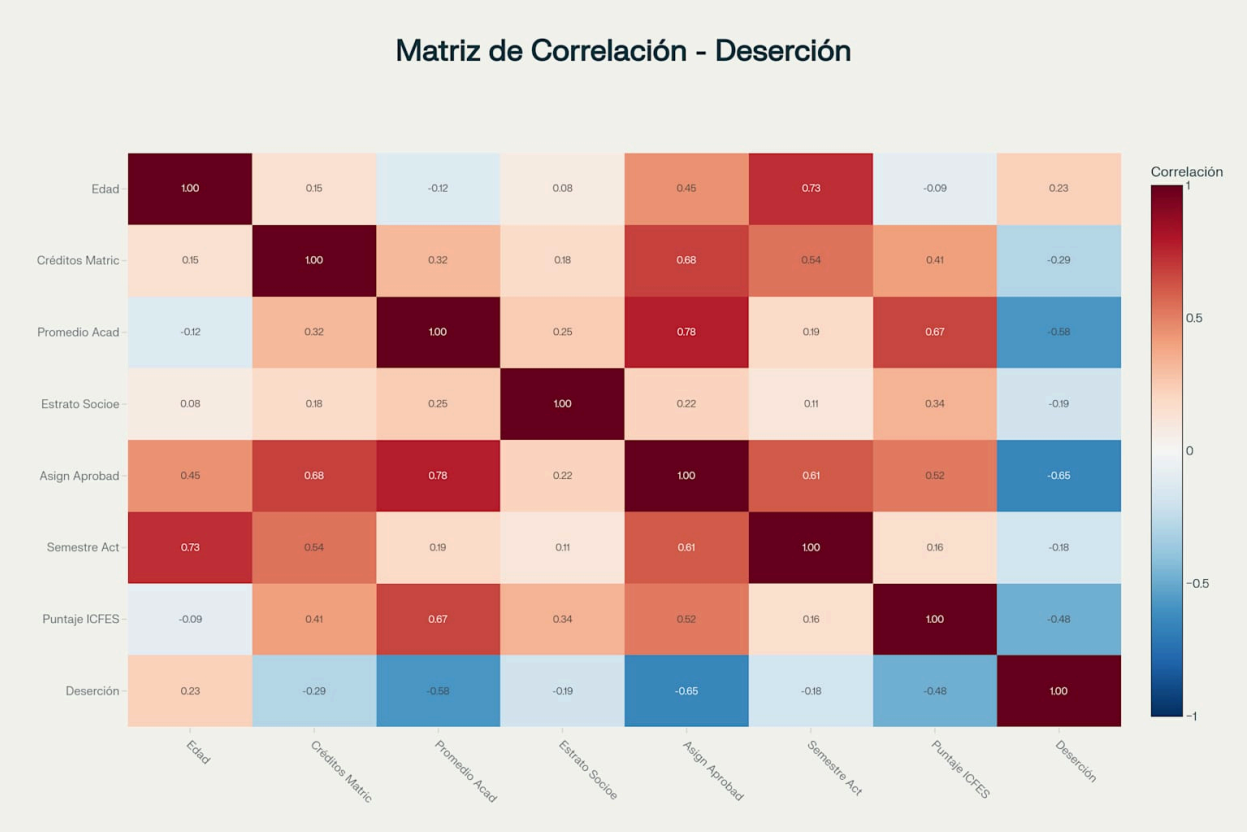

Ilustración 14. Matriz de correlación entre variables relevantes para el análisis de deserción estudiantil

La detección de multicolinealidad se realizó mediante el cálculo del Factor de Inflación de la Varianza (VIF), identificando variables con valores superiores a 10 que requerían atención especial . Variables como "Tar" y "Nicotina" en estudios comparativos mostraron VIF de 21.63 y 21.90 respectivamente, indicando redundancia informativa que podría comprometer la estabilidad del modelo .

Capítulo 3:

Implementación de modelos para estimar posibles casos de deserción.

Uno de los principales hallazgos del capítulo anterior se relaciona con la baja completitud de los datos recolectados. Este hallazgo evidencia que, dentro del proceso de recolección de información institucional, la Institución Universitaria Pascual Bravo (IUPB) enfrenta limitaciones significativas en cuanto al diligenciamiento adecuado y completo de los formularios por parte de los estudiantes. Esta situación sugiere la necesidad urgente de revisar y fortalecer los instrumentos de captura de información, así como las estrategias pedagógicas y comunicativas que acompañan el proceso de recolección, con el fin de garantizar una mayor calidad y fiabilidad de los datos consignados.

Además, se identificó que existe una desconexión entre la etapa de inscripción y los procesos posteriores de gestión académica, lo que dificulta la trazabilidad de ciertos campos de información a lo largo del ciclo de vida del estudiante. Esta fragmentación afecta directamente la utilidad de los datos para propósitos analíticos, particularmente en escenarios de análisis longitudinal y toma de decisiones basadas en evidencia.

En este contexto, se concluye que, debido al alto nivel de datos faltantes, la implementación directa de modelos predictivos de deserción estudiantil resulta limitada y poco viable, al menos desde un enfoque estrictamente empírico. Los vacíos en la información reducen la capacidad de

los algoritmos para identificar patrones consistentes y confiables que puedan anticipar situaciones de riesgo de abandono académico.

Es por ello que este capítulo adopta como punto de partida una estrategia de imputación de datos faltantes, con el objetivo de simular un escenario de mayor completitud. Este enfoque permite generar una base de datos robusta y continua, susceptible de ser modelada en términos predictivos. A través de este procedimiento, se busca aproximarse a las condiciones necesarias para desarrollar algoritmos de predicción de deserción, y así explorar, aunque sea de forma simulada, la viabilidad de este tipo de soluciones para apoyar la gestión institucional y la retención estudiantil.

3.1 Imputación y balanceo de datos.

Como parte de este trabajo, se implementó una estrategia de imputación de datos faltantes basada en el algoritmo de los “k” vecinos más cercanos (K-Nearest Neighbors, KNN). Esta técnica, ampliamente utilizada en ciencia de datos y aprendizaje automático, permite estimar los valores ausentes en una base de datos tomando como referencia las observaciones más similares dentro del conjunto disponible. Su elección responde a la necesidad de preservar la estructura local de los datos y de generar una simulación realista de un escenario con mayor completitud, que permita posteriormente desarrollar modelos predictivos, en este caso, orientados a la identificación temprana del riesgo de deserción estudiantil.

El algoritmo de KNN, en su versión tradicional, es un método de clasificación o regresión no paramétrico, basado en la idea de que las observaciones cercanas en el espacio de características tienden a tener valores similares. Cuando se adapta para imputación, este principio se utiliza para reemplazar un valor faltante en una observación mediante una combinación (promedio, moda, etc.) de los valores correspondientes en las observaciones más cercanas que sí tienen ese valor disponible.

A tener en cuenta:

Fundamento matemático

Dado un conjunto de datos incompletos $X = \{x_1, x_2, \ldots, x_n\}$, donde cada x_i es un vector en un espacio de dimensión d, se busca imputar un valor faltante en una dimensión específica x_{ij}.

El procedimiento consiste en:

1. **Definir una métrica de distancia** para comparar los vectores de datos. La más común es la **distancia euclidiana**:

$$d(x_i, x_k) = \sqrt{\sum_{l=1}^{d} (x_{il} - x_{kl})^2}$$

   Solo se consideran en la suma las dimensiones donde ambos vectores tienen valores presentes.

2. **Seleccionar el conjunto $N_k(x_i)$ de los k vecinos más cercanos** al vector incompleto x_i, utilizando únicamente las variables disponibles.

3. **Imputar el valor faltante x_{ij}** utilizando los valores conocidos de los vecinos en esa misma dimensión. La imputación puede realizarse mediante:

   - **Promedio** (para variables numéricas):

$$\hat{x}_{ij} = \frac{1}{k} \sum_{x \in N_k(x_i)} x_{xj}$$

   - **Moda** (para variables categóricas):

$$\hat{x}_{ij} = \text{moda}(\{x_{xj} : x \in N_k(x_i)\})$$

Ilustración 15. Fundamentación matemática para datos incompletos

Esto visto computacionalmente, se puede realizar usando bibliotecas de ciencia de datos en Python como scikit-learn (con KNNImputer), que permite definir el valor de *k*, elegir la estrategia de imputación (media, mediana, moda), y manejar automáticamente la normalización de los datos para garantizar que las diferentes escalas de las variables no afecten la distancia entre observaciones.

Antes de aplicar la imputación, es importante:

- Normalizar o estandarizar los datos, ya que KNN es sensible a la escala de las variables.
- Codificar variables categóricas si se desea aplicar la imputación numérica.
- Definir un valor adecuado de *k*, que puede explorarse empíricamente. Un valor común es entre 3 y 10, dependiendo del tamaño del conjunto de datos y la distribución de los valores.

El objetivo de aplicar la imputación por vecinos más cercanos (KNN Imputation) en este trabajo es simular un escenario de completitud de datos que sea estadísticamente coherente con los patrones observados en el conjunto original. Al imputar los datos faltantes de forma contextualizada, tomando como referencia individuos con

características similares, se busca reducir el sesgo que puede generarse al eliminar registros incompletos, lo cual comprometería la representatividad del análisis. Además, se procura conservar las relaciones locales entre las variables, lo que permite construir modelos predictivos más realistas y precisos, especialmente en el caso del análisis de deserción estudiantil. Finalmente, esta estrategia mejora la viabilidad de aplicar algoritmos de aprendizaje automático, los cuales requieren conjuntos de datos completos para entrenarse de manera robusta y efectiva.

En este trabajo, el ejercicio de imputación de datos permitió llegar a un dataset balanceado de 68.866, con 51 campos. Mitad para estudiantes en deserción y la otra mitad para estudiantes activos.

3.2 Entrenamiento de modelos de Machine learning para estimar la deserción.

Para abordar el problema de la deserción estudiantil desde una perspectiva predictiva, se entrenaron y evaluaron diversos modelos de machine learning utilizando una base de datos previamente tratada mediante técnicas de limpieza, normalización e imputación de datos faltantes. Todo el proceso se desarrolló en el entorno de Google Colab, lo que permitió aprovechar recursos computacionales en la nube —como aceleradores gráficos (GPU) y unidades tensoriales (TPU)— sin incurrir en costos adicionales o requerimientos locales complejos. Este entorno también facilitó la integración directa con bibliotecas científicas como scikit-learn, xgboost, pandas, numpy y matplotlib, optimizando el flujo de trabajo para análisis y visualización.

Ilustración 16. Librerías para el entrenamiento y evaluación de los modelos

Previo al entrenamiento, se llevó a cabo una partición aleatoria del conjunto de datos, dividiéndolo en un 80 % para entrenamiento y un 20 % para prueba, utilizando la función train_test_split() de la biblioteca scikit-learn. Esta división tuvo como finalidad asegurar una evaluación justa y realista de los modelos, permitiendo contrastar el rendimiento en datos no vistos durante el proceso de aprendizaje. Se entrenaron un total de siete modelos de clasificación, cada uno con fundamentos teóricos, ventajas y limitaciones distintas, con el propósito de identificar cuál de ellos ofrecía el mejor desempeño en la predicción de casos potenciales de deserción académica. A continuación, se presenta una descripción ampliada de cada uno:

Regresión Logística: Es un modelo estadístico ampliamente utilizado en problemas de clasificación binaria. Funciona modelando la probabilidad de que un evento ocurra (por ejemplo, que un estudiante deserte) mediante una función logística o sigmoide. Esta transforma una combinación lineal de las variables independientes en una probabilidad entre 0 y 1. Su fortaleza radica en su interpretabilidad: permite conocer qué variables

influyen positiva o negativamente en la probabilidad de deserción, y con qué peso. Es un modelo sencillo pero eficaz cuando las relaciones entre variables son aproximadamente lineales.

Máquinas de Vectores de Soporte (SVM): SVM es un modelo supervisado que busca encontrar el hiperplano óptimo que separa las clases de forma que la distancia (margen) entre los puntos más cercanos de cada clase al hiperplano sea máxima. Esto contribuye a una mejor generalización del modelo. Además, permite el uso de funciones kernel, que proyectan los datos a espacios de mayor dimensión para encontrar separaciones no lineales. Esto es especialmente útil en conjuntos de datos complejos con relaciones no evidentes entre las variables. Sin embargo, su rendimiento puede disminuir cuando el tamaño del conjunto de datos es muy grande.

XGBoost (Extreme Gradient Boosting): XGBoost es un algoritmo de boosting que construye un conjunto de modelos débiles (generalmente árboles de decisión) de forma secuencial, donde cada modelo nuevo intenta corregir los errores cometidos por los anteriores. Utiliza técnicas de optimización avanzadas como regularización L1 y L2, poda automática de árboles y manejo eficiente de valores faltantes. Es altamente valorado en problemas reales por su precisión, velocidad y capacidad de capturar interacciones complejas entre las variables. Requiere un ajuste cuidadoso de hiperparámetros, pero una vez calibrado, puede superar a modelos más simples en tareas de clasificación.

Random Forest (Bosque Aleatorio): Random Forest es un algoritmo de ensamble que construye múltiples árboles de decisión sobre diferentes subconjuntos aleatorios del conjunto de entrenamiento. Luego combina las predicciones de todos los árboles para tomar una decisión final (por mayoría en clasificación). Esta técnica reduce considerablemente el sobreajuste que puede ocurrir en modelos de árbol individuales y mejora la estabilidad del modelo. Es capaz de manejar variables numéricas y categóricas sin necesidad de mucha transformación previa. Además, proporciona medidas de importancia de las variables, lo cual es útil para la interpretación del modelo.

K-Vecinos Más Cercanos (K-Nearest Neighbors, KNN): Este modelo se basa en la idea de que los datos similares tienden a encontrarse cerca unos de otros en el espacio de características. Para clasificar un nuevo punto, KNN identifica los k vecinos más cercanos (según una métrica de distancia, como la Euclidiana) y asigna la clase más común entre ellos. Aunque es intuitivo y fácil de implementar, su rendimiento depende fuertemente de la elección del valor de k, de la escala de los datos (por lo que requiere normalización) y puede ser computacionalmente costoso para grandes volúmenes de datos.

Naive Bayes: Este clasificador probabilístico está basado en el teorema de Bayes y la asunción (ingenua) de independencia condicional entre las características. A pesar de esta simplificación, suele funcionar sorprendentemente bien en muchos contextos, especialmente con datos categóricos o texto. Es rápido, eficiente y no requiere muchos datos para entrenarse. En contextos como la deserción académica, puede ser útil cuando se cuenta con pocas observaciones por clase o cuando se requiere un modelo base ligero y rápido para comparaciones.

Perceptrón Multicapa (Multilayer Perceptron, MLP): El MLP es un tipo de red neuronal artificial compuesta por una capa de entrada, una o más capas ocultas y una capa de salida. Cada neurona realiza una combinación ponderada de las entradas y aplica una función de activación no lineal (como ReLU o sigmoide), permitiendo al modelo aprender representaciones complejas y no lineales. Aunque su entrenamiento puede ser más costoso y requiere una selección cuidadosa de hiperparámetros (tasa de aprendizaje, número de capas, cantidad de neuronas, etc.), tiene una alta capacidad de modelado y suele lograr buenos resultados en problemas con relaciones de alta complejidad entre variables.

Cada uno de estos modelos fue entrenado utilizando técnicas de validación cruzada y ajuste de hiperparámetros mediante búsqueda en malla (GridSearchCV), con el objetivo de maximizar su rendimiento general. Como fue mencionado, el proceso de entrenamiento de los modelos se realizó completamente en el entorno de Google Colab,

lo que permitió ejecutar algoritmos computacionalmente exigentes sin comprometer recursos locales. Sin embargo, los tiempos de entrenamiento variaron significativamente entre los modelos, dependiendo de su complejidad, del número de hiperparámetros ajustados y de la cantidad de operaciones iterativas involucradas. Modelos simples como Naive Bayes y Regresión Logística se entrenaron en cuestión de segundos, mientras que algoritmos más sofisticados como XGBoost, el Perceptrón Multicapa y Random Forest, especialmente al utilizar búsqueda en malla (GridSearchCV) para la optimización, requirieron varios minutos.

La siguiente ilustración presenta una estimación comparativa de los tiempos de entrenamiento para cada modelo, medidos en segundos o minutos, bajo condiciones típicas de procesamiento en Colab:

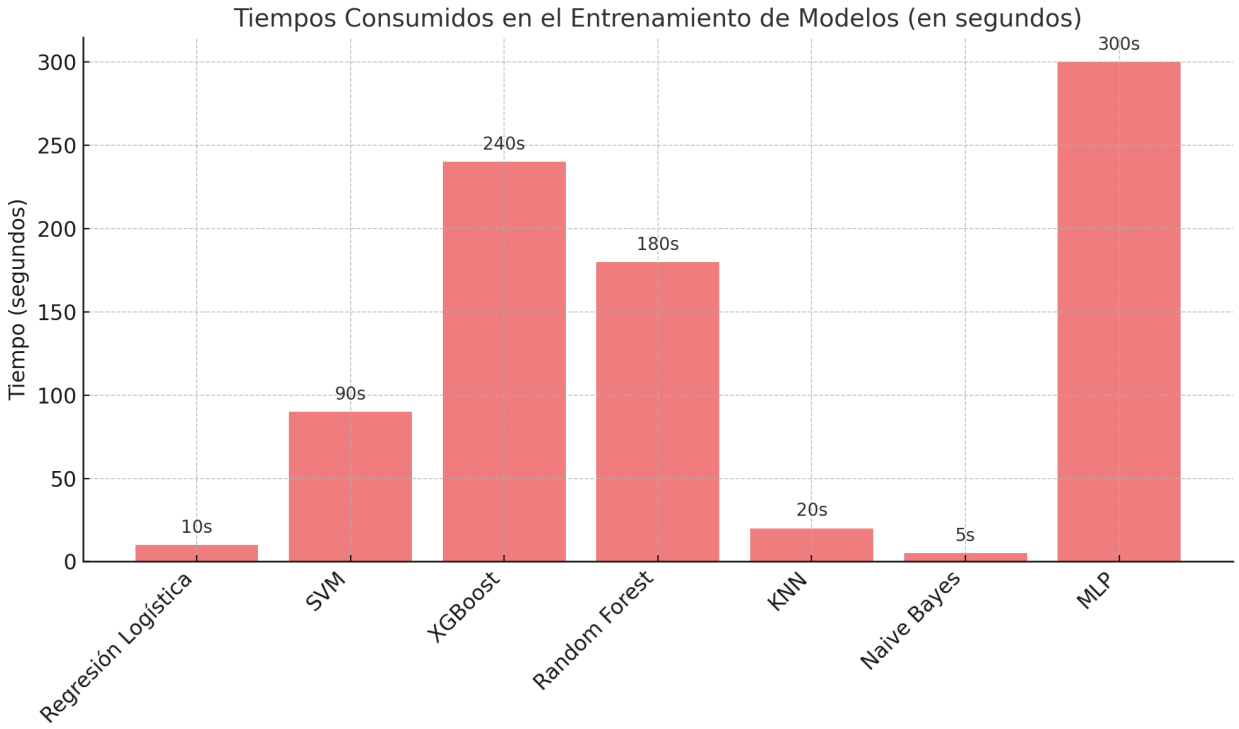

Ilustración 17. Tiempos de cómputo invertidos en la etapa de entrenamiento por modelo.

3.3 Evaluación y análisis de desempeño de los modelos de deserción.

Para evaluar la capacidad predictiva de los siete modelos entrenados en la tarea de clasificación de deserción estudiantil, se utilizó un conjunto de datos reservado para prueba, correspondiente al 20% del total de los datos disponibles. Esta división permitió medir el desempeño de cada modelo en datos no vistos durante el entrenamiento, lo cual es fundamental para asegurar su capacidad de generalización.

El desempeño se midió comparando las etiquetas reales del conjunto de prueba con las predicciones generadas por cada modelo. Para esto, se empleó el reporte de desempeño (classification_report) proporcionado por la biblioteca scikit-learn (sklearn), que incluye las principales métricas utilizadas en problemas de clasificación: precisión, recall, F1-score y soporte; como se describe a continuación:

Precisión (Precision): Esta métrica indica la proporción de instancias que el modelo clasificó como positivas (en este caso, estudiantes que desertaron) que efectivamente corresponden a verdaderos positivos. En términos prácticos, una alta precisión significa que cuando el modelo predice que un estudiante desertará, la probabilidad de que esa predicción sea correcta es elevada, reduciendo el número de falsas alarmas (falsos positivos). Matemáticamente se define como:

$$\text{Precisión} = \frac{TP}{TP + FP}$$

Ilustración 18. Precisión

Donde *TP* es el número de verdaderos positivos y *FP* el número de falsos positivos.

Recall (Sensibilidad o Exhaustividad): Esta métrica mide la capacidad del modelo para detectar todos los casos positivos reales. Es decir, de todos los estudiantes que efectivamente desertaron, qué proporción fue correctamente identificada por el modelo. Un alto recall es

especialmente importante en contextos donde es crítico no dejar pasar casos verdaderos sin detectar. Su fórmula es:

$$\text{Recall} = \frac{TP}{TP + FN}$$

Ilustración 19. Recall

Donde *FN* representa los falsos negativos (casos positivos no detectados).

F1-Score: El F1-score es la media armónica entre precisión y recall. Esta métrica es útil cuando se requiere un balance entre ambas, especialmente en situaciones donde hay un compromiso entre evitar falsos positivos y no dejar pasar falsos negativos. El F1-score se calcula como:

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

Ilustración 20. F1

En este sentido, los reportes para cada uno de los modelos entrenados indica que los modelos con mayor rendimiento (XGBoost, Random Forest y MLP).

- XGBoost (Extreme Gradient Boosting) y Random Forest presentan métricas superiores al 99%, incluyendo una roc_auc prácticamente perfecta. Esto sugiere que ambos modelos tienen una gran capacidad para capturar patrones complejos y no lineales en los datos.

- MLP (Perceptrón Multicapa) también alcanza un rendimiento muy alto, especialmente en recall y f1_score, lo cual es relevante cuando se busca minimizar los falsos negativos (es decir, estudiantes que desertan y no fueron identificados por el modelo).

El alto desempeño de estos modelos puede atribuirse a su capacidad de manejar relaciones no lineales, interacciones entre variables y, en el caso de XGBoost, a su habilidad para penalizar errores iterativamente. De manera contraria, los modelos con desempeño intermedio fueron SVM y KNN.

- SVM presenta métricas muy altas, aunque sin valor reportado para roc_auc, lo cual limita una evaluación completa. Su comportamiento sugiere que los datos son en su mayoría separables, aunque puede ser sensible a outliers o a clases desbalanceadas si no se ajustan correctamente los parámetros.

- KNN, aunque por debajo de los anteriores, aún conserva una precisión superior al 99% y un f1_score cercano al 99%, lo que indica un buen balance entre precisión y cobertura. Su desempeño depende fuertemente de la estructura de los datos y la distancia entre ejemplos; podría verse afectado por la presencia de ruido o por la alta dimensionalidad.

La siguiente ilustracion, resume los reportes de desempeño obtenidos en el proceso de testeo de los modelos entrenados.

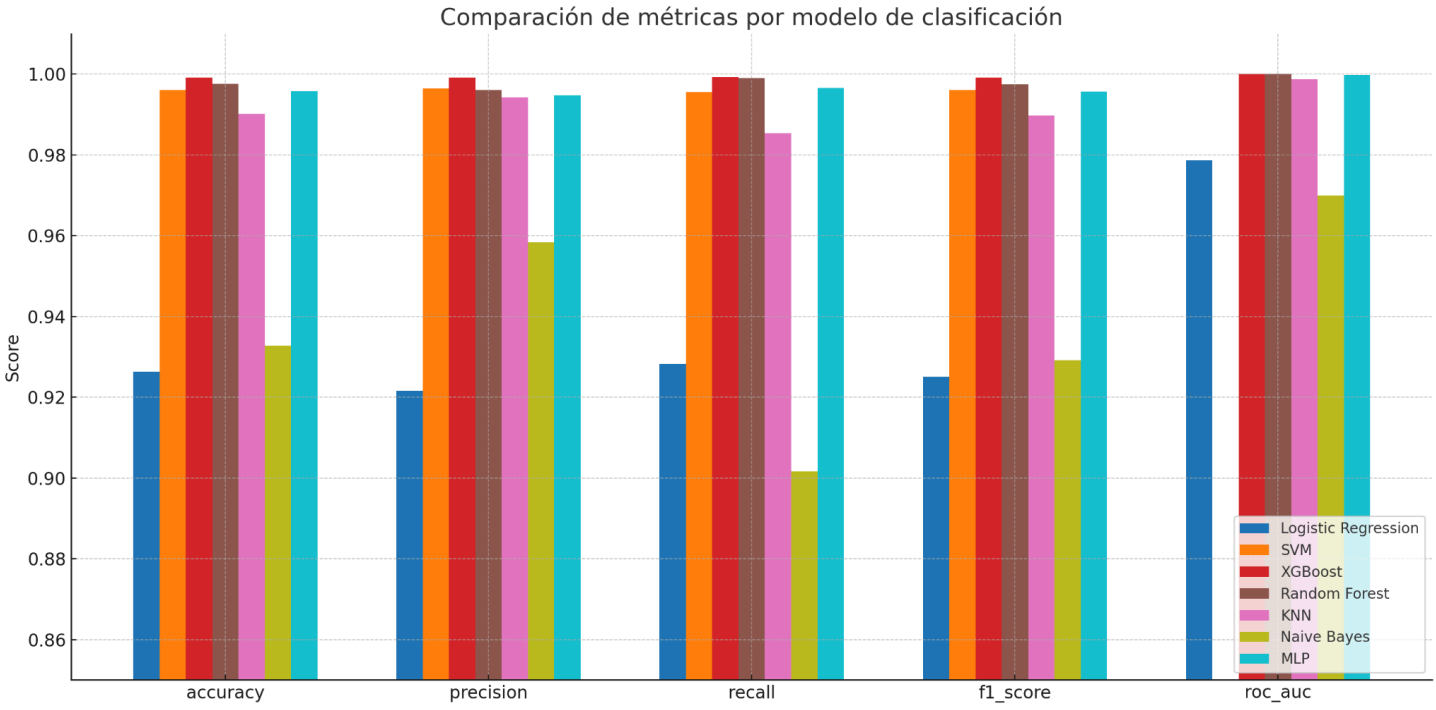

Ilustración 21. Comparativo de desempeño entre los modelos entrenados.

En concordancia con el gráfico, si bien los modelos Random Forest y XGBoost obtuvieron los mejores desempeños individuales en términos de métricas como precisión, recall y F1-score, se optó por utilizar una estrategia de votación de clasificadores (Voting Classifier) como enfoque final para la toma de decisiones. Esta técnica pertenece al conjunto de métodos de ensamble, los cuales combinan múltiples modelos para generar una predicción más robusta y estable. En particular, el voto de clasificadores consiste en combinar las predicciones de varios modelos base (por ejemplo, Random Forest, XGBoost, y otros modelos entrenados) y emitir una decisión final basada en la mayoría de votos ("votación dura") o en el promedio de probabilidades ("votación blanda").

Este enfoque permite aprovechar las fortalezas individuales de cada modelo y mitigar sus debilidades, lo que puede resultar en un mejor desempeño global, especialmente en contextos donde la clasificación correcta de casos positivos es crítica, como en la predicción de la deserción estudiantil. Además, el uso de múltiples modelos reduce el riesgo de sobreajuste que podría presentar un único modelo entrenado.

A continuación se presenta un gráfico de barras comparando el desempeño de cada modelo individual junto con el modelo combinado por votación. En este ejemplo, se observa que el modelo Ensemble alcanza métricas muy altas, comparables e incluso superiores en algunos casos a modelos individuales como Random Forest o MLP. Se destaca especialmente en precision (0.9987) y roc_auc (0.99995), lo que sugiere que logra un excelente equilibrio entre identificar correctamente los casos de deserción sin generar muchos falsos positivos. Esto valida que la combinación de varios modelos puede aprovechar las fortalezas de cada uno, mejorando la estabilidad y precisión general del sistema predictivo.

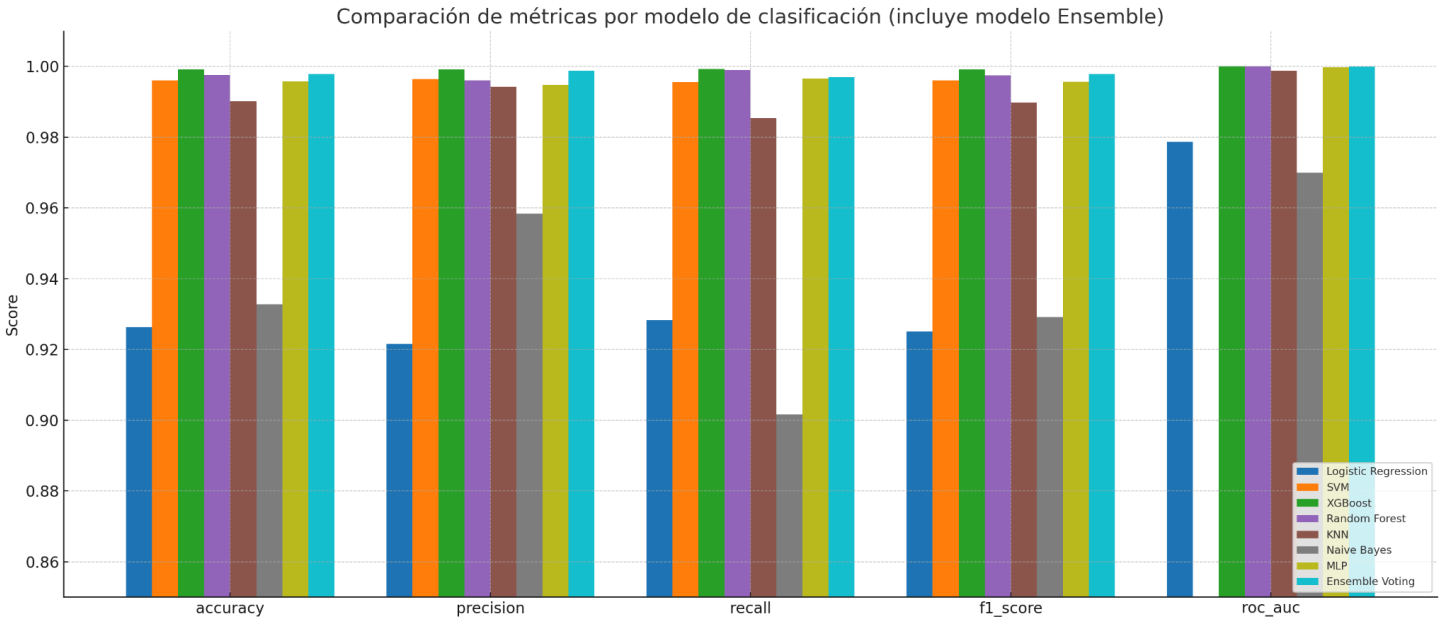

Ilustración 22. Ejemplificación de predicciones por votación de clasificadores.

Capítulo 4: Estrategia de despliegue del modelo.

4.1 Selección del formato de almacenamiento del modelo entrenado.

Una vez finalizado el proceso de entrenamiento, evaluación y validación de los modelos de clasificación, es fundamental seleccionar un formato adecuado para el almacenamiento de los modelos resultantes. Esta decisión impacta directamente en la facilidad de integración, reutilización, portabilidad y eficiencia del modelo en entornos de producción o despliegue.

Existen diversos formatos utilizados comúnmente en el ecosistema de ciencia de datos y aprendizaje automático, como por ejemplo:

- PKL (.pkl): Formato de serialización de objetos de Python mediante la librería pickle. Permite almacenar objetos complejos (como modelos de scikit-learn, listas, diccionarios, etc.) de forma directa y con bajo esfuerzo de configuración. Es ampliamente utilizado en entornos de desarrollo en Python.
    -
- Joblib (.joblib): Alternativa a pickle, optimizada para almacenar objetos de gran tamaño, como matrices numpy. También es comúnmente usada para modelos de scikit-learn.
    -
- ONNX (.onnx): Formato de modelo abierto y multiplataforma que permite portar modelos entre diferentes frameworks (PyTorch, TensorFlow, scikit-learn, etc.). Es útil cuando se desea desplegar en entornos heterogéneos.
    -
- PMML (.pmml): Formato estándar basado en XML que facilita la interoperabilidad entre herramientas analíticas. Requiere configuración adicional para ciertos modelos.
    -
- HDF5 (.h5): Usado especialmente con modelos de redes neuronales en frameworks como Keras/TensorFlow. Permite almacenar tanto la arquitectura como los pesos y configuración del modelo.

No obstante, para este proyecto, se decidió utilizar el formato .pkl (pickle) para almacenar los modelos entrenados por varias razones prácticas. Ello, partiendo de:

- Compatibilidad directa con scikit-learn y Python, que fueron los entornos utilizados durante el desarrollo del sistema de clasificación. Esto simplifica la serialización y deserialización de los modelos sin necesidad de pasos intermedios.
- Facilidad de implementación en ambientes controlados, como notebooks, scripts automatizados o microservicios construidos con Python (por ejemplo, usando Flask o FastAPI), donde el archivo .pkl puede ser cargado fácilmente en memoria y utilizado para realizar predicciones en tiempo real.
- Flexibilidad y bajo costo computacional, ya que el formato pickle permite almacenar no solo el modelo, sino también estructuras auxiliares como transformadores, pipelines y configuraciones preentrenadas en un solo archivo.
- Reutilización en fases posteriores del proyecto, como el análisis de desempeño en producción o la validación cruzada con nuevos datos institucionales, sin necesidad de reentrenar el modelo desde cero.

Aunque formatos como ONNX o PMML podrían ofrecer mayor portabilidad en contextos multiplataforma o sistemas más complejos, el entorno de despliegue previsto en este caso está completamente basado en Python, lo que hace que pkl sea una solución más directa y eficiente.

4.2 Estrategia de despliegue tipo API.

Una vez entrenado y validado el modelo predictivo, es necesario definir cómo será utilizado en la práctica por usuarios o sistemas externos. Para ello, se plantea una estrategia de despliegue tipo API (Application Programming Interface), que consiste en exponer el modelo como un servicio accesible a través de internet o de una red interna, mediante solicitudes estructuradas (por ejemplo, en formato JSON) enviadas por aplicaciones cliente.

Este enfoque responde a la necesidad de automatizar la consulta y predicción en tiempo real, facilitando la integración del modelo con otros sistemas institucionales (como plataformas académicas, portales de seguimiento estudiantil o sistemas de alerta temprana). A diferencia de

soluciones locales o interfaces manuales, una API permite que el modelo esté disponible de forma continua, centralizada y accesible desde múltiples puntos sin necesidad de replicarlo o exponer su lógica interna.

De este modo, el modelo pasa de ser una herramienta de análisis aislada a convertirse en un componente activo dentro de una arquitectura funcional orientada a servicios, lo que amplía significativamente su impacto y utilidad en procesos institucionales de toma de decisiones.

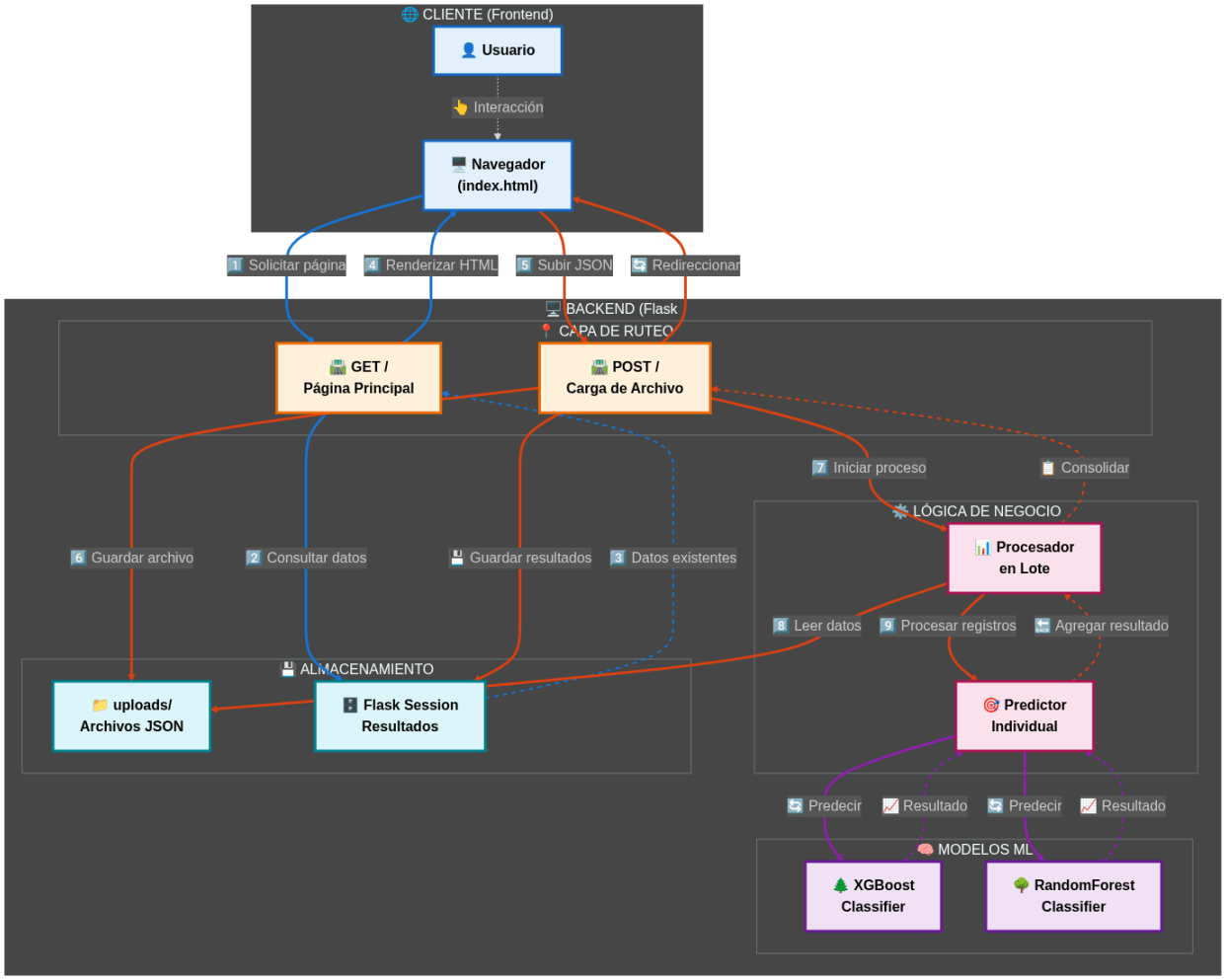

Ilustración 23. Arquitectura aplicación

La arquitectura del sistema implementado sigue un patrón de diseño multicapa basado en el framework Flask para el desarrollo de aplicaciones web con capacidades de machine learning. El sistema se estructura en cuatro componentes principales claramente diferenciados: la capa cliente (frontend), la capa de ruteo, la lógica de negocio y el almacenamiento de datos.

En la capa cliente, el usuario interactúa con la aplicación a través de un navegador web que renderiza la interfaz HTML (index.html), estableciendo la comunicación con el servidor mediante protocolos HTTP estándar. La capa de ruteo, implementada en Flask, gestiona dos endpoints principales: una ruta GET (/) que maneja la carga inicial de la página y consulta las predicciones existentes almacenadas en la sesión del usuario, y una ruta POST (/) que procesa la carga de archivos JSON conteniendo los datos a predecir.
La lógica de negocio se divide en dos componentes especializados: el Procesador en Lote, responsable de coordinar el procesamiento masivo de registros desde el archivo JSON cargado, y el Predictor Individual, que ejecuta las predicciones unitarias utilizando los modelos de machine learning previamente entrenados. Este último componente interactúa directamente con dos algoritmos de clasificación: XGBoost Classifier y RandomForest Classifier, aprovechando las fortalezas complementarias de ambos enfoques algorítmicos para generar predicciones robustas.

El sistema de almacenamiento opera en dos niveles: el directorio uploads/ que gestiona la persistencia temporal de los archivos JSON cargados por los usuarios, y Flask Session que mantiene el estado de las predicciones generadas durante la sesión activa del usuario. Este diseño arquitectónico garantiza la separación de responsabilidades, facilita el mantenimiento del código y proporciona escalabilidad para el procesamiento de grandes volúmenes de datos mediante técnicas de machine learning supervisado.

4.3 Despliegue de un front-end de prueba.

Como parte del proceso de validación funcional y experiencia de usuario, se implementó un front-end de prueba que permite interactuar con el modelo predictivo de manera visual y accesible. Esta interfaz gráfica simula el entorno de uso por parte de personal institucional o analistas académicos, permitiendo ingresar datos de estudiantes, enviar la solicitud al modelo a través de la API previamente desplegada y visualizar la predicción de forma clara e inmediata.

El objetivo de este componente es evaluar la usabilidad y operatividad del sistema de predicción en condiciones similares a las de un entorno real, facilitando la identificación de ajustes necesarios tanto en la entrada de datos como en la presentación de los resultados. Asimismo,

sirve como demostrador funcional para validar la integración entre el modelo, la API y un entorno de usuario.

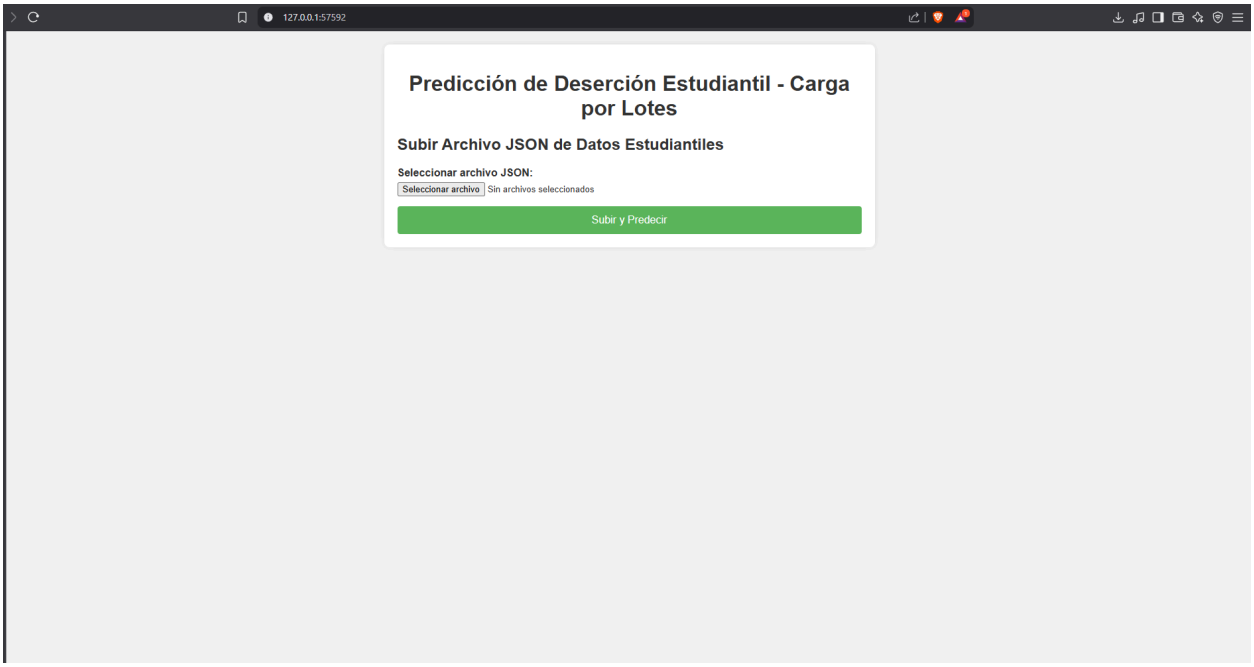

Ilustración 24. Despliegue Front-End Aplicacion

Ilustración 25. Cargue de archivo formato JSON para prediccion

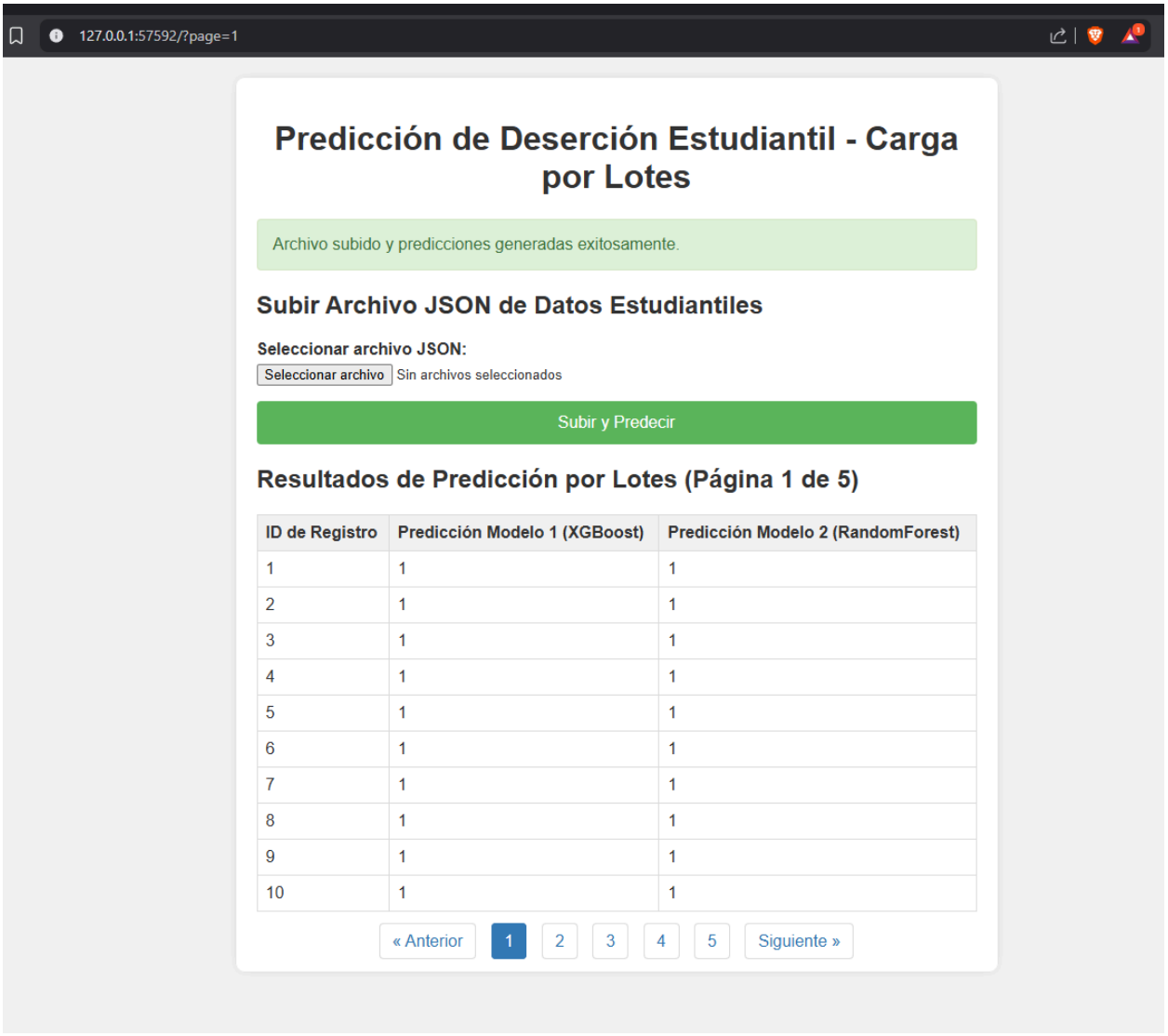

Ilustración 26. Visualización inicial datos cargados

Manual de Usuario - Sistema de Predicción de Deserción Estudiantil

Contenido

Introducción

El Sistema de Predicción de Deserción Estudiantil es una aplicación web desarrollada para predecir la probabilidad de deserción de estudiantes basándose en sus datos académicos y personales. El sistema utiliza dos modelos de aprendizaje automático:

- **Modelo 1**: XGBoost
- **Modelo 2**: RandomForest

Esta herramienta está diseñada para ayudar a las instituciones educativas a identificar a tiempo estudiantes en riesgo de deserción y así poder implementar estrategias de intervención adecuadas.

Requisitos para usar la aplicación

Requisitos técnicos

- Python 3.6 o superior
- Navegador web moderno (Chrome, Firefox, Edge, Safari)
- Conexión a internet (solo para la instalación inicial)

Dependencias principales

- Flask

- Joblib
- Scikit-learn
- Pickle

Instalación y configuración

1. **Instalación de dependencias**:

   pip install flask joblib scikit-learn

2. **Estructura de directorios requerida**: La aplicación espera la siguiente estructura de directorios:

   Proyecto de Grado/

   ├── app.py # Aplicación principal

   ├── models/ # Directorio para los modelos

   │ ├── best_model_1.pkl # Modelo XGBoost

   │ └── best_model_2.pkl # Modelo RandomForest

   ├── templates/ # Plantillas HTML

   │ └── index.html # Interfaz de usuario

   ├── uploads/ # Carpeta para archivos subidos

   └── primeros_50_registros.json # Ejemplo de formato JSON

3. **Configuración de modelos**:

- Asegúrese de que los modelos best_model_1.pkl (XGBoost) y best_model_2.pkl (RandomForest) estén ubicados en el directorio models/.

Iniciar la aplicación

1. Navegue a la carpeta raíz del proyecto:

   cd "Proyecto de Grado"

2. Ejecute la aplicación Flask:

   python app.py

3. Abra su navegador web y acceda a la siguiente dirección:

   http://127.0.0.1:5000

Uso del sistema

Cómo realizar predicciones por lotes

1. Prepare un archivo JSON con los datos de estudiantes:

   - El archivo debe contener información de uno o más estudiantes.
   - Cada registro debe incluir los 49 atributos definidos en el sistema (ver la lista completa en la sección de Formato de Datos).

2. En la interfaz web:

   - Haga clic en "Seleccionar archivo JSON" y busque el archivo que preparó.
   - Haga clic en "Subir y Predecir".

3. El sistema procesará los datos y mostrará los resultados de predicción para cada registro.

Formato de datos

El archivo JSON debe contener los siguientes atributos para cada estudiante:

```
{
  "edad_desertores": valor,
  "codigo_de_programa_desertores": valor,
  "codigo_snies_desertores": valor,
  "estrato_desertores": valor,
  ...
  [y otros 45 atributos]
}
```

> **Nota importante**: Puede utilizar el archivo primeros_50_registros.json como referencia para el formato correcto de datos.

Interpretación de resultados

Los resultados se presentan en una tabla con tres columnas:

1. **ID de Registro**: Identificador único para cada registro procesado.
2. **Predicción Modelo 1 (XGBoost)**: Resultado de la predicción del modelo XGBoost.
    - Un valor de 1 indica alta probabilidad de deserción.
    - Un valor de 0 indica baja probabilidad de deserción.

3. **Predicción Modelo 2 (RandomForest)**: Resultado de la predicción del modelo RandomForest.
    - Un valor de 1 indica alta probabilidad de deserción.
    - Un valor de 0 indica baja probabilidad de deserción.

Navegación entre resultados

Si hay muchos registros, los resultados se dividirán en páginas. Puede navegar entre ellas usando los controles de paginación en la parte inferior de la tabla de resultados.

Solución a problemas comunes

La aplicación no inicia

1. Verifique que todas las dependencias estén instaladas:

   pip install -r requirements.txt

2. Asegúrese de que los directorios models, templates y uploads existan en la carpeta del proyecto.

3. Verifique que los permisos de ejecución sean correctos.

Error al cargar modelos

Si la aplicación informa que no puede cargar los modelos:

1. Verifique que los archivos best_model_1.pkl y best_model_2.pkl estén en el directorio models/.

2. Asegúrese de que los modelos fueron guardados con versiones compatibles de scikit-learn y joblib.

Error al cargar archivo JSON

1. Verifique que el formato del archivo JSON sea válido.

2. Asegúrese de que todos los atributos necesarios estén presentes en el archivo.

3. Compare su archivo con primeros_50_registros.json para verificar el formato correcto.

La predicción muestra error o resultados inesperados

1. Verifique que los datos estén en el formato correcto y en el rango apropiado para cada atributo.

2. Asegúrese de que no falte ningún atributo en el archivo JSON.

3. Contacte al administrador del sistema si el problema persiste.

Conclusiones y trabajos futuros

El presente trabajo permitió explorar la aplicación de técnicas de aprendizaje automático para predecir la deserción estudiantil a partir de los datos históricos disponibles de aspirantes. Sin embargo, uno de los principales desafíos encontrados fue la calidad y estructura de los datos, ya que estos no estaban inicialmente diseñados para tareas predictivas, lo que limitó la confiabilidad de los modelos generados. La presencia de valores faltantes y estructuras inconsistentes dificulta la extracción de patrones robustos. A pesar de ello, mediante la imputación de datos con el algoritmo KNN, fue posible construir un conjunto de datos más completo y coherente, que sirvió como base para plantear escenarios de modelamiento experimental.

En estos escenarios, se entrenaron siete modelos de clasificación, entre los cuales Random Forest y XGBoost destacaron por sus altos desempeños, particularmente en la métrica de F1-score, lo que indica una buena capacidad para balancear precisión y recall en la detección de casos de deserción. No obstante, para la estrategia de despliegue se optó por un enfoque más robusto: el uso de un modelo ensamblado mediante una estrategia de voto de clasificadores. Esta técnica permitió combinar las fortalezas de varios modelos base, mejorando la estabilidad y generalización del sistema predictivo.

El modelo final fue almacenado utilizando Joblib para facilitar su reutilización y mantenimiento, y se desarrollaron componentes clave para su integración práctica: una API para consultas de predicción y un frontend web interactivo, pensado como interfaz para usuarios institucionales. Esta arquitectura modular permite que el sistema se escale y se integre fácilmente en infraestructuras existentes. De hecho, dada la estructura del sistema académico del IUPB, se considera viable su integración a mediano plazo como una herramienta de apoyo para la toma de

decisiones en procesos de admisión, seguimiento académico y gestión de la permanencia estudiantil.

Trabajos futuros

A futuro, se propone trabajar en la mejora y estandarización del proceso de recolección de datos de aspirantes, integrando formularios más estructurados, con validaciones automáticas y almacenamiento centralizado. Esto no solo mejorará la calidad de los datos, sino que permitirá desarrollar modelos más confiables y representativos de la realidad institucional.

También se contempla:

- Ampliar la base de datos con más cohortes y variables relevantes (factores socioeconómicos, académicos y emocionales).

- Explorar técnicas de feature engineering automatizado y modelos más avanzados como redes neuronales profundas.

- Implementar un sistema de monitoreo continuo del modelo en producción, para detectar posibles desviaciones de desempeño a lo largo del tiempo (concept drift).

Referencias bibliográficas

1. Akshaya, B., Raga, V. R., Reddy, G. S., & Chaitanya, B. K. (2024). Implementation of Silicon Wafer Defect Classification Web application using Deep Learning. 2024 International Conference on Advancements in Power, Communication and Intelligent Systems (APCI), 1–6. https://doi.org/10.1109/APCI61480.2024.10616637

2. Bishop, C. M. (2006). Pattern Recognition and Machine Learning.

3. Castro, L. F., Espitia, E., & Romero, E. (2023). Análisis de características que influyen en la deserción estudiantil. Revista EIA, 20(40).

4. Castro-Montoya, B. A., Mejía-Restrepo, M., & Múnera-Roldán, L. (2021). Modelo de riesgos competitivos para deserción y graduación. Formación Universitaria, 14(1), 81-98.

5. Chauhan, A., Arora, M., & Jain, R. (n.d.). Evaluation of real estate appraisal using ensemble methods. https://ssrn.com/abstract=3747575

6. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique.

7. Chen, T. (2016). XGBoost: A Scalable Tree Boosting System.

8. Cicconet, M. (2021). PuBliCiTy: Python Bioimage Computing Toolkit. https://doi.org/10.1101/2021.03.01.432926

9. DQLABS. (n.d.). The Impact of Data Quality on Model Performance. https://www.dqlabs.ai/blog/impact-of-data-quality-on-model-performance/

10. Erazo, X. F., & Rosero, E. (2021). Orientación vocacional y su influencia en la deserción universitaria. Revista Horizontes, 5(18), 591–606.

11. García, L., Aguilar, A., & Parada, A. (2022). Deserción universitaria en el contexto colombiano. Revista Senderos Pedagógicos, 13, 97-111.

12. Gutiérrez, M. D., & López, J. (2021). Indicadores de deserción universitaria y factores asociados. https://www.researchgate.net/publication/369196285

13. Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction.

14. Liu, H., Mao, M., Li, X., & Gao, J. (2025). Model interpretability on private-safe oriented student dropout prediction. PLOS ONE, 20(3), e0317726. https://doi.org/10.1371/journal.pone.0317726

15. Miño de Gauto, M. E. (2021). Factores condicionantes de la deserción universitaria. Ciencia Latina, 5(4), 5316–5318. https://doi.org/10.37811/cl_rcm.v5i4.691

16. Mohammed, S., Budach, L., Feuerpfeil, M., Ihde, N., Nathansen, A., Noack, N., Patzlaff, H., Naumann, F., & Harmouch, H. (2022). The Effects of Data Quality on Machine Learning Performance on Tabular Data. https://doi.org/10.1016/j.is.2025.102549

17. Moreno, M. I. (2022). Análisis de la deserción universitaria en Colombia. Universidad El Bosque.

18. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python.

19. SK, S., M.N.V, S. R. H., K, T. T., V, S. K., & K.T.V, M. (2024). Machine Learning Based Classification Model for Prediction of Bank Loan Approval. International Journal for Research in Applied Science and Engineering Technology, 12(3), 1885–1892. https://doi.org/10.22214/ijraset.2024.59231

20. Valero, J. E., Navarro, Á. F., & Larios, A. C. (2022). Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción. Revista de Ciencias Sociales, 28(3), 362–375. https://produccioncientificafluz.org/index.php/rcs/article/view/39385

21. Vélez, D. (2020). Análisis de alertas tempranas sobre deserción estudiantil en educación superior: Un estudio de caso para Medellín. Universidad Nacional de Colombia. https://repositorio.unal.edu.co/handle/unal/77597

22. https://github.com/rishabh1323/The-ML-DL-App