COMPARACIÓN DE MODELOS ESTADÍSTICOS Y DE APRENDIZAJE AUTOMÁTICO PARA LA PREDICCIÓN Y PREVENCIÓN DE LA DESERCIÓN ESTUDIANTIL EN LOS PROGRAMAS DE PREGRADO DE LA UNIVERSIDAD EAFIT

Comparison of Statistical and Machine Learning Models for Predicting and Preventing Student Dropout in Undergraduate Programs at Universidad EAFIT

JUAN CAMILO OSORIO GONZÁLEZ

Trabajo de Grado

Asesora

PhD. Paula María Almonacid Hurtado

UNIVERSIDAD EAFIT

ESCUELA DE CIENCIAS APLICADAS E INGENIERÍA

MAESTRÍA EN CIENCIAS DE LOS DATOS Y LA ANALÍTICA

MEDELLÍN

2024

CONTENIDO

RESUMEN

La deserción estudiantil es un fenómeno que afecta significativamente a estudiantes, instituciones educativas y la sociedad en general. En esta investigación se aborda el caso de la Universidad EAFIT y se exploran variables sociodemográficas, académicas e institucionales, entre otras, para proporcionar una comprensión integral de la deserción estudiantil. El objetivo principal es analizar las dinámicas de permanencia y deserción en la universidad, generando conocimientos que orienten la toma de decisiones y la formulación de estrategias efectivas de retención.

Se aplican modelos estadísticos y de aprendizaje automático utilizando una adaptación del proceso CRISP-DM para explicar y predecir la deserción estudiantil. Los modelos incluyen Regresión Logística, Árboles de Decisión y Random Forest, evaluados con métricas estándar y enfocándose en el F1-Score para medir la precisión del modelo. Además, se incorpora un análisis de supervivencia como aporte adicional para capturar el tiempo hasta la deserción y sus factores asociados.

Palabras clave: deserción, aprendizaje automático, modelos de clasificación, análisis de supervivencia.

ABSTRACT

This research focuses on the issue of student attrition at Universidad EAFIT, a phenomenon significantly impacting students, educational institutions, and society at large. It explores sociodemographic, academic, and institutional variables, among others, to provide a comprehensive understanding of student attrition. The primary objective is to analyze the dynamics of retention and attrition within the university, generating insights to inform decision-making and the development of effective retention strategies.

Statistical and machine learning models are applied using an adaptation of the CRISP-DM process to explain and predict student attrition. The models include Logistic Regression, Decision Trees, and Random Forest, evaluated with standard metrics and emphasizing the F1-Score to measure model accuracy. Additionally, survival analysis is incorporated as an additional contribution to capture time until attrition and its associated factors.

Keywords: attrition, machine learning, classification models, survival analysis.

INTRODUCCIÓN

La deserción estudiantil es uno de los desafíos más críticos que enfrentan las instituciones de educación superior en todo el mundo. Este fenómeno no solo afecta la vida académica de los estudiantes, sino que también tiene implicaciones significativas para las universidades, incluyendo la reducción en la tasa de graduación, pérdida de ingresos y deterioro de la reputación institucional. Identificar y prevenir la deserción estudiantil se ha convertido en una prioridad para muchas universidades, incluyendo la Universidad EAFIT.

En este contexto, los avances en el análisis de datos y la disponibilidad de información han permitido el desarrollo de modelos predictivos que pueden identificar con antelación a los estudiantes en riesgo de desertar. Estos modelos, basados en técnicas estadísticas tradicionales y de aprendizaje automático, ofrecen una herramienta poderosa para apoyar la toma de decisiones en la gestión educativa.

El presente trabajo de grado tiene como objetivo principal comparar la eficacia de diferentes enfoques predictivos para la deserción estudiantil en los programas de pregrado de la Universidad EAFIT. En particular, y partir de la información de los modelos más utilizados encontrados en la bibliografía, se evaluarán modelos estadísticos como la Regresión Logística, para abordar el problema de clasificación, y un Análisis de Supervivencia, para analizar la probabilidad de permanencia de los estudiantes en el tiempo. También se evaluarán modelos de aprendizaje automático, como Árboles de Decisión y Random Forest, con el fin de determinar cuál de ellos proporciona mejores resultados en términos de precisión predictiva y utilidad práctica para la intervención temprana.

El documento inicia con el planteamiento del problema y la justificación de este; posteriormente, se formulan los objetivos generales y específicos. Luego, se continúa con la revisión de literatura y el marco teórico, donde se exponen los

diferentes enfoques para abordar la deserción, las variables que influyen en el fenómeno y los modelos más comunes y útiles en el tema. A continuación, se presenta el diseño metodológico, que consta de una adaptación de la conocida metodología CRISP-DM, seguido del desarrollo del trabajo, donde se describen las diferentes etapas para la consolidación de la base de datos final y los detalles de la implementación de los diferentes modelos propuestos. Finalmente, se presentan los resultados obtenidos, las conclusiones del estudio y las referencias utilizadas.

PLANTEAMIENTO DEL PROBLEMA

Las Instituciones de Educación Superior tienen como uno de sus propósitos la formación de ciudadanos profesionales. Sin embargo, la dinámica de permanencia y deserción de los estudiantes en sus programas resulta ser un desafío crucial. De acuerdo con el mencionado propósito, se espera que un estudiante que ingrese a un programa de educación superior desarrolle su plan académico de forma satisfactoria y lo culmine dentro del tiempo previsto para ello. Pero se reconoce que diversos factores, tanto internos de la institución, como externos y propios del estudiante, pueden incidir negativamente en la permanencia y que interrumpa su ciclo académico o deserte. Que los estudiantes no logren culminar sus estudios va en contra del cumplimiento de los propósitos institucionales y, por consiguiente, incidirá negativamente en el desarrollo de la sociedad y el país.

El problema central, desde una perspectiva conceptual, es comprender las dinámicas de permanencia y deserción de los estudiantes de la Universidad EAFIT y, con base en ello, promover acciones que ayuden a mitigar la deserción. Desde una perspectiva técnica, el problema se traduce en identificar, con ayuda de la implementación de aprendizaje automático o Machine Learning, las variables que permitirán predecir, con algún grado de precisión, la deserción o permanencia de un estudiante en los programas de la Universidad.

JUSTIFICACIÓN

La deserción estudiantil en la educación superior es un problema que afecta a los estudiantes, las instituciones y la sociedad en general, pues implica pérdidas de recursos, oportunidades y desarrollo humano. Por ello, es necesario e importante investigar las causas y consecuencias de este fenómeno, así como diseñar e implementar estrategias para prevenirlo y reducirlo (Phan et al., 2023).

En la Universidad EAFIT, la media armónica de deserción durante los periodos comprendidos entre el 2009-1 y el 2019-1, fue del **38.48 %**, lo que significa que aproximadamente 38 de cada 100 estudiantes desertaron en algún momento a lo largo del desarrollo de su programa. Además, en los reportes de **EAFIT en Cifras** se aprecia que la deserción acumulada por cohorte es mayor en los primeros semestres de estudio, por lo que es válido adoptar un enfoque en la deserción temprana.

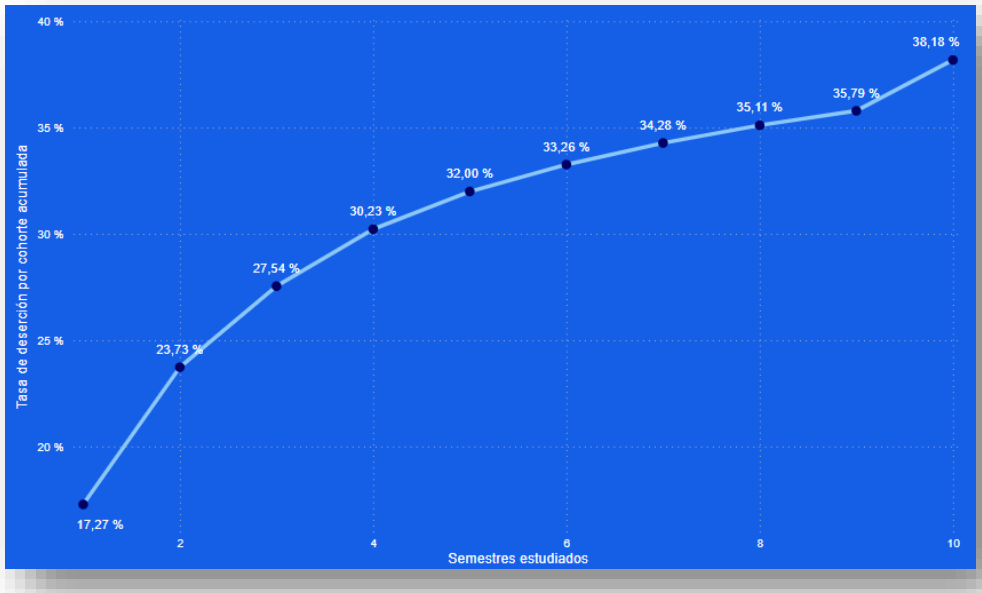

Ilustración 1: Tasa de deserción por cohorte acumulada. Fuente: EAFIT en Cifras.

Por lo anterior, realizar un estudio de deserción es necesario e importante para identificar los factores de riesgo y protección que afectan la permanencia de los estudiantes en la educación superior; diseñar e implementar acciones que promuevan la resiliencia de los estudiantes y les permitan enfrentar los desafíos académicos y personales; mejorar la calidad y la equidad de la educación superior, así como su impacto social y económico.

OBJETIVOS

GENERAL

Desarrollar e implementar modelos estadísticos y de aprendizaje automático que permitan identificar los factores más relevantes y predecir la deserción estudiantil en la Universidad EAFIT a nivel de pregrado, incluyendo un análisis de supervivencia para capturar el tiempo hasta la deserción y sus factores asociados, con el propósito de generar información y conocimiento que contribuyan a la toma de decisiones y al diseño de estrategias de retención efectivas.

ESPECÍFICOS

- Consolidar la base de datos final para hacer la modelación a partir de los conjuntos de datos suministrados por la Universidad.
- Implementar los modelos propuestos en la metodología, utilizando los datos disponibles y las librerías de Python necesarias.
- Incorporar y ejecutar el análisis de supervivencia para capturar el tiempo hasta la deserción y sus factores asociados.
- Analizar y presentar de forma informativa los resultados obtenidos tras la modelación y una comparación entre modelos.

REVISIÓN DE LITERATURA Y MARCO TEÓRICO

Esta revisión se estructura en dos secciones principales. La primera explora las definiciones y características de la deserción académica, proporcionando un marco conceptual claro. La segunda sección se centra en las variables utilizadas en los modelos que buscan comprender y predecir este fenómeno. Dentro de esta sección, se abordan tanto los modelos estadísticos, como la regresión logística y el análisis de supervivencia, como los modelos de aprendizaje automático basados en árboles.

DESERCIÓN UNIVERSITARIA

La deserción universitaria es un fenómeno complejo y multidimensional, que se refiere al abandono definitivo o temporal de los estudios superiores por parte de los estudiantes matriculados en una institución o programa educativo. Este fenómeno tiene implicaciones tanto individuales como sociales, ya que afecta el desarrollo personal y profesional de los estudiantes, así como el cumplimiento de las metas y políticas educativas de los países y las regiones (González, 2005).

El concepto de deserción universitaria ha sido abordado desde diferentes perspectivas teóricas y metodológicas, que han tratado de explicar sus causas, consecuencias y posibles soluciones. Entre las teorías más influyentes se encuentran la interaccionista de Tinto (1975, 1989) y la sociológica de Bourdieu (1979, 1984), como bien lo estudia Romito et al. (2020).

Según Tinto (1975, 1989), la deserción es el resultado de un proceso de desvinculación que ocurre cuando el estudiante no se integra bien al sistema académico y social de la institución. El compromiso del estudiante con sus metas e institución determina su integración o aislamiento, que influye en su persistencia o deserción. Tinto identifica varios factores individuales e institucionales que afectan este proceso, como las características personales, las experiencias previas, las expectativas, el rendimiento académico, el apoyo familiar, la interacción con pares, etc.

Bourdieu (1979, 1984) propone que la deserción es el resultado de una falta de correspondencia entre el habitus y el campo académico. El habitus es el conjunto de disposiciones, gustos, valores y prácticas que el individuo adquiere por su socialización en un contexto, y que guía sus acciones y percepciones. El campo es el espacio social donde se dan las relaciones de poder y de lucha entre los agentes que tienen diferentes tipos y cantidades de capital (económico, cultural, social, simbólico, etc.). La deserción ocurre cuando el habitus del estudiante no se adapta al campo académico, y cuando el estudiante no tiene suficiente capital para competir y legitimarse en ese campo.

Spady (1970) propone un modelo basado en Durkheim que se enfoca en la interacción entre los atributos del estudiante y el entorno universitario. Acevedo (2021) plantea que la deserción depende de la definición, la perspectiva y el contexto de los actores involucrados. Otras aproximaciones (Díaz, 2008) han intentado explicar la deserción desde diferentes perspectivas, aportando elementos para comprenderla y prevenirla. Sin embargo, no hay una definición, una medición y una comparación únicas y consensuadas de la deserción universitaria, sino que dependen de las características y condiciones de los sistemas educativos, las instituciones y los estudiantes, así como de sus procesos y trayectorias educativas (Acevedo 2021). Por lo tanto, se requiere una perspectiva más amplia y contextualizada de la deserción, que considere la diversidad y la complejidad de los factores y las dimensiones que la conforman. Según Gonzalez (2005), el abandono universitario se puede entender desde dos focos, con respecto al tiempo (inicial, temprana y tardía) y con respecto al espacio (institucional, interna y del sistema educativo). Entendiendo esto y desde el punto de vista interaccionista de Tinto (1989) en la cual se entiende la deserción desde tres puntos de vista: el individual, institucional y estatal, se puede tener en cuenta las siguientes dimensiones de la deserción para comprender mejor este fenómeno:

La dimensión individual, que se refiere a las características, motivaciones, actitudes, expectativas, comportamientos y decisiones de los estudiantes que desertan o persisten en sus estudios.

La dimensión institucional, que se refiere a las características, políticas, prácticas, recursos, servicios y resultados de las instituciones de educación superior donde se matriculan y se forman los estudiantes.

La dimensión estatal, que se refiere a las características, normas, planes, programas, indicadores y evaluaciones de los sistemas y subsistemas de educación superior de los países y las regiones donde se ubican las instituciones y los estudiantes.

La dimensión temporal, que se refiere al momento o al período en que se produce la deserción, que puede ser al inicio, durante o al final de los estudios, o en función de la duración o la intensidad de la interrupción.

La dimensión espacial, que se refiere al lugar o al nivel en que se produce la deserción, que puede ser en una institución, en un programa, en una carrera, en una modalidad, en una sede, en una facultad, en un departamento, etc.

Otro aspecto para considerar es que, debido a este carácter polisémico asociado a la deserción y entendiendo que esta puede variar de un contexto a otro, Tinto propone que los investigadores e instituciones deben elegir con cuidado las definiciones que mejor se ajusten a sus intereses y metas, teniendo en cuenta la amplia gama y variantes de la deserción en sus diferentes dimensiones.

Definiciones de deserción

La definición de deserción varía según la institución y depende de sus políticas internas y objetivos específicos en el análisis de este fenómeno. Sin embargo, algunos estudios han establecido marcos temporales concretos para definir a un estudiante como desertor. Por ejemplo, Fernández- Martín et al. (2018) consideran que un estudiante es desertor si no ha vuelto a matricularse en el mismo programa

después de dos años. Por otro lado, Hoyos Osorio y Daza Santacoloma (2023) definen como desertor a un estudiante que no ha renovado su matrícula durante al menos dos semestres consecutivos en programas específicos. Asimismo, Lopes et al. (2023) clasifican a un estudiante como desertor si se ha ausentado durante un semestre o periodo académico.

Para la Universidad EAFIT, la definición más adecuada se alinea con la propuesta de Hoyos y Daza. En este contexto, se considera desertor a un estudiante que ha estado al menos dos periodos consecutivos sin matricularse en su programa, sin haber realizado algún reingreso o reintegro, ya sea al mismo programa u otro.

MODELOS Y ENFOQUES PARA LA EXPLICACIÓN Y PREDICCIÓN DE LA DESERCIÓN UNIVERSITARIA

Los modelos predictivos son herramientas que estiman la probabilidad de deserción de los estudiantes, basándose en técnicas estadísticas o de aprendizaje automático, que analizan datos históricos o actuales de los estudiantes y los contextos de aprendizaje. Estos modelos se enmarcan en las Analíticas Académicas (AA), que son una herramienta valiosa para la educación superior, ya que permiten recopilar y analizar información estática y dinámica sobre los procesos educativos y proporcionar información útil, oportuna y personalizada, que pueda orientar la gestión y la intervención educativa. El objetivo principal de los modelos predictivos es facilitar la toma de decisiones y la implementación de acciones preventivas o correctivas, que puedan mejorar la retención y el éxito académico de los estudiantes. Sin embargo, la construcción y la aplicación de estos modelos implica una serie de desafíos y limitaciones, tanto conceptuales como metodológicos, que deben ser considerados y superados (Norambuena et al, 2022).

Norambuena et al (2022) realizaron una revisión sistemática sobre los modelos predictivos de deserción en educación superior, en la que analizaron estudios de 9 países (Países bajos, Australia, Italia, España, Turquía, Taiwan, Israel, Ecuador y Estados Unidos). Los resultados muestran que las variables utilizadas en los

modelos se agrupan en tres categorías: sociodemográficas, extraídas de LMS y sociocognitivas. Además, el tamaño de muestra más común es el rango entre 101 a 500 estudiantes. El estudio concluye que los modelos predictivos que usan AA tienen un gran poder predictivo del rendimiento académico y que es necesario incorporar variables psicoeducativas para comprender mejor el abandono universitario. **También resalta la escasez de investigaciones sobre modelos predictivos para estudiar la deserción en América Latina**, en contraste con los países anglosajones. Finalmente, el estudio señala las limitaciones de los métodos predictivos tradicionales y su falta de efectividad para predecir el riesgo de deserción.

Variables utilizadas en los modelos predictivos de deserción

Los modelos predictivos de deserción universitaria utilizan variables de tres categorías: sociodemográficas, académicas y sociocognitivas. Las variables sociodemográficas son características personales y sociales de los estudiantes, como el género, la edad, el nivel socioeconómico, etc., que no se pueden modificar por la intervención educativa. Las variables académicas son características, resultados y comportamientos de los estudiantes en el ámbito educativo, como el promedio, las calificaciones, la asistencia, el uso de los recursos, etc., que se pueden modificar por la intervención educativa. Las variables sociocognitivas son características, actitudes y creencias de los estudiantes relacionadas con el aprendizaje, como la motivación, la satisfacción, la autoeficacia, el compromiso, etc., que también se pueden modificar por la intervención educativa. Norambuena et al (2022) señalan que las variables sociocognitivas son una brecha en la literatura y que no muchos modelos las tienen en consideración.

Algunas de las variables que deben ser tenidas en cuenta de acuerdo con la revisión de estudios de modelos predictivos como Norambuena et al (2022) y Villareal et al (2023) se pueden evidenciar en la Tabla 1, en la cual se puede observar más detalladamente por categorías las variables susceptibles a ser utilizadas en un modelo de predicción:

Categoría	Variable
Demográficas y Socioeconómicas	Edad
	Sexo
	Estado civil
	Estrato social
	Ingresos familiares
	Migración
	Dependencia económica
	Ciudad de residencia
	Tipo de colegio (público o privado)
	Puntaje en pruebas estandarizadas (Saber 11, PSU, etc.)
	Beneficios obtenidos al ingresar
	Número de hermanos
	Educación y ocupación de los padres
	Modalidad de ingreso (apoyo económico, recursos propios, cuotas, méritos, etc.)
Académicas	Notas del semestre anterior
	Notas de matemáticas en la prueba de admisión
	Rendimiento académico
	Número de asignaturas perdidas
	Promedio del primer semestre
	Resultado de la prueba de admisión
	Número de créditos aprobados
	Reprobación de años durante el bachillerato
	Tipo de programa o carrera
Comportamiento y Adaptación	Participación en actividades deportivas, culturales o de liderazgo
	Relación docente-estudiante
	Necesidades y adaptación curricular
	Intención de abandono
	Tiempo de permanencia en la universidad
	Rendimiento en el primer semestre de estudio
	Interrupción de los estudios
Relacionadas con el Proceso de Aprendizaje	Flujo de clics en el entorno virtual de aprendizaje (LMS)
	Actividad en el entorno virtual de aprendizaje
	Calificación en tareas y exámenes
	Estilo de aprendizaje individual
Salud y Bienestar	Nivel de depresión y ansiedad
	Pruebas de detección de alcohol, tabaquismo y sustancias involucradas

	Autoeficacia social cognitiva y conductual
	Agotamiento emocional
Institucionales y Contexto Universitario	Infraestructura universitaria
	Flexibilidad curricular
	Organización administrativa
	Relación con el entorno universitario
Participación Estudiantil	Participación estudiantil (debates, asambleas, congresos)
	Integración social y sentido de pertenencia
	Participación en grupos estudiantiles / o semilleros de investigación
	Relación con el entorno universitario

Tabla 1: Variables utilizadas en modelos predictivos de deserción universitaria.

Modelos estadísticos y de aprendizaje automático utilizados

En el ámbito de la predicción y análisis de deserción estudiantil, diversos modelos y enfoques han sido explorados con el fin de comprender y anticipar patrones de abandono en instituciones educativas. Uno de los enfoques destacados es el empleo de técnicas de ensamble por apilamiento (Talamás-Carvajal & Ceballos, 2023). Estos autores desarrollaron dos modelos distintos: uno enfocado en el período post-admisión y otro en el post-primer semestre. Ambos modelos utilizaron un clasificador de ensamble por apilamiento con componentes base que incluyeron un clasificador de árbol de decisión, K-vecinos más cercanos, Naive Bayes y regresión logística. La evaluación de estos modelos se llevó a cabo considerando métricas clave como precisión, exactitud, exhaustividad, puntaje F1 y área bajo la curva ROC – AUC.

En otro contexto, se exploraron modelos explicativos de regresión multinomial y predictivos de aprendizaje automático para identificar patrones de deserción estudiantil. Se evaluaron seis métodos, entre ellos boosting trees, random forest, árboles de decisión, redes neuronales, support vector machines y regresión logística. Destacaron que el algoritmo "random forest" demostró ser el más eficaz al identificar correctamente a potenciales desertores con una probabilidad del 83%,

capturando al mismo tiempo el 34% de la deserción real (Fernández- Martín et al., 2018).

Hoyos Osorio y Daza Santacoloma (2023) contribuyeron al campo mediante el uso de técnicas de validación cruzada y métodos para datos desbalanceados. Su enfoque incluyó la codificación de variables categóricas, detección y eliminación de valores atípicos, sobremuestreo de la clase minoritaria y selección de características recursivas. Además de métricas convencionales, introdujeron la media geométrica (G-mean) para evaluar el equilibrio entre sensibilidad y especificidad del modelo.

En el ámbito de la educación superior online, Rodríguez Velasco et al. (2023) emplearon un algoritmo de eliminación recursiva de características y diversas técnicas para abordar el desafío de datos desbalanceados, como SMOTE, ADASYN y penalización de hiperparámetros. Utilizando regresión logística, bosques aleatorios y redes neuronales, destacaron la eficacia del ajuste del umbral de probabilidad óptimo, alcanzando valores de sensibilidad de 0.75, 0.67 y 0.6 para los respectivos clasificadores.

Lopes et al. (2023) por su parte, se centraron en analizar la deserción de estudiantes de pregrado utilizando modelos de regresión logística. No se abordó explícitamente el desbalance de datos. Identificaron algunos factores asociados a la deserción, resaltando la diversidad de covariables entre diferentes centros de enseñanza, resaltando que hay diferentes efectos asociados al departamento al que pertenezcan los programas.

Finalmente, un estudio aplicado en la Universidad Nacional de Moquegua (UNAM), Perú, presentó ocho modelos predictivos basados en minería de datos. Utilizando WEKA y aplicando técnicas como SMOTE para balancear las clases, encontraron que el modelo basado en Random Forest demostró la mayor precisión y robustez, superando resultados de trabajos relacionados y ofreciendo una herramienta valiosa para prevenir el abandono estudiantil (Flores et al., 2022).

En conjunto, estos enfoques y modelos aportan una rica perspectiva al campo de investigación sobre deserción estudiantil, destacando la diversidad de métodos utilizados y sus respectivos éxitos en la identificación y predicción de patrones de abandono. En la siguiente imagen se aprecia, en la revisión de literatura realizada, los modelos y técnicas utilizadas en el aprendizaje estadístico y automático, para el problema de la deserción estudiantil, resaltando como más relevantes por frecuencia y obtención de resultados adecuados la regresión logística, random forest, árboles de decisión y redes neuronales, específicamente, las ANN (redes neuronales artificiales):

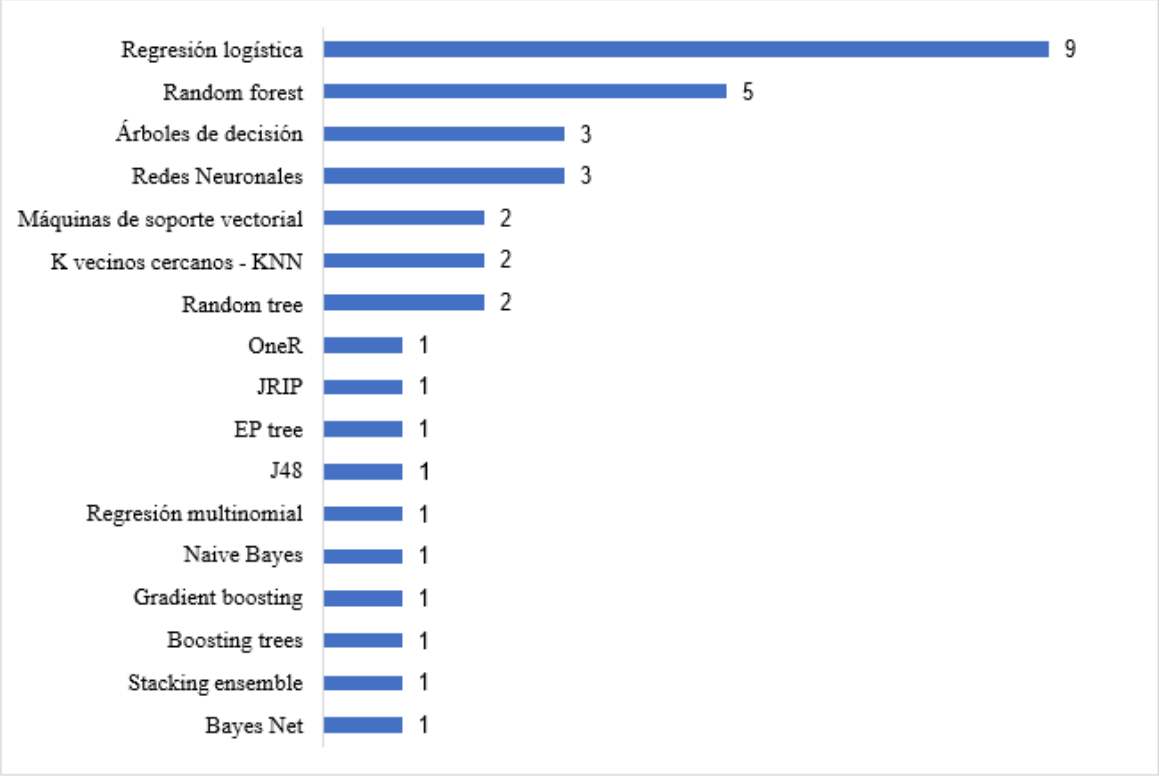

Ilustración 2: Modelos y técnicas más utilizadas. Fuente: elaboración propia.

MODELOS DE SUPERVIVENCIA

Los modelos de supervivencia han demostrado ser herramientas valiosas para estudiar la deserción estudiantil. Un artículo publicado por el *International Journal of*

Educational Development destaca que estos modelos permiten medir el tiempo hasta que ocurre un evento crítico, como la finalización de un curso o la deserción, proporcionando una comprensión más profunda de los factores que influyen en la permanencia de los estudiantes en la educación superior (Silva & Sampaio, 2023). El estudio de la Universidad SIMAD resalta la capacidad de los modelos de supervivencia para identificar factores de riesgo significativos, como el nivel de grado académico, ausentismo, dificultades de aprendizaje e interés del estudiante en la escuela, cruciales para predecir y mitigar la deserción (Ali & Hussein, 2024). La PUCC encontró que el uso de estos modelos fue valioso para identificar factores de riesgo asociados con la deserción estudiantil y las graduaciones tardías, permitiendo a las autoridades universitarias inspirar políticas para mitigar estos eventos (Vallejos & Steel, 2017). MDPI resaltó la importancia de las variables académicas y temporales en la predicción de la deserción, destacando el rendimiento de modelos como DeepSurv en el análisis de supervivencia (Gutierrez-Pachas et al., 2023). En el Higher Education Research and Development demostraron que los modelos de supervivencia permiten examinar la deserción estudiantil y la finalización de grados a lo largo del tiempo, identificando patrones de deserción y finalización en función de variables como el trastorno psicológico (Cvetkovski et al., 2018). En una publicación realizada en el *Journal of Latinos and Education* se concluye que estos modelos son valiosos para predecir y prevenir la deserción estudiantil al considerar el tiempo hasta el evento de interés, permitiendo identificar factores de riesgo y tomar medidas preventivas para mejorar las tasas de retención estudiantil (Arias et al., 2024). En otra investigación de este journal destacan que el uso de modelos de supervivencia, tanto clásicos como extendidos, ha permitido modelar la complejidad de los datos de estudio, como la censura, variables que varían con el tiempo y observaciones ampliamente heterogéneas, proporcionando información relevante para la implementación de políticas institucionales y públicas (Buenaño et al., 2023).

Regiones de estudio

Los estudios se han llevado a cabo en diversas regiones, proporcionando una perspectiva global sobre la deserción universitaria. Algunos artículos mencionan que hay investigaciones en Brasil, Colombia, Italia y Alemania (Silva & Sampaio, 2023). La Universidad SIMAD realizó estudios en Ghana y Somalia (Ali & Hussein, 2024). En Chile, el estudio de la PUCC se centró en programas de pregrado seleccionados durante el período 2000-2011 (Vallejos & Steel, 2017). El estudio de MDPI se enfoca en el contexto latinoamericano, destacando la necesidad de estrategias personalizadas para abordar la deserción en esta región (Gutierrez-Pachas et al., 2023). Cvetkovski, Jorm y Mackinnon llevaron a cabo sus investigaciones en Australia, utilizando una muestra nacional de estudiantes universitarios a través de una encuesta longitudinal llamada HILDA (Cvetkovski et al., 2018). Arias, Linares-Vásquez y Héndez-Puerto aplicaron su estudio en diferentes regiones de Colombia, incluyendo universidades como la Universidad de Antioquia, la Universidad Pedagógica y Tecnológica de Colombia, y la Universidad de Caldas (Arias et al., 2024). Buenaño, Beletanga y Mancheno se centraron en Ecuador, en instituciones de educación superior (Buenaño et al., 2023). Esta diversidad regional permite comparar y contrastar los factores de deserción en diferentes contextos culturales y educativos.

Relevancia de algunas variables de segmentación

El género y la edad son factores críticos en la deserción universitaria, según diversos estudios. El International Journal of Educational Development (2021) destaca su influencia en las probabilidades de deserción y en los tiempos de graduación. La Universidad SIMAD encontró diferencias de género significativas en los patrones de deserción (Ali & Hussein, 2024), mientras que en la PUCC se observaron variaciones en las tasas de deserción y graduación según la edad y el género de los estudiantes (Vallejos & Steel, 2017). El estudio de MDPI resalta que el género y la edad afectan el rendimiento académico y las decisiones de permanencia (Gutierrez-Pachas et al., 2023), y Higher Education Research and

Development sugiere que los estudiantes mayores tienen más probabilidades de abandonar sus estudios (Cvetkovski et al., 2018). En términos de grado y tipo de institución educativa, estas variables también influyen significativamente. Silva & Sampaio (2023) señalan que los estudiantes en grados más avanzados tienen menos probabilidades de desertar, mientras que el tipo de institución puede presentar desafíos diferentes para los estudiante. La investigación de la PUCC destaca el impacto del tipo de escuela secundaria y la preferencia de aplicación en las tasas de deserción (Vallejos & Steel, 2017), y MDPI subraya la importancia de estas variables en la predicción de la deserción (Gutierrez-Pachas et al., 2023). Estos estudios enfatizan la necesidad de adaptar las políticas de retención a las características específicas de los estudiantes y las instituciones (Arias et al., 2024; Buenaño et al., 2023).

Horizonte de observación y cantidad de datos

Los estudios sobre deserción universitaria varían significativamente en cuanto a la cantidad y granularidad de datos utilizados, así como en los horizontes temporales evaluados. Silva & Sampaio (2023) emplearon datos anuales de panel desde 2010 hasta 2021, abarcando características estudiantiles como becas, ingresos familiares y actividades remuneradas a lo largo de varios años. La Universidad SIMAD (Ali & Hussein, 2024) se basó en un conjunto específico de datos de 70 estudiantes de economía para el año académico 2017-2018, ofreciendo una visión detallada dentro de un contexto más limitado. La investigación de la PUCC (Vallejos & Steel, 2017) analizó datos extensos de 27,189 estudiantes de pregrado durante el período 2000-2011, registrando eventos de graduación, deserción y observaciones censuradas. MDPI (Gutierrez-Pachas et al., 2023) utilizó datos de múltiples instituciones en América Latina para evaluar modelos predictivos y correlacionales con horizontes temporales amplios y variados. Cvetkovski et al. (2018) emplearon una muestra nacional amplia y económica para datos longitudinales desde 2001 hasta 2011, observando la deserción y finalización de grados a lo largo de más de una década. Arias et al. (2024) y Buenaño et al. (2023) también utilizaron horizontes temporales

específicos desde el ingreso de los estudiantes hasta posibles eventos de deserción, cada uno enfocándose en aspectos particulares de la trayectoria académica.

DISEÑO METODOLÓGICO

El diseño metodológico de este estudio sobre la deserción académica en la Universidad Eafit sigue una adaptación de la metodología CRISP-DM, excluyendo el despliegue final debido a los objetivos planteados. Así:

- **Comprensión del Negocio y de los Datos:** se describe el contexto del problema de la deserción académica, ya abordado en las secciones de descripción del problema y justificación. Se exploran los datos disponibles y se evalúa su origen y fidelidad, identificando variables relevantes y abordando problemas como datos faltantes o inconsistencias.
- **Preparación de los Datos:** se prepara la base de datos final a partir de diferentes fuentes de datos recopiladas. Esto incluye la limpieza y transformación de los datos, así como la creación de nuevas variables cuando es necesario. Se identifican y abordan los retos específicos relacionados con la implementación de los modelos propuestos.
- **Modelado:** los modelos seleccionados se eligieron en función de su prevalencia en la revisión bibliográfica. Se implementaron métodos supervisados de clasificación, entre los cuales se incluyen:
    - **Regresión Logística:** Se utiliza tanto la regresión logística clásica como versiones regularizadas con L1 (Lasso) y L2 (Ridge). También se implementan métodos de selección de características como el Stepwise.
    - **Árboles de Decisión y Random Forest:** Se utilizan para capturar relaciones no lineales y complejas entre las variables.

- **Optimización de Modelos:** se aplica el método de Grid Search Cross Validation para seleccionar los mejores modelos y configurar los hiperparámetros óptimos mediante la evaluación del rendimiento del modelo utilizando validación cruzada.
- **Análisis de Supervivencia:** se implementa un análisis de supervivencia para entender las probabilidades de supervivencia de los estudiantes y predecir sus curvas de supervivencia. Este enfoque ayuda a identificar cuándo es más probable que ocurra la deserción.
- **Evaluación de Resultados:**
    - Para los modelos de regresión logística, árboles de decisión y random forest, se utilizan métricas clásicas como precisión, sensibilidad, especificidad y especialmente el F1 Score para tener en cuenta el desbalance de clases.
    - Para los modelos de supervivencia, se utilizan el índice de concordancia, la significancia de los parámetros y las curvas de supervivencia predichas.

Finalmente, se comparan los resultados de los distintos modelos y se presentan las conclusiones del estudio.

DESARROLLO DEL TRABAJO

CONTEXTO DEL NEGOCIO Y LOS DATOS

A finales del año 2022, la Universidad EAFIT contaba con aproximadamente 44.479 estudiantes en todos los grados académicos, incluyendo pregrado, posgrado, idiomas, educación continua, alta dirección, saberes de vida, Instituto Confucio, universidad de los niños y Nodo. Y solo el nivel de **pregrado representa el 21,1% con 9.383 estudiantes**. Como se ha mencionado anteriormente, es crucial para la Universidad entender las dinámicas de la deserción académica. La revisión de la

literatura proporciona una base para enfocar el estudio a entender cómo ciertas variables se relacionan con este fenómeno y en qué medida son relevantes.

En este trabajo, la deserción académica se define específicamente como la deserción temprana, es decir, la de los estudiantes que abandonan sus estudios en los primeros semestres. Esta definición se ha adoptado debido a la limitada disponibilidad de algunos datos considerados para el estudio, específicamente los relacionados con la información sociodemográfica y de ingreso, debido a la entrada en vigor de un nuevo sistema de información en el año 2022, a pesar de que se cuenta con datos académicos de mayor antigüedad. La deserción académica temprana se refiere a los estudiantes que abandonan voluntariamente o por bajo rendimiento académico en los primeros cuatro semestres de su carrera y que permanecen retirados por al menos dos semestres consecutivos y no realizan un reintegro o reingreso al mismo programa u otro. El estudio abarca datos de un total de 1.181 estudiantes de pregrado admitidos en los dos ciclos lectivos del año 2022, es decir, semestres 2022-1 y 2022-2, analizándolos desde el momento de su admisión hasta el semestre 2024-1. Esto proporciona un período de observación de cuatro semestres para los estudiantes admitidos en 2022-1 y de tres semestres para los admitidos en 2022-2. No se eligieron estudiantes de semestres anteriores debido a la ausencia de los tipos de datos mencionados anteriormente.

Los datos disponibles para el estudio incluyen aspectos personales, familiares, sociodemográficos, académicos, y el uso de los programas de bienestar y salud de los estudiantes. La variable de interés para los modelos de clasificación es ***Deserción***, que es binaria: toma el valor de 1 si el estudiante desertó y 0 si continuó. Para el modelo de supervivencia, se utiliza la variable ***Tiempo***, que es numérica y define el tiempo de observación del estudiante, ya sea que haya desertado o continúe en el sistema.

Dado que la deserción universitaria tiene múltiples dimensiones, se evalúan diversos aspectos del perfil del estudiante. Desde la dimensión académica, se

analiza el desempeño general del estudiante. Desde la dimensión socioeconómica, se consideran factores como el nivel de ingresos y el apoyo financiero familiar. Y desde la dimensión institucional, se tienen en cuenta factores como la pertenencia a grupos estudiantiles y el uso de los servicios de bienestar. Con base en estos datos, se construyen modelos predictivos que integren estas dimensiones y proporcionen información sobre la probabilidad de deserción.

PREPARACIÓN DE LOS DATOS

Los siguientes conjuntos de datos fueron proporcionados para el desarrollo de este estudio por las áreas que administran los diferentes tipos de datos, bajo un compromiso de confidencialidad firmado con la Universidad:

Dataset	**Descripción**
Balance Académico	Contiene los datos de promedios, créditos cursados y condición académica de los estudiantes de pregrado de la Universidad.
Planilla de citas	Contiene la información de los estudiantes que han hecho uso del servicio médico de la Universidad, puede extraerse la cantidad de veces que un estudiante asistió por primera vez y control durante un periodo específico.
Historia Académica	Contiene la información de las calificaciones obtenidas por todos los estudiantes por ciclo lectivo o semestre.
Formulario de Inscripción	Contiene la información que los estudiantes registran en el formulario de inscripción al programa en la Universidad. Contiene información personal y de familiares.
Servicios Académicos	Contiene la información de todos los servicios académicos solicitados por los estudiantes. Interesa conocer los servicios de cancelación de asignaturas solicitados y gestionados por ciclo lectivo y las asignaturas canceladas.

Becas	Contiene la información de las becas o apoyos económicos, por valor en pesos o porcentaje, asignados a los estudiantes beneficiados.
Atención	Contiene la información de los estudiantes que solicitaron apoyo de bienestar en tres servicios: apoyo psicosocial, técnicas de estudio y orientación vocacional.
Grupos Estudiantiles	Contiene la información de los grupos estudiantiles de la Universidad y los datos de los integrantes de estos.
Títulos Externos	Contiene información de si los estudiantes que ingresaron tienen algún otro título antes de ingresar
Estudiantes Retirados	Contiene la información de los estudiantes que han sido retirados de los programas de la Universidad por cualquier causa
Estudiantes Activos	Contiene la información de los estudiantes que siguen activos en los programas de la Universidad

Tabla 2: Descripción de los conjuntos de datos a utilizar en el estudio.

Ensamble de la base de datos final

Muestra de datos y variable respuesta

Se realizaron las transformaciones necesarias para que los datos sean adecuados para el análisis y la modelación. Para ello se tiene en cuenta la definición de deserción que se estableció, lo que implica seleccionar los periodos de los cuales se extraen los datos de los estudiantes. Además, se debe generar la variable respuesta, ***Deserción***, que indica si el estudiante desertó (1) o continuó (0), según el criterio definido, así como el tiempo que el estudiante lleva en el sistema.

Se eligió el conjunto muestra a partir de la tabla de Formulario de Inscripción, seleccionando los estudiantes cuyo ciclo de admisión fuera 2022-1 y 2022-2 y cuyo estado de matrícula fuera "Activa". Esto resultó en un total de 1,181 estudiantes distribuidos así:

2022-1: 770

2022-2: 411

Se construyó la variable respuesta, "Deserción", a partir de las tablas de Estudiantes Retirados y Estudiantes Activos. Se creó una función para determinar cuidadosamente el estado real del estudiante, considerando aquellos que pudieron haberse trasladado a otros programas. Tras este proceso, se identificaron 1,178 estudiantes con 972 no desertores y 206 desertores.

Se añadieron varias variables adicionales, como promedios académicos, relación de créditos aprobados/cursados, cancelaciones de clase, apoyos económicos, uso del servicio médico y participación en grupos estudiantiles. Se seleccionaron datos del Formulario de Inscripción, incluyendo Programa Académico, Tipo Colegio, Estrato, Tipo Vivienda, Ingresos Familiares, entre otros.

Manejo de datos nulos

Se evaluaron y trataron los datos faltantes. Variables con altos niveles de datos nulos, como Tipo de Colegio, Labora Actualmente y Posición entre Hermanos, fueron eliminadas por el riesgo que trae hacer una imputación de datos no cuidadosa. Para variables con pocos datos nulos, como promedios académicos y relación de créditos aprobados/cursados, se utilizó la imputación basada en la mediana.

Variable	Cantidad de datos nulos
ID Estudiante	0
y: deserción	0
Promedio Semestral	14
Promedio Acumulado	14
Relación Créditos: Aprobados/Cursados	15
Programa Académico	0
Tipo Colegio	306
Labora Actualmente	425
Estrato	0
Tipo Vivienda	0

Ingresos Familiares	0
Número de Hermanos	0
Posición entre Hermanos	247
Víctima Conflicto	0
Víctima Desplazamiento	0
Grupos Minoritarios	0
Limitaciones Discapacidad	0
Cancelaciones de Clase Aceptadas	0
Cantidad de Apoyos Económicos	0
Uso Servicio Médico	0
Grupo Estudiantil	0

Tabla 3: Resumen del conteo de la cantidad de datos nulos en el set de datos final.

Categorización de variables no numéricas

Se agruparon y codificaron las variables categóricas, como Programa Académico (agrupado por escuelas) y Tipo Vivienda, utilizando el método OneHot, mientras que otras variables fueron convertidas a binarias o numéricas según correspondiera:

Variable	**Método**
Programa Académico	Método OneHot
Tipo Vivienda	Método OneHot
Ingresos Familiares	Se convierte a numérica
Limitación Discapacidad	Binaria: manualmente
Víctima Conflicto	Binaria: manualmente
Víctima Desplazamiento	Binaria: manualmente
Grupos Minoritarios	Binaria: manualmente

Tabla 4: Resumen de los métodos de categorización de variables.

Construcción de datos para el modelo de supervivencia

Variable tiempo

Se creó una función para calcular el número de semestres que cada estudiante ha estado en la Universidad desde su ingreso hasta la deserción o el final del periodo de observación.

Tiempo (semestres)	Cantidad estudiantes
5	645
4	377
2	70
3	50
1	36

Tabla 5: Balance de la cantidad de muestras (estudiantes) según el tiempo de observación.

Variable evento

Esta variable es simplemente la misma variable **Deserción** ya definida anteriormente.

Balance del set de datos final

Al ejecutar las diversas estrategias de depuración y construcción del set de datos final, este quedó con una cantidad de registros igual a 1178 muestras y 21 variables, en el que se presentan 206 estudiantes desertores y 972 no desertores.

Análisis descriptivo y de correlación de variables

Las variables consideradas inicialmente tienen los siguientes resultados descriptivos:

	count	mean	std	min	25%	50%	75%	max
EA_PROMEDIO_SEM	1178.0	3.854593	0.665084	0.03	3.62	3.970	4.27	4.96
EA_PROMEDIO_ACU	1178.0	3.876265	0.656693	0.00	3.67	3.995	4.27	4.96
CRED_APR/CURS	1178.0	0.931045	0.188022	0.00	1.00	1.000	1.00	1.00
ESTRATO	1178.0	4.114601	1.456106	1.00	3.00	4.000	5.00	6.00
INGRES_FAMIL	1178.0	9.613752	5.427929	1.00	5.00	10.000	15.00	17.00
NUM_HERMANOS	1178.0	1.220713	0.965248	0.00	1.00	1.000	2.00	4.00
TOTAL_CANCLA_ACEPTADA	1178.0	0.685908	0.799206	0.00	0.00	0.000	1.00	3.00
CANT_TOTAL_APOYOS_ECONOM	1178.0	0.481324	1.100741	0.00	0.00	0.000	0.00	6.00
VECES_USO_SERV_MED	1178.0	0.103565	0.433677	0.00	0.00	0.000	0.00	5.00

Ilustración 3: variables no binarias

	count	unique	top	freq
TIPO_VVDA_Arriendo	1178	2	0.0	919
LIMIT_DISCAPACIDAD	1178	2	0	1171
APOYO_ESTUDIANTIL	1178	2	0.0	963
GRUPO_ESTUDIANTIL	1178	2	0.0	1047
VICTIM_CONFL_DESPL	1178	2	0	1128
PROG_ACAD_Escuela de Administración	1178	2	0.0	802
PROG_ACAD_Escuela de Artes y Humanidades	1178	2	0.0	979
PROG_ACAD_Escuela de Ciencias Aplicadas e Ingeniería	1178	2	0.0	828
PROG_ACAD_Escuela de Derecho	1178	2	0.0	1053
PROG_ACAD_Escuela de Finanzas, Economía y Gobierno	1178	2	0.0	1050
TIPO_VVDA_Arriendo	1178	2	0.0	919
TIPO_VVDA_Familiar	1178	2	0.0	924
TIPO_VVDA_Propio	1178	2	1.0	665

Ilustración 4: variables binarias

Del gráfico de correlación se puede concluir lo siguiente:

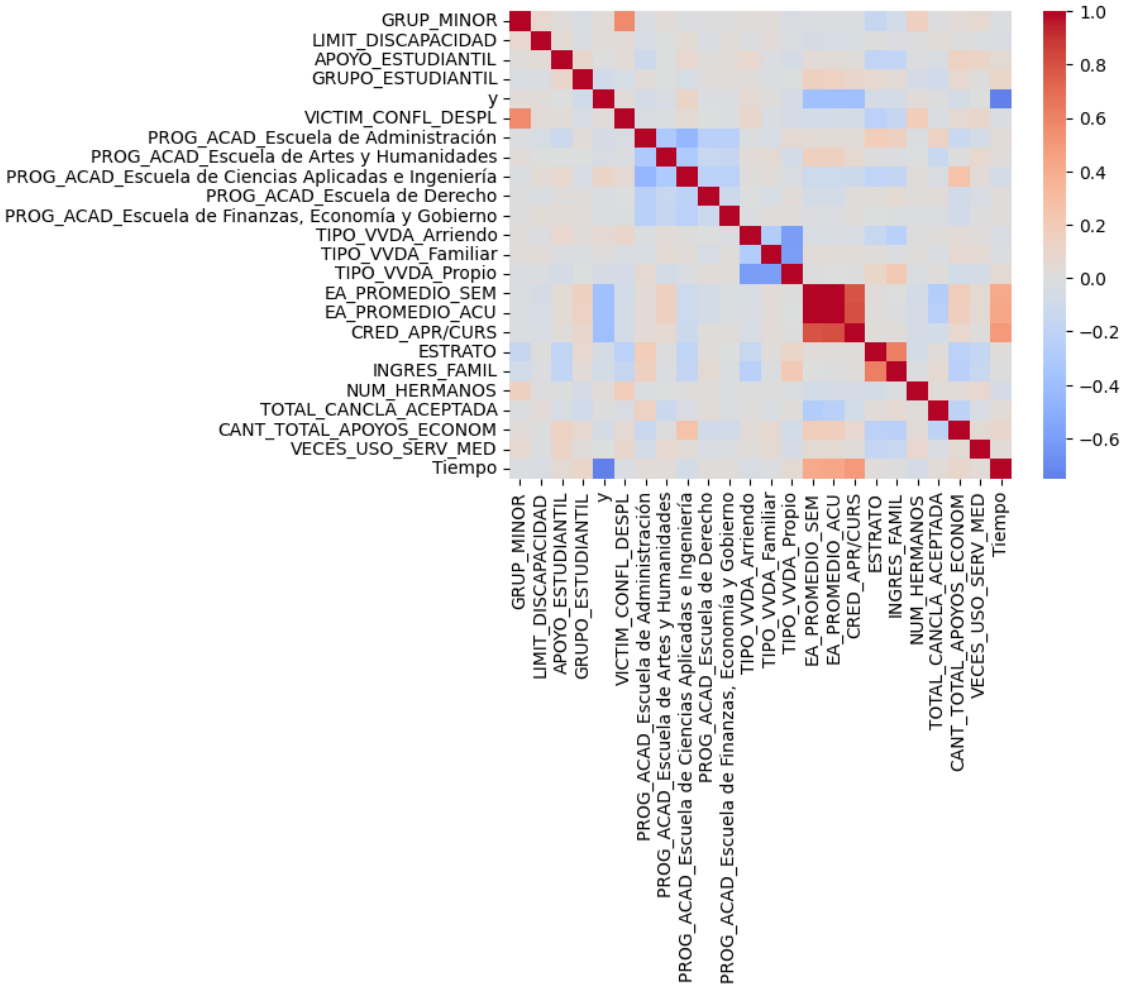

Ilustración 5: Mapa de calor de la matriz de correlación entre variables. Fuente: elaboración propia.

- Existen relaciones inversas moderadas entre variables del mismo tipo como los tipos de vivienda, los promedios y las escuelas. Esto es un indicio de multicolinealidad entre las variables. Puede ser útil retirar algunas para mejor el desempeño de algunos modelos.
- Hay relaciones que pueden resultar lógicas, como víctimas del conflicto y el desplazamiento.
- Existe una correlación alta entre las variables académicas: promedio acumulado, promedio semestral y relación créditos aprobados/cursados, sugiriendo redundancia.

- Se aprecia también una relación inversa entre el tiempo y el evento de deserción, lo cual es lógico debido a la disminución de desertores con el tiempo.

Es interesante revisar cómo el Factor de Inflación de Varianza puede ayudar a tomar una decisión referente a las variables que conviene reducir para corregir los efectos de la multicolinealidad. Para este caso, todas las variables relacionadas con los programas académicos (escuelas) y de tipo de vivienda:

	feature	VIF
0	GRUP_MINOR	1.532468
1	LIMIT_DISCAPACIDAD	1.024676
2	APOYO_ESTUDIANTIL	1.083081
3	GRUPO_ESTUDIANTIL	1.056439
4	VICTIM_CONFL_DESPL	1.591440
5	PROG_ACAD_Escuela de Administración	inf
6	PROG_ACAD_Escuela de Artes y Humanidades	inf
7	PROG_ACAD_Escuela de Ciencias Aplicadas e Inge...	inf
8	PROG_ACAD_Escuela de Derecho	inf
9	PROG_ACAD_Escuela de Finanzas, Economía y Gobi...	inf
10	TIPO_VVDA_Arriendo	inf
11	TIPO_VVDA_Familiar	inf
12	TIPO_VVDA_Propio	inf
13	EA_PROMEDIO_SEM	45.604214
14	EA_PROMEDIO_ACU	47.805105
15	CRED_APR/CURS	3.120478
16	ESTRATO	1.769811
17	INGRES_FAMIL	1.759694
18	NUM_HERMANOS	1.058717
19	TOTAL_CANCLA_ACEPTADA	1.214726
20	CANT_TOTAL_APOYOS_ECONOM	1.228447
21	VECES_USO_SERV_MED	1.063141

Ilustración 6: Resultados de la prueba del Factor de Inflación de Varianza - VIF.

MODELACIÓN

En el proceso de modelación e implementación de los diferentes algoritmos para resolver el problema de clasificación de este estudio, se aborda el desafío del

desbalance de clases utilizando métodos de sobremuestreo y submuestreo. Para cada modelo implementado, excepto en el análisis de supervivencia, se estiman los parámetros utilizando conjuntos de datos de entrenamiento (**70 % de los datos que corresponde a 825 muestras**) y prueba (**30% de los datos que corresponde a 353 muestras**) obtenidos del set original y ajustados mediante técnicas de sobremuestreo, que consiste en duplicar aleatoriamente las muestras de la clase minoritaria sin reemplazo para equilibrar el conjunto de datos, y de submuestreo, que, por el contrario, consiste en eliminar aleatoriamente muestras de la clase mayoritaria.

Regresión Logística

La regresión logística es un método clásico y efectivo para problemas de clasificación. En este estudio, se implementa la regresión logística estándar junto con las técnicas de regularización L1 (Lasso) y L2 (Ridge).

Para implementar este modelo, se parte de la función logística, que calcula la probabilidad de que *x* tome un valor específico, como *x* = 1 para representar la deserción, en función de los coeficientes de la regresión lineal. La función se expresa como:

$$p(x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

Los coeficientes β_0 y β_1 se estiman utilizando el método de máxima verosimilitud, que consiste en maximizar la función de verosimilitud, definida como:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i\prime:y_{i\prime}=0} \left(1 - p(x_{i\prime})\right)$$

El objetivo es encontrar los valores de β_0 y β_1 que, al sustituirlos en el modelo, produzcan una probabilidad cercana a 1 para los individuos que desertaron y cercana a 0 para los que no.

Este enfoque se puede extender fácilmente al caso multivariado, incrementando la cantidad de parámetros en función del número de predictores o variables.

Regularización L1 (Lasso)

Además de prevenir el sobreajuste, Lasso realiza selección automática de variables al reducir algunos coeficientes exactamente a cero. Esta propiedad es beneficiosa para abordar la multicolinealidad al eliminar variables altamente correlacionadas. La característica de las variables cuyo coeficiente se reduce a cero es precisamente que son las menos relevantes en el modelo.

Regularización L2 (Ridge)

Reduce los coeficientes del modelo sin eliminar variables, lo que ayuda a prevenir el sobreajuste mientras retiene todas las características que pueden ser importantes.

Árboles de Decisión

Los árboles de decisión son altamente interpretables y capaces de identificar patrones complejos en los datos, incluyendo interacciones no lineales entre variables. Son flexibles en la manipulación de diferentes tipos de datos, incluyendo variables categóricas, lo que los hace adecuados para una amplia gama de escenarios académicos.

Los árboles de decisión funcionan como diagramas de flujo, donde cada nodo representa una pregunta basada en una característica de los datos, como, por ejemplo, si un estudiante desertará o no. Cada rama del árbol corresponde a una posible respuesta a esa pregunta.

Para determinar la mejor manera de dividir los datos en un nodo, se puede utilizar la impureza de Gini, que mide la probabilidad de clasificar incorrectamente un elemento si se asigna aleatoriamente según la distribución de clases en ese nodo. Esta impureza se calcula con la fórmula:

$$gini = 1 - \sum_{i=1}^{k} p_i^2$$

Donde p_i es la proporción de elementos o datos que pertenecen a la clase i en ese nodo.

La decisión de hacia qué rama seguir se toma buscando minimizar la impureza de Gini. Si el valor de Gini es 0, significa que el nodo es puro y todos los elementos pertenecen a una única clase. Por el contrario, un valor elevado de Gini indica una alta impureza, lo que sugiere que las clases están mezcladas en ese nodo.

Random Forest

El Random Forest se destaca por su eficacia en conjuntos de datos con numerosas características y desequilibrios. Este modelo es resistente al sobreajuste y proporciona información sobre la importancia relativa de las características. Aunque la interpretación directa de los resultados puede ser limitada, es versátil y adaptable a diversas situaciones académicas.

Este modelo combina múltiples árboles de decisión independientes. Para su construcción, el conjunto de entrenamiento se divide en N subconjuntos de datos de igual tamaño mediante una técnica llamada *bagging*. A partir de cada subconjunto, se genera un árbol de decisión y, en cada nodo del árbol, se consideran de forma aleatoria solo algunas variables, en lugar de todas, para determinar la mejor decisión o clase.

Una vez entrenados los árboles, cada uno emite una predicción para un nuevo dato, y la predicción final se decide por votación mayoritaria entre todos los árboles. Matemáticamente, esto se expresa como:

$$\hat{y} = moda(\{h_1(x), h_2(x), h_3(x) \dots h_k(x)\}$$

donde $h_i(x)$ es la predicción del *i*-ésimo árbol para la entrada *x*, en un bosque compuesto por *k* árboles.

Método *Stepwise* de selección de características para regresión logística

Método iterativo para encontrar el mejor conjunto de características, añadiendo o eliminando características en cada iteración, el cual solo se aplica a la regresión logística.

- **Backward:** comienza con un modelo completo y elimina características una a una, evaluando el impacto en el rendimiento del modelo según el F1-score.
- **Forward:** comienza con un modelo nulo y agrega características una a una, seleccionando aquellas que mejoran significativamente el rendimiento según el F1-score.

Análisis de supervivencia

Los modelos de supervivencia, más que predecir el valor que una variable tomará dados unos datos de entrada, estiman la probabilidad de ocurrencia del evento. Inicialmente, se estima la función de supervivencia utilizando el estimador de Kaplan-Meier. Este estimador proporciona una manera no paramétrica de calcular la probabilidad de que un individuo sobreviva más allá de un cierto tiempo *t*, a partir de los tiempos de eventos observados y las censuras.

La función de supervivencia de Kaplan-Meier se calcula como:

$$S(t) = \prod_{t_i < t} \left(\frac{n_i - d_i}{n_i} \right)$$

donde t_i son los tiempos en los que ocurre el evento, d_i es el número de eventos en t_i, y n_i es el número de individuos en riesgo justo antes de t_i (James et al. 2023). Las censuras se refieren a los casos donde no se ha observado el evento de interés (deserción) durante el período de estudio. Esto puede ocurrir porque los estudiantes siguen matriculados al final del período de observación.

Luego, se estima el modelo de supervivencia utilizando el modelo de riesgos proporcionales de Cox. Este modelo evalúa el efecto de las covariables en el tiempo hasta la ocurrencia del evento, permitiendo identificar la significancia de las

variables independientes en la predicción de la deserción. Las censuras son cruciales en este análisis, ya que el modelo de Cox puede manejar datos censurados adecuadamente, diferenciando entre aquellos que han experimentado el evento y aquellos que no.

La significancia de las variables se identifica mediante pruebas de hipótesis sobre los coeficientes del modelo de Cox. Las predicciones de las curvas de supervivencia se generan para diferentes grupos de estudiantes y se comparan con las curvas de supervivencia observadas, para validar la precisión del modelo (James et al. 2023).

Finalmente, el modelo se evalúa utilizando el índice de concordancia (C-index), que mide la capacidad del modelo para discriminar entre individuos que experimentan el evento en diferentes tiempos. Un C-index cercano a 1 indica un buen poder predictivo del modelo, mientras que un valor cercano a 0.5 indica que es un modelo aleatorio o al azar.

RESULTADOS

Dado que estamos frente a un problema de desbalance de clases, aunque se presenten todas las métricas clásicas de evaluación de modelos de clasificación, es conveniente enfocarse primordialmente en el F1-Score. Es especialmente útil en este contexto porque combina la precisión y la exhaustividad (recall) en una sola métrica, proporcionando una medida más equilibrada del desempeño del modelo cuando las clases no están igualmente representadas. Esto es crucial para asegurar que el modelo no solo identifica correctamente los casos de deserción, sino que también minimiza los falsos positivos y falsos negativos.

En primer lugar, se entrenaron y evaluaron tres modelos principales: Regresión Logística, Árboles de Decisión y Random Forest, usando la base de datos consolidada, la cual presenta una estructura de corte transversal. Los resultados iniciales de estos modelos se resumen a continuación:

Modelo	Enfoque	Precisión	Exhaustividad (Recall)	F1 Score
Regresión Logística	Muestra original	80%	67%	69%
	Sobremuestreo	67%	72%	68%
	Submuestreo	64%	71%	65%
Lasso	Muestra original	75%	57%	59%
	Sobremuestreo	65%	71%	67%
	Submuestreo	61%	68%	61%
Ridge	Muestra original	72%	55%	56%
	Sobremuestreo	68%	72%	70%
	Submuestreo	63%	69%	64%
Arboles de Decisión	Normal	62%	60%	61%
Random Forest	Normal	78%	59%	61%

Tabla 6: Resumen de resultados de los modelos estimados.

Como se puede observar, aunque todos los modelos muestran un rendimiento aceptable en términos de precisión y exhaustividad, la Regresión Logística destacó con el F1-Score más alto, usando una regularización de Ridge y un enfoque de sobremuestreo, lo que indica un mejor equilibrio entre la capacidad del modelo para identificar casos de deserción y su precisión general. Además, la evaluación del F1-score en la regresión logística, tanto con o sin regularización, respalda la observación de que el conjunto de datos con sobremuestreo demuestra un rendimiento superior. Este hallazgo refuerza la efectividad de abordar el desbalance de clases mediante sobremuestreo en el contexto del problema de deserción.

El modelo de Ridge con sobremuestreo nos enseña que las siguientes son las características que inciden en la predicción:

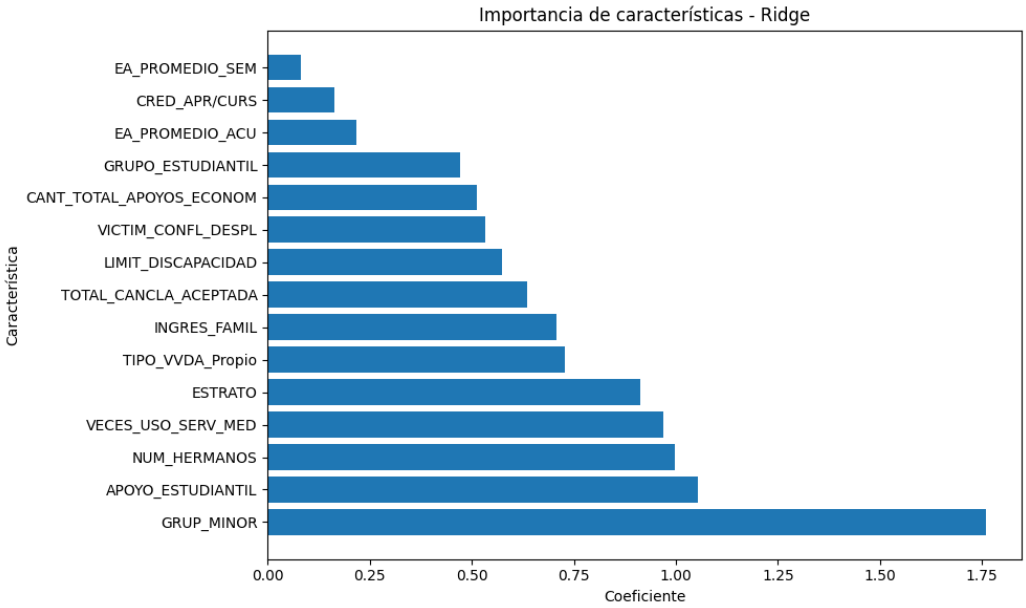

Ilustración 7: Relevancia de las variables para la predicción en el modelo Logístico.

Complementariamente se utilizó el método de Grid Search Cross Validation, el cual es un método de selección de modelos que busca identificar la mejor configuración de hiperparámetros mediante la evaluación del rendimiento del modelo utilizando validación cruzada. Este enfoque es crucial para prevenir el sobreajuste y proporciona una estimación más precisa del rendimiento del modelo en datos no observados. Al aplicar este método a los tres modelos utilizados, los resultados revelan que el mejor modelo, con hiperparámetros optimizados, sigue siendo la Regresión Logística, logrando un F1-score del 69 %. Este desempeño supera al Árbol de Decisión y al Random Forest, a pesar de que estos, en general, siempre presentan un desempeño mayor para el problema de la deserción:

Modelo	Precisión	Exhaustividad (Recall)	F1 Score
Regresión Logística	67%	73%	69%
Arboles de Decisión	71%	63%	65%
Random Forest	80%	61%	64%

Tabla 7: Resumen de los resultados de los modelos empleando GS Cross Validation

También se llevó a cabo un método de selección de características Stepwise para seleccionar la mejor configuración basada en el F1-Score para la Regresión

Logística. Aunque se puede observar que el mejor modelo seleccionado es el que usa el método forward-stepwise con sobremuestreo, las variables seleccionadas pueden llevar a obtener unas predicciones sesgadas y erróneas; por lo tanto, se opta por el segundo mejor modelo, el cual corresponde con el backward-stepwise con submuestreo, pues se identifican variables que pueden ser más adecuadas para el análisis del fenómeno:

Stepwise	Enfoque	F1-Score	Características seleccionadas
Backward	Muestra original	36.11%	Grupo minoritario, Discapacidad, Apoyo Estudiantil, Grupo Estudiantil, Víctima de Conflicto, Escuela de Administración, Escuela de Ciencias Aplicadas e Ingeniería, Tipo de vivienda de Arriendo y Propio, Promedio Primer Semestre, Créditos Aprobados/Cursados, Ingresos Familiares y Número de Hermanos
	Sobremuestreo	68.89%	Grupo Estudiantil y Promedio Primer Semestre
	Submuestreo	69.21%	Discapacidad, Gupo Estudiantil, Víctima de Conflicto, Escuela de Administración, Escuela de Artes y Humanidades, Escuela de Finanzas, Economía y Gobierno, Tipo de Vivienda Propia, Promedio Primer Semestre, Estrato, Número de Hermanos, Recibir Apoyo Económico, Servicio Médico
Forward	Muestra original	36.10%	Grupo Minoritario, Discapacidad, Apoyo Estudiantil, Víctima de Conflicto, Escuela de Administración, Escuela de Artes y Humanidades, Escuela de Finanzas, Economía y Gobierno, Tipo de Vivienda de arriendo y Familiar, Promedio Primer Semestre, Créditos Aprobados/Cursados, Estrato, Ingresos Familiares, Número de Hermanos y Apoyo Económico
	Sobremuestreo	70.09%	Grupo Estudiantil, Escuela de Artes y Humanidades, Escuela de Derecho y Promedio Primer Semestre
	Submuestreo	68.23%	Grupo Minoritario, Discapacidad, Grupo Estudiantil, Escuela de Ciencias Aplicadas e Ingeniería, Escuela de Derecho, Escuela de Finanzas, Economía y Gobierno, Tipo de Vivienda de Arriendo, Promedio Primer Semestre, Apoyo Económico, Servicio Médico

Tabla 8: Resultado de la selección de características con el método stepwise para el modelo Logístico.

Por último, se realizó el análisis de supervivencia para comprender mejor los factores asociados con el tiempo hasta la deserción y la forma como esta se comporta.

Para ello primero se construye la curva de supervivencia. En esta se aprecia que la probabilidad de deserción es particularmente alta desde el principio y desciende hasta cerca del 85 % en el cuarto semestre en general:

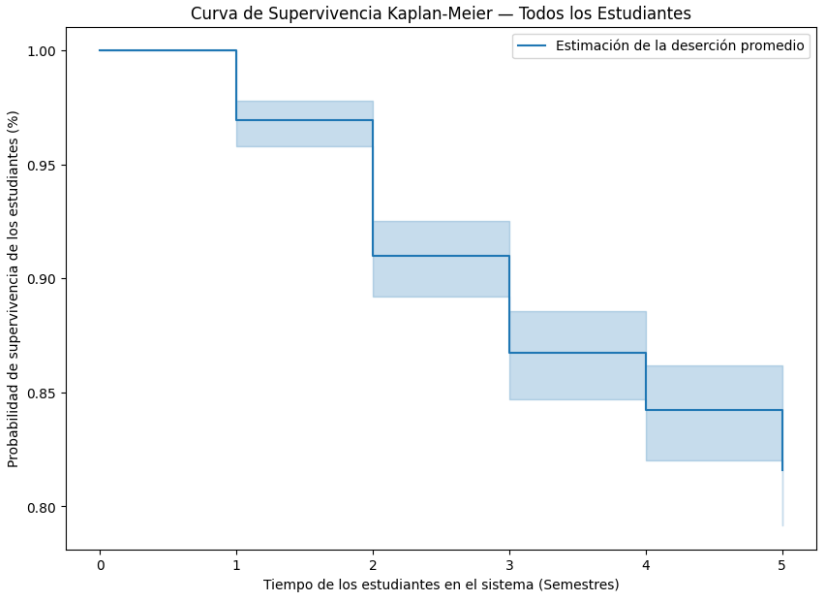

Ilustración 8: Curva de supervivencia de Kaplan-Meier. Fuente: elaboración propia.

Al estimar el modelo de riesgos proporcional de Cox, se aprecia que las variables con mayor efecto sobre la deserción según el modelo de supervivencia son Grupos Minoritarios, Víctimas del Conflicto, Tipo de Vivienda Propio, Relación de Créditos Aprobados/Cursados y Total Cancelaciones Aprobadas. En la siguiente tabla se muestra el resumen y, en las columnas *coef* y *exp(coef),* la magnitud del efecto de cada variable en el modelo de regresión estimado:

	coef	exp(coef)	se(coef)	coef lower 95%	coef upper 95%	exp(coef) lower 95%	exp(coef) upper 95%	cmp to	z	p	-log2(p)
GRUP_MINOR	1.64	5.17	0.69	0.30	2.99	1.35	19.84	0.00	2.39	0.02	5.91
LIMIT_DISCAPACIDAD	0.38	1.46	0.76	-1.11	1.86	0.33	6.45	0.00	0.49	0.62	0.69
APOYO_ESTUDIANTIL	0.18	1.19	0.20	-0.21	0.57	0.81	1.77	0.00	0.89	0.38	1.41
GRUPO_ESTUDIANTIL	-0.48	0.62	0.35	-1.16	0.21	0.31	1.23	0.00	-1.36	0.17	2.53
VICTIM_CONFL_DESPL	-1.29	0.28	0.50	-2.27	-0.30	0.10	0.74	0.00	-2.56	0.01	6.58
TIPO_VVDA_Propio	-0.32	0.72	0.16	-0.65	-0.00	0.52	1.00	0.00	-1.97	0.05	4.34
EA_PROMEDIO_SEM	-0.54	0.58	0.42	-1.37	0.29	0.25	1.33	0.00	-1.28	0.20	2.32
EA_PROMEDIO_ACU	0.08	1.09	0.42	-0.74	0.91	0.47	2.48	0.00	0.19	0.85	0.24
CRED_APR/CURS	-0.34	0.71	0.10	-0.53	-0.14	0.59	0.87	0.00	-3.40	<0.005	10.54
ESTRATO	-0.08	0.92	0.10	-0.29	0.12	0.75	1.13	0.00	-0.81	0.42	1.26
INGRES_FAMIL	-0.12	0.88	0.10	-0.32	0.08	0.72	1.08	0.00	-1.20	0.23	2.13
NUM_HERMANOS	-0.00	1.00	0.08	-0.15	0.15	0.86	1.16	0.00	-0.04	0.97	0.04
TOTAL_CANCLA_ACEPTADA	-0.20	0.82	0.08	-0.36	-0.04	0.70	0.96	0.00	-2.45	0.01	6.14
CANT_TOTAL_APOYOS_ECONOM	-0.16	0.85	0.10	-0.35	0.04	0.70	1.04	0.00	-1.59	0.11	3.17
VECES_USO_SERV_MED	-0.09	0.91	0.10	-0.28	0.10	0.75	1.11	0.00	-0.94	0.35	1.53

Ilustración 9: Resultados de la estimación del modelo de supervivencia. Fuente: elaboración propia.

Se presenta el efecto de los parámetros de forma gráfica para facilitar su comprensión:

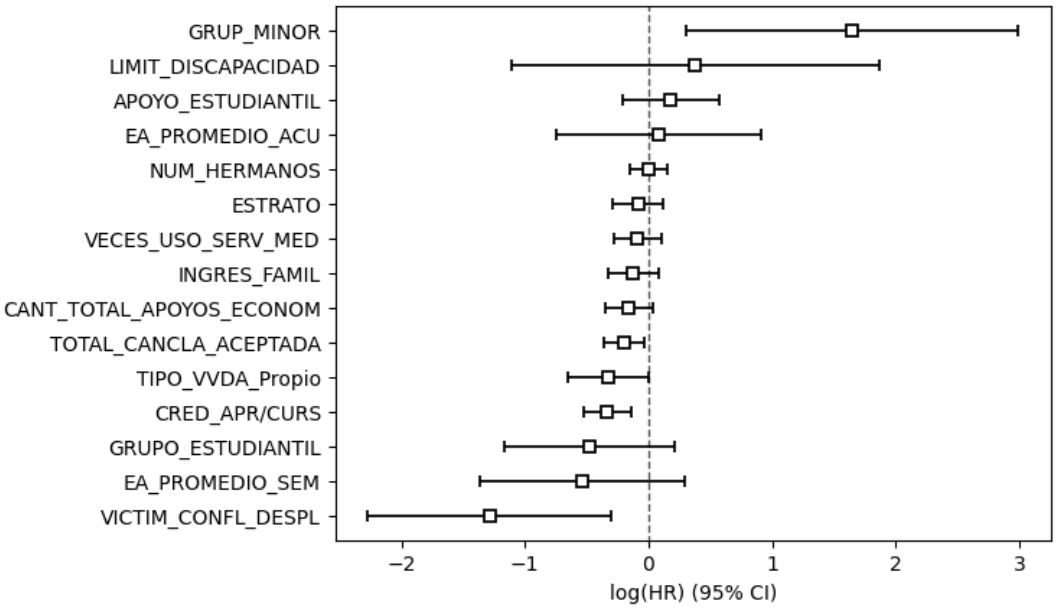

Ilustración 10: Visualización de la significancia de las variables para el modelo de supervivencia. Fuente: elaboración propia.

Esta estimación del modelo permite obtener una medida de la capacidad de predicción a través del índice de concordancia, C-index. Para este caso, se puede observar que el modelo tiene una capacidad potente de predicción pues el C-index se calcula en **0.769**, es decir, se comporta mucho mejor que un modelo aleatorio.

Por último, se presentan las curvas de supervivencia predichas para 10 de los individuos en el conjunto de prueba y se contrastan con la curva de supervivencia observada. Se observan comportamientos cercanos a la curva observada, la pendiente de cada curva predicha puede darnos una idea de si un individuo puntual tendrá mayor o menor probabilidad de permanencia (no deserción) en el corto plazo.

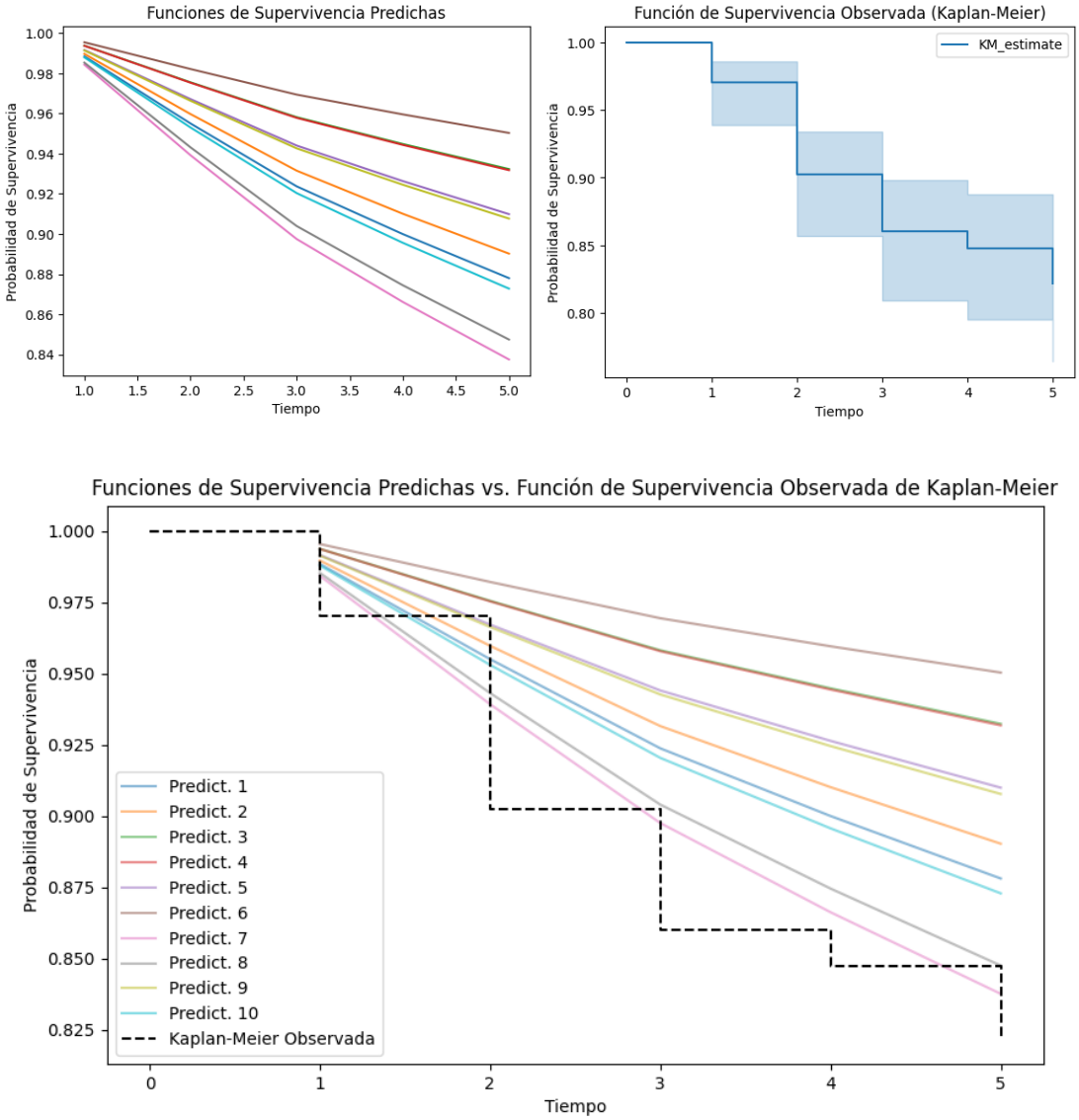

Ilustración 11: Comparación de las curvas de supervivencia predichas y observada. Fuente: elaboración propia.

CONCLUSIONES

A lo largo de esta investigación, se han logrado alcanzar los objetivos planteados inicialmente, obteniendo datos y hallazgos significativos que aportan al conocimiento y comprensión del problema de deserción estudiantil. Los puntos clave se resumen a continuación:

Eficacia de la Regresión Logística: a pesar de ser un modelo clásico y tradicional, la regresión logística sigue demostrando su potencia y efectividad para resolver problemas de clasificación binaria. Sorprendentemente, en muchos casos, supera a modelos más modernos como los basados en árboles de decisión, destacando su relevancia y utilidad en la práctica. Las variables que contribuyen a la predicción fueron, en primer lugar, Grupos Minoritarios, Apoyo Estudiantil, Número de Hermanos, Uso del Servicio Médico, Estrado, Vivienda Propia, Ingresos Familiares y Cancelaciones de Clases Aceptadas; de forma menos relevante, pero también incluidas, se encuentran Limitación de Discapacidad, Víctima del Conflicto, Apoyos Económicos, Grupo Estudiantil, Promedio Acumulado, Relación de Créditos Aprobados/Cursados y Promedio Semestral.

Manejo del desbalance de clases: en el contexto del problema de deserción, el desbalance de clases puede corregirse de manera efectiva utilizando métodos de sobremuestreo. Este enfoque ha mostrado una mejora en el desempeño global del modelo. Sin embargo, es crucial explorar métodos más novedosos para tratar el desbalance de clases y evaluar los modelos en estas circunstancias, garantizando una robustez mayor en los resultados obtenidos.

Variables relevantes en el análisis de deserción: abordar el problema de la deserción de manera integral requiere considerar una variedad de variables académicas, personales, sociodemográficas, entre otras. Durante este estudio, se observó que las variables sociodemográficas, más que las académicas, desempeñan un papel más relevante en la predicción de la deserción. Esto resalta

la necesidad de considerar un enfoque multidimensional para capturar las diversas facetas del problema.

Importancia de los modelos de supervivencia: Los modelos de supervivencia han demostrado ser herramientas valiosas, ya que proporcionan información adicional crucial al capturar la probabilidad de no deserción (supervivencia) a lo largo del tiempo. Desde la perspectiva de esta investigación, resulta más valioso conocer las probabilidades de deserción de un estudiante en un tiempo t específico, en lugar de simplemente determinar si desertará o no con cierta incertidumbre, como ocurre con los modelos de clasificación. Esta perspectiva destaca la importancia y relevancia de implementar modelos de supervivencia para abordar de manera más efectiva el problema de la deserción estudiantil. Este modelo, a diferencia de la regresión logística, consideró que la variable de Relación Créditos Aprobados/Cursados sí era relevante para el cálculo de la probabilidad de no deserción; además, el modelo de deserción también incluye como relevantes a las variables Grupos Minoritarios, Víctima del Conflicto, Vivienda Propia y Cancelaciones de Clase Aceptadas.

En resumen, esta investigación subraya la vigencia y eficacia de los modelos tradicionales como la regresión logística, la necesidad de métodos avanzados para tratar el desbalance de clases y la importancia de considerar múltiples variables y enfoques para comprender y predecir la deserción estudiantil de manera integral.

REFERENCIAS

Acevedo, F. (2021). Concepts and measurement of dropout in higher education: A critical perspective from Latin America. *Issues in Educational Research, 31*(3), 661-678.

Ali, D. A., & Hussein, A. M. (2024). Analysis of Cox proportional hazard model for dropout students in university: Case study from SIMAD University. *Journal of Applied Research in Higher Education, 16*(3), 820–830. https://doi.org/10.1108/JARHE-03-2023-0103

Arias, A., Linares-Vásquez, M., & Héndez-Puerto, N. R. (2024). Undergraduate Dropout in Colombia: A systematic literature review of causes and solutions. *Journal of Latinos and Education, 23*(2), 612–627. https://doi.org/10.1080/15348431.2023.2171042

Buenaño, E., Beletanga, M. J., & Mancheno, M. (2023). What factors are relevant to understanding dropout? Analysis at a co-financed university in Ecuador and policy implications, using survival Cox models. *Journal of Latinos and Education*. https://doi.org/10.1080/15348431.2023.2271570

Cvetkovski, S., Jorm, A. F., & Mackinnon, A. J. (2018). Student psychological distress and degree dropout or completion: A discrete-time, competing risks survival analysis. *Higher Education Research and Development, 37*(3), 484–498. https://doi.org/10.1080/07294360.2017.1404557

Díaz Peralta, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios Pedagógicos (Valdivia), 34*(2), 65-86.

Fernández-Martín, T., Solís-Salazar, M., Hernández-Jiménez, M. T., & Moreira-Mora, T. E. (2018). A multinomial and predictive analysis of factors associated with university dropout. *Revista Electrónica Educare, 23*(1). https://doi.org/10.15359/ree.23-1.5

Flores, V., Heras, S., & Julian, V. (2022). Comparison of predictive models with balanced classes using the SMOTE method for the forecast of student dropout in higher education. *Electronics, 11*(3), 457. https://doi.org/10.3390/electronics11030457

González, L. E. (2005). Estudio sobre la repitencia y deserción en la educación superior chilena. Digital Observatory for Higher Education in Latin America and The Caribbean. IESALC-UNESCO.

Hoyos Osorio, J. K., & Daza Santacoloma, G. (2023). Predictive model to identify college students with high dropout rates. *Revista Electrónica de Investigación Educativa, 25*, 1–10. https://doi.org/10.24320/redie.2023.25.e13.5398

James, G., Witten, D., Hastie, T., Tibshirani, R., & Taylor, J. (2023). *An introduction to statistical learning*. Springer International Publishing. https://doi.org/10.1007/978-3-031-38747-0

Lopes, R., Ribeiro, G., Lisboa, L. S., Silva, J. L. P. da, & Taconeli, C. A. (2023). Fatores associados à evasão de calouros no ensino superior: Um estudo com dados da Universidade Federal do Recôncavo da Bahia. *Revista Brasileira de Educação, 28*. https://www.redalyc.org/articulo.oa?id=27574386015

Norambuena, J. M., Badilla-Quintana, M. G., & Angulo, Y. L. (2022). Modelos predictivos basados en uso de analíticas de aprendizaje en educación superior: Una revisión sistemática. *Texto Livre, 15*, e36310.

Phan, M., De Caigny, A., & Coussement, K. (2023). A decision support framework to incorporate textual data for early student dropout prediction in higher education. *Decision Support Systems, 168*, 113940. https://doi.org/10.1016/j.dss.2023.113940

Rodríguez Velasco, C. L., García Villena, E., Brito Ballester, J., Durántez Prados, F. Á., Silva Alvarado, E., & Crespo Álvarez, J. (2023). Forecasting of post-graduate students' late dropout based on the optimal probability threshold adjustment technique for imbalanced data. *International Journal of Emerging Technologies in Learning (IJET), 18*(04), 120–155. https://doi.org/10.3991/ijet.v18i04.34825

Romito, M., Pilutti, S., & Contini, D. (2020). Why do students leave university? Qualitative research at an Italian higher education institution. *European Journal of Education, 55*(3), 456-470.

Silva, P. T. de F. e., & Sampaio, L. M. B. (2023). Does student aid make a degree more likely? Evidence of the Permanence Scholarship Program from survival models. *International Journal of Educational Development, 96*. https://doi.org/10.1016/j.ijedudev.2022.102697

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*(1), 64-85.

Talamás-Carvajal, J. A., & Ceballos, H. G. (2023). A stacking ensemble machine learning method for early identification of students at risk of dropout. *Education and Information Technologies, 28*(9). https://doi.org/10.1007/s10639-023-11682-z

Tinto, V. (1989). Definir la deserción: Una cuestión de perspectiva. *Revista de Educación Superior, 71*(18), 1-9.

Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research.

Villarreal-Torres, H., Ángeles-Morales, J., Marín-Rodriguez, W., Andrade-Girón, D., Cano-Mejía, J., Mejía-Murillo, C., ... & Palomino-Márquez, M. (2023). Classification model for student dropouts using machine learning: A case study. *EAI Endorsed Transactions on Scalable Information Systems*.