

7. Factores del abandono universitario: Análisis con LightGBM y la Teoría de juegos cooperativos de Shapley

College Dropout Factors: Analysis with LightGBM and Shapley's Cooperative Game Theory

Hugo Roger Paz* @ 

Universidad Nacional de Tucumán, Tucumán, Argentina

RESUMEN

Un tema preocupante en la Carrera de Ingeniería Civil de la Facultad de Ciencias Exactas y Tecnología de la Universidad Nacional de Tucumán (FACET-UNT) es el abandono estudiantil. En el periodo que va desde 2005 hasta 2019, un total de 900 alumnos dejaron la carrera, de un total de 1615 ingresantes. El presente estudio se basó en el análisis de datos de historias académicas de estudiantes de ingeniería civil en la FACET-UNT. Nuestro objetivo principal fue determinar las variables de rendimiento académico que tienen un impacto significativo en la deserción de la carrera. Para ello implementamos un modelo de correlación utilizando LightGBM (Barbier et al. 2016; Ke et al. 2017; Shi et al. 2022). Utilizamos este modelo para identificar las variables claves que influyen en la probabilidad de deserción de los estudiantes. Además, empleamos la teoría de juegos para interpretar los resultados obtenidos. Específicamente, utilizamos la biblioteca SHAP (Lundberg et al., 2018, 2020; Lundberg & Lee, 2017) en Python para calcular los números de Shapley. Los resultados de nuestro estudio revelaron las variables más importantes que influyen en el abandono. Se identificaron diferencias significativas en términos de edad, tiempo transcurrido en los estudios y rendimiento académico, que incluye la cantidad de cursos aprobados y el número de exámenes realizados. Estos resultados pueden ser de utilidad para desarrollar estrategias más efectivas de retención estudiantil y mejorar el éxito académico en esta disciplina.

Palabras clave: Deserción escolar; Procesamiento de datos; Teoría de los juegos

College Dropout Factors: Analysis with LightGBM and Shapley's Cooperative Game Theory

ABSTRACT

A concerning issue in the Civil Engineering program at the Faculty of Exact Sciences and Technology of the National University of Tucumán is student attrition. From 2005 to 2019, a total of 900 students dropped out of the program out of 1615 initial enrollees. This study was based on data analysis of academic histories of civil engineering students at FACET-UNT. Our main objective was to determine the academic performance variables that have a significant impact on the dropout of the career. To do this, we implemented a correlation model using LightGBM (Barbier et al., 2016; Ke et al., 2017; Shi et al., 2022). We use this model to identify the key variables that influence the probability of student dropout. In addition, we use game theory to interpret the results obtained. Specifically, we use the SHAP library (Lundberg et al., 2018, 2020; Lundberg & Lee, 2017) in Python to calculate the Shapley numbers. The results of our study revealed the most important variables that influence the dropout from the civil engineering career. Significant differences were identified in terms of age, time spent in studies, and academic performance, which includes the number of courses passed and the number of exams taken. These results may be useful to develop more effective student retention strategies and improve ACADEMIC success in this discipline.

Keywords: Academic failure; Data processing; Game theory

Fatores de evasão universitária: análise com LightGBM e teoria dos jogos cooperativos de Shapley

RESUMO

Uma questão preocupante na Carreira de Engenharia Civil da Faculdade de Ciências Exatas e Tecnologia da Universidade Nacional de Tucumán é a evasão estudantil. No período de 2005 a 2019, um total de 900 alunos abandonaram o curso, de um total de 1615 ingressantes. O presente estudo baseou-se na análise de dados de históricos acadêmicos de estudantes de engenharia civil da FACET-UNT. Nosso principal objetivo foi determinar as variáveis de desempenho acadêmico que impactam significativamente na evasão do programa. Para isso, implementamos um modelo de correlação utilizando LightGBM (Barbier et al., 2016; Ke et al., 2017; Shi et al., 2022). Utilizamos este modelo para identificar as principais variáveis que influenciam a probabilidade de abandono escolar. Além disso, utilizamos a teoria dos jogos para interpretar os resultados obtidos. Especificamente, usamos a biblioteca SHAP (Lundberg et al., 2018,

2020; Lundberg & Lee, 2017) em Python para calcular números de Shapley. Os resultados do nosso estudo revelaram as variáveis mais importantes que influenciam a evasão da carreira de engenharia civil. Foram identificadas diferenças significativas em termos de idade, tempo de estudo e desempenho acadêmico, incluindo o número de cursos aprovados e o número de exames realizados. Esses resultados podem ser úteis para desenvolver estratégias mais eficazes de retenção de alunos e melhorar o sucesso acadêmico nesta disciplina.

Palavras-chave: Evasão escolar; Processamento de dados; Teoria do jogo

Facteurs d'abandon universitaire : analyse avec LightGBM et la théorie des jeux coopératifs de Shapley

RÉSUMÉ

Un sujet préoccupant dans la filière du Génie Civil de la Faculté des Sciences Exactes et de la Technologie de l'Université Nationale de Tucumán est l'abandon des études par les étudiants. Sur la période allant de 2005 à 2019, un total de 900 étudiants ont abandonné leurs études, sur un total de 1615 inscrits. La présente étude s'appuie sur l'analyse des données issues des parcours académiques des étudiants en génie civil de la FACET-UNT. Notre objectif principal était de déterminer les variables de performance académique qui ont un impact significatif sur l'abandon du programme. Pour ce faire, nous avons mis en œuvre un modèle de corrélation utilisant LightGBM (Barbier et al., 2016 ; Ke et al., 2017 ; Shi et al., 2022). Nous utilisons ce modèle pour identifier les variables clés qui influencent la probabilité de décrochage scolaire. De plus, nous utilisons la théorie des jeux pour interpréter les résultats obtenus. Plus précisément, nous avons utilisé la bibliothèque SHAP (Lundberg et al. 2018; Lundberg y Lee 2017; Lundberg et al. 2020). Les résultats de notre étude ont révélé les variables les plus importantes qui influencent l'abandon de la carrière en génie civil. Des différences significatives ont été identifiées en termes d'âge, de temps passé à étudier et de résultats scolaires, notamment le nombre de cours réussis et le nombre d'examens passés. Ces résultats peuvent être utiles pour développer des stratégies de rétention des étudiants plus efficaces et améliorer la réussite scolaire dans cette discipline.

Mots clés: Décrochage scolaire ; Abandon en cours d'études; Théorie des jeux

1. INTRODUCCIÓN

El abandono estudiantil representa un tema de gran preocupación en la Carrera de Ingeniería Civil de la Facultad de Ciencias Exactas y Tecnología (FACET)

de la Universidad Nacional de Tucumán (UNT), Argentina. Desde la implementación del Plan de Estudios de Ingeniería Civil en 2005, hasta el año 2020, solo 145 estudiantes han completado sus estudios, lo que se traduce en un promedio de aproximadamente 9 egresados por año. Durante este mismo período, se registraron 1615 ingresos a la carrera. Aunque esta cifra incluye universitarios que aún están cursando, queda claro que menos del 10% de los ingresantes finalizan sus estudios, un porcentaje alarmantemente bajo. El resto de los estudiantes corresponde a aquellos que abandonan o tienen altas probabilidades de hacerlo. Comprender en profundidad las causas de esta situación podría permitirnos desarrollar propuestas de mejora para reducir el abandono.

El presente estudio se centró en el análisis de datos de historias académicas de estudiantes de ingeniería civil en la FACET-UNT, con el objetivo de identificar las variables de rendimiento académico que tienen un impacto significativo en la deserción. Para ello, se implementó un modelo de correlación utilizando LightGBM, siguiendo las metodologías propuestas por Barbier et al. (2016), Ke et al. (2017) y Shi et al. (2022). Este modelo nos permitió identificar las variables clave que influyen en la probabilidad de abandono de los estudiantes.

Además, se empleó la teoría de juegos para interpretar los resultados obtenidos. Específicamente, se utilizó la biblioteca SHAP (Lundberg et al., 2018, 2020; Lundberg & Lee, 2017) en Python para calcular los valores de Shapley, proporcionando una interpretación detallada de la contribución de cada variable a la deserción estudiantil.

2. ESTUDIO DE ANTECEDENTES Y PRECISIONES CONCEPTUALES

La deserción académica en instituciones de educación superior representa un desafío crucial en la búsqueda de una educación de calidad y equidad. Este fenómeno no solo tiene un impacto significativo en la vida de los estudiantes, sino que también plantea preocupaciones importantes para las instituciones educativas y la sociedad en general (Cabrera et al. 2006; Tinto 1975).

En la última década, se han redoblado los esfuerzos para investigar la causa del problema en su contexto. Algunos autores señalan factores económicos, institucionales y sociales como desencadenantes (Abdala Leiva, Castiglione, y Infante 2008; Arancibia et al. 2013; Arce et al. 2017). Otros autores han identificado factores sociofamiliares como enfermedades personales o familiares,

lejanía de la vida universitaria en función de las becas o la necesidad de integración en el mercado laboral (Gonzalez-Ramirez y Pedraza-Navarro 2017).

Se identificaron factores relacionados con la adaptación profesional, social y académica, la dedicación y el uso de tecnologías de aprendizaje (Bernardo et al. 2016; Castillo-Sánchez et al., 2020; Chacín et al., 2020; Da Re y Clerici 2017; Lázaro Alvarez et al., 2020). Asimismo se investigó la relación entre la satisfacción, el rendimiento académico y la probabilidad de abandonar los estudios (De Coninck et al., 2019; Ponce, 2021; Tomás y Gutiérrez, 2019).

Fonseca y García analizan el desgranamiento desde la teoría organizacional, critican la tendencia de la investigación basadas sólo en teorías de la adaptabilidad de los estudiantes y revelan la necesidad de incluir recomendaciones para que las instituciones se adapten a los nuevos grupos de estudiantes. También piden la inclusión de estudios cualitativos y mixtos en los análisis de deserción, ya que esta es un fenómeno complejo y multifacético (Fonseca y García 2016). Estos elementos se han considerado y tienen mucha importancia.

En otros estudios, la formación previa en matemáticas relacionada con los puntajes de las pruebas de ingreso y el rendimiento académico del primer año, particularmente en matemáticas introductorias e informática, se ha identificado como predictor de deserción (Kori et al., 2018; Niitsoo et al. 2014).

De igual forma, se analizaron variables no evaluadas en los estudios anteriores, tales como conocimientos teóricos y prácticos de informática (Araque et al., 2009; Xenos et al., 2002) y desarrollo del pensamiento lógico y analítico (Salar-Fernández et al. 2019; Mitic et al. 2021). Si bien se examinaron variables socioeconómicas, ninguno de estos estudios utilizó un enfoque económico.

Con relación a las técnicas de investigación utilizadas para analizar los procesos de desgranamiento y abandono, la utilización de bases de datos para la gestión de información relativa al desarrollo del curriculum (historias académicas de alumnos, gestión de aulas virtuales, datos de cursadas en cada asignatura, entre otros aspectos) ha promovido el desarrollo de herramientas de análisis de datos que actualmente se categorizan como Educational Data Mining (EDM) y Learning Analytics (LA).

Esta última puede definirse como la medición, recolección, análisis y reporte de datos sobre los alumnos y sus contextos, con el fin de comprender y optimizar el aprendizaje y los entornos en los que se desarrolla. (Lang et al. 2017).

Existen muchos estudios que analizan y describen estas técnicas (Armatas y Spratt 2019; Gagliardi et al., 2018; Hilliger et al. 2019; Ozdemir et al., 2019; Romero y Ventura 2020; Volungeviciene et al. 2019). La mayor parte de la investigación que aplica LA en la educación superior se ha centrado en el estudio de la retención y el abandono (Siemens et al., 2014). Con relación a la aplicación de técnicas específicas al estudio de esta problemática, Wan Yaacob et al., (2020) proponen técnicas de minería de datos para comprender qué estudiantes están en riesgo de abandonar y cuáles son los factores que contribuyen a mayores tasas de deserción.

Namoun y Alshanqiti realizaron una revisión sistemática de la literatura con relación a la predicción del rendimiento de los estudiantes mediante la minería de datos y técnicas de análisis de aprendizaje. El trabajo sintetiza los modelos y paradigmas aplicados de la minería de datos aplicados en la educación para predecir el desempeño de los estudiantes; también, identifica varios desafíos y proporciona recomendaciones para futuras investigaciones en el campo de la minería de datos educativos (Namoun y Alshanqiti, 2021).

Alban y Mauricio abordaron la literatura sobre la predicción de la deserción estudiantil universitaria como objeto de estudio, mediante técnicas de minería de datos. Encontraron que este objeto de estudio es un problema que afecta a universidades de todo el mundo, con consecuencias tales como reducción de matrícula, de ingresos para la universidad y pérdidas financieras para el Estado que financia los estudios, y además constituye un problema social para los estudiantes, sus familias y la sociedad en general. Es por eso que concluyen relevando la importancia de la predicción de este fenómeno sobre deserción universitaria; esto es, identificar con anticipación a los estudiantes desertores, para diseñar estrategias que proyecten soluciones al problema (Alban y Mauricio 2019). Consideramos significativo destacar como conclusiones de este estudio los principales factores que causan el abandono universitario:

1. Factores económicos, institucionales y sociales.
2. Factores sociofamiliares como enfermedades personales o familiares, lejanía de la vida universitaria debido a becas, y la necesidad de integrarse al mercado laboral.
3. Factores relacionados con la adaptación profesional, social y académica, la dedicación y el uso de tecnologías de aprendizaje.

4. Relación entre la satisfacción, el rendimiento académico y la probabilidad de abandonar los estudios.

En el presente artículo se analizan factores principalmente académicos y relacionados con el desempeño y progreso de los estudiantes en sus estudios, identificando patrones que influyen en la decisión de abandono. Asimismo, se analiza la variable tiempo de permanencia en la universidad y cómo influye este factor en la deserción.

3. MÉTODO

3.1. Población y Muestra

Los datos objeto de análisis provienen de las trayectorias académicas de 1615 estudiantes matriculados en la carrera de Ingeniería Civil, abarcando un espectro diverso de avances en la carrera, desde estudiantes de primer año hasta aquellos en etapas cercanas a la graduación, así como individuos que han discontinuado su formación y aquellos que han culminado su proceso educativo. Estos datos se obtuvieron a través del Sistema de Gestión Administrativa y de Estudiantes (SIGEA), que recopila y centraliza información relevante sobre el progreso académico y la gestión de los estudiantes en el programa.

En el marco de esta investigación, se ha llevado a cabo un exhaustivo examen del rendimiento académico de estudiantes universitarios a lo largo de un período de catorce años, desde 2004 hasta 2019. Esta delimitación temporal se justifica en base a la implementación de un nuevo plan de estudios para la carrera de ingeniería civil, el cual modifica completamente el currículo anterior, impidiendo el análisis coherente de datos de historias académicas anteriores a dicha fecha. Además, es significativo destacar que se decidió detener el análisis en el año 2019 por dos razones fundamentales. En primer lugar, la pandemia de COVID-19 plantea la posibilidad de que se hayan producido variaciones en el rendimiento académico producto de este hecho. En segundo lugar, se realizó un cambio en la base de datos del estudio a partir del año 2020, lo que justifica la conclusión del período de análisis en 2019 para mantener la coherencia metodológica.

Con el propósito de asegurar la precisión de los resultados, se efectuó una exclusión de aquellos estudiantes que habían obtenido la aprobación de asignaturas mediante un sistema de equivalencias, debido a modificaciones en

su plan de estudios, trayectoria académica o institución educativa. Dicha medida se implementó considerando que estos alumnos podrían haber ingresado al programa con asignaturas ya aprobadas, lo cual podría impactar en su progreso y desempeño en comparación con sus pares. Para el análisis, se procedió a seleccionar un conjunto de 1343 historias académicas de un total de 1615.

En el presente contexto, se clasifica como «egresados» a aquellos estudiantes que han finalizado sus estudios dentro del período bajo evaluación (2005-2019). Por su parte, se considera «permanentes» a los estudiantes que han mantenido su inscripción en el programa académico hasta el año 2019 y aún no han concluido su formación.

En última instancia, se define la situación de «deserción» para aquellos alumnos que no registran ninguna inscripción en el año 2020 y no han culminado su proceso educativo. Al adoptar este enfoque, seguimos la propuesta de González Fiegehen (2007) para los alumnos que se matricularon en 2019, representando únicamente el 5% del total analizado (Paz, 2022). Para el resto de los estudiantes, se emplea un criterio menos riguroso, incluyendo a aquellos que retoman su inscripción en la carrera y reanudan sus estudios tras un período de inactividad. El cálculo de los lapsos temporales se ha basado en la diferencia entre la última actividad documentada en el sistema SIGEA y la fecha de ingreso al programa académico.

El programa de estudios de cinco años y medio, distribuido en un total de once semestres, sigue un Sistema de Correlativas que establece condiciones necesarias para poder inscribirse en los cursos. Estas condiciones pueden ser «regularizadas», lo que significa que se han superado evaluaciones parciales o tareas prácticas, o «aprobadas», lo que indica que se ha completado satisfactoriamente el examen final del curso, garantizando así su aprobación.

Para el análisis se utilizaron las historias académicas (ver apartado 2.1.), de las cuales se extrajeron los siguientes datos para 1343 estudiantes:

- Género
- Edad Última Actividad
- Tiempo Facultad
- Abandono
- Promedio Notas
- Número Total Cursadas
- Número Regulares

- Número Recursadas
- Número Libres
- Número de Exámenes
- Número Promociones
- Número Aprobados
- Número de Reprobados
- Número de Ausentes

3.2. Técnica de análisis: LightGBM

La elección de las herramientas y metodologías adecuadas en la investigación es un proceso crucial que impacta directamente en la calidad y eficacia de los resultados obtenidos. En el caso de este estudio, la selección de LightGBM como algoritmo de aprendizaje automático y la integración de la teoría de juegos se basa en varias razones fundamentales:

- **Eficiencia Computacional:** LightGBM es ampliamente reconocido en la comunidad de aprendizaje automático por su eficiencia computacional y su capacidad para manejar conjuntos de datos grandes y complejos. Su implementación de crecimiento de árboles por hojas (leaf-wise) permite un proceso de entrenamiento más rápido en comparación con otros algoritmos basados en árboles, lo que resulta especialmente beneficioso al analizar un gran volumen de datos de historias académicas de estudiantes.
- **Alto Rendimiento:** LightGBM está diseñado para proporcionar un alto rendimiento en términos de precisión de predicción. Esto es esencial al abordar problemas de clasificación y predicción en el contexto de la deserción académica, donde se busca identificar patrones y factores determinantes con la mayor precisión posible.
- **Interpretabilidad:** A pesar de su complejidad, LightGBM ofrece herramientas y métodos para interpretar y comprender los resultados del modelo. Esto es crucial para nuestro objetivo de identificar las variables de rendimiento académico más influyentes en la deserción estudiantil.
- **Teoría de Juegos:** La incorporación de la teoría de juegos en el análisis añade un enfoque analítico adicional para entender las dinámicas detrás de las decisiones de los estudiantes en cuanto a la deserción. Esta teoría

nos permite considerar cómo las acciones de un estudiante pueden verse influenciadas por las de otros, lo que puede tener implicaciones significativas en la toma de decisiones académicas.

- **Enfoque Holístico:** La combinación de LightGBM y la teoría de juegos nos brinda un enfoque holístico para abordar el problema de la deserción académica. Al utilizar un modelo de aprendizaje automático avanzado junto con la teoría de juegos, estamos mejor posicionados para explorar y comprender las interacciones complejas entre las variables de rendimiento académico y los factores de decisión de los estudiantes.

A continuación, se detalla el proceso, incluyendo los parámetros utilizados en la aplicación del modelo:

División de los Datos:

Se inició el proceso de modelado dividiendo los datos en conjuntos de entrenamiento y testeo. Para ello, se utilizó la función `train_test_split` de la biblioteca scikit-learn (Pedregosa et al. 2011), con una proporción del 70% de los datos para entrenamiento y un 30% para prueba. Esta división permite evaluar el rendimiento del modelo en datos no utilizados durante el entrenamiento y garantizar su capacidad de generalización.

Creación de los Conjuntos de Datos LightGBM:

Luego, se crearon los conjuntos de datos específicos de LightGBM, denominados `d_train` y `d_test`, utilizando la clase `lgb.Dataset`. Estos conjuntos se construyeron a partir de los datos de entrenamiento y prueba previamente divididos.

Definición de Parámetros del Modelo:

Los parámetros del modelo LightGBM se definieron cuidadosamente para lograr un equilibrio entre rendimiento y regularización. Los parámetros utilizados son los siguientes:

- `max_bin`: 512
- `learning_rate`: 0.05
- `boosting_type`: «gbdt»
- `objective`: «binary»
- `metric`: «binary_logloss»
- `num_leaves`: 10
- `verbose`: -1

- ``min_data``: 100
- ``boost_from_average``: True

Estos parámetros se configuran de acuerdo con las recomendaciones de LightGBM y con base en la experiencia previa en la selección de hiperparámetros. El objetivo es optimizar la precisión del modelo y controlar la complejidad.

Entrenamiento del Modelo:

Finalmente, el modelo LightGBM se entrenó utilizando la función ``lgb.train``. Se especificaron los conjuntos de datos de entrenamiento y se indicó el número de iteraciones (en este caso, 10,000). Además, se evaluó el rendimiento en el conjunto de prueba mediante la opción ``valid_sets=[d_test]``.

3.3. Técnica de análisis: Teoría de Juegos Cooperativos de Shapley

En este estudio, hemos incorporado la teoría de juegos como un marco analítico valioso para comprender las dinámicas detrás de las decisiones de los estudiantes en relación con la deserción académica. Antes de profundizar en la explicación acerca de cómo se aplicó la teoría de juegos en nuestro análisis, es trascendental presentar los conceptos clave y herramientas utilizadas en esta metodología.

La teoría de juegos es un campo de estudio que se centra en el análisis de las decisiones estratégicas tomadas por los agentes racionales cuando sus acciones afectan tanto sus resultados como los resultados de otros agentes. En el contexto de la educación superior, esta teoría puede utilizarse para modelar cómo las decisiones individuales de los estudiantes, como la decisión de abandonar o continuar con sus estudios, pueden estar influenciadas por las decisiones de otros estudiantes y factores contextuales.

Los números Shapley son una herramienta fundamental en la teoría de juegos cooperativos que se utiliza para asignar un valor numérico a la contribución relativa de cada jugador en un juego cooperativo. En el contexto de nuestro análisis, los números Shapley se utilizan para comprender cuánto contribuye cada variable (por ejemplo, el rendimiento académico) en la toma de decisiones de los estudiantes con respecto a la deserción académica.

Para incorporar la teoría de juegos en nuestro análisis, aplicamos la biblioteca SHAP (SHapley Additive exPlanations), que nos permite calcular los números Shapley y visualizar la contribución de las variables al modelo LightGBM.

4. PRESENTACION DE RESULTADOS Y DISCUSION

4.1. Presentación de resultados

En este estudio, hemos aplicado diversas métricas de evaluación para calcular el rendimiento de nuestro modelo LightGBM y para evaluar la efectividad de nuestras predicciones en el contexto de la deserción académica en estudiantes de ingeniería civil de la FACET-UNT. Estas métricas son estándar en la evaluación de modelos de clasificación binaria y proporcionan una comprensión completa de cómo se desempeña nuestro modelo. A continuación, discutimos las métricas utilizadas:

Los resultados son los siguientes:

- Accuracy (la proporción de predicciones correctas.): 0.7816
- Precision (la proporción de verdaderos positivos entre todas las predicciones positivas.): 0.7957
- Recall (la proporción de verdaderos positivos entre todos los casos positivos reales.): 0.8170
- F1 Score (una métrica que combina precisión y recall.): 0.8062
- ROC AUC Score (el área bajo la curva ROC (Receiver Operating Characteristic)): 0.8699
- Log Loss (la pérdida logarítmica binaria.): 0.6720

Los resultados de las métricas de evaluación para el modelo LightGBM indican un rendimiento sólido en la tarea de clasificación de deserción académica en los estudiantes objeto del estudio. Con un valor de precisión (Precision) de 0.7957, el modelo muestra una alta capacidad para identificar correctamente a los estudiantes que realmente abandonarán sus estudios, minimizando así los falsos positivos. Además, el valor de recall de 0.8170 indica que el modelo es eficaz en capturar una gran proporción de los casos reales de deserción, lo que es fundamental para la detección temprana de estudiantes en riesgo.

El F1 Score de 0.8062, que combina precisión y recall, sugiere un equilibrio adecuado entre la identificación precisa y la minimización de errores. El alto valor del área bajo la curva ROC (ROC AUC Score) de 0.8699 indica que el modelo tiene una buena capacidad para discriminar entre las clases positivas y negativas. Por último, el valor de pérdida logarítmica binaria (Log Loss) de 0.6720 es coherente con un modelo bien calibrado. En conjunto, estos resultados sugieren que el modelo LightGBM es preciso en la identificación de la

deserción académica y muestra un rendimiento robusto en este contexto específico. En Tabla 1 se pueden observar los resultados de los números de Shapley para cada variable.

Tabla 1. Valores de Shapley para las variables de importancia

	Número Shapley
NúmeroPromociones_shap	3.651
NúmeroRegulares_shap	2.867
PromedioNotas_shap	2.569
TiempoFacultad_shap	2.533
EdadUltimaActividad_shap	2.510
NúmeroRecursadas_shap	1.923
NúmeroTotalCursadas_shap	1.873
MaxRegAcum_shap	1.726
NúmeroLibres_shap	1.658
NúmeroAprobados_shap	1.636
NúmeroReprobados_shap	1.490
NúmeroAusentes_shap	1.117
NúmeroExámenes_shap	0.786
Genero_shap	0.578

Fuente: Elaboración propia.

Los resultados obtenidos al aplicar la teoría de juegos de Shapley para analizar la incidencia de las variables en la deserción estudiantil son esclarecedores en términos de la contribución relativa de cada variable en la toma de decisiones de los estudiantes sobre abandonar o continuar con sus estudios. Aquí se presentan las observaciones clave basadas en los números Shapley:

- NúmeroPromociones_shap (3.651): Esta variable tiene el número Shapley más alto, lo que sugiere que es la característica más influyente en las decisiones de deserción. Un mayor número de promociones parece estar asociado con una menor probabilidad de deserción.
- NúmeroRegulares_shap (2.867): La cantidad de cursos regulares aprobados también tiene una influencia significativa en la toma de decisiones de los estudiantes. Un mayor número de cursos regulares aprobados se asocia con una menor probabilidad de deserción.
- PromedioNotas_shap (2.569): El promedio de calificaciones es otra variable relevante en la toma de decisiones. Un promedio de notas más alto parece ser un factor importante en la retención estudiantil.
- TiempoFacultad_shap (2.533): El tiempo que un estudiante ha permanecido en la facultad también es una variable influyente. Cuanto

más tiempo haya estado un estudiante en la facultad, es menos probable que abandone sus estudios.

- **EdadUltimaActividad_shap (2.510):** La edad en la que un estudiante realizó su última actividad académica tiene un impacto significativo. Los estudiantes más jóvenes en su última actividad tienden a tener una menor probabilidad de deserción.

Estos cinco factores enumerados anteriormente (número de promociones, número de cursos regulares aprobados, promedio de calificaciones, tiempo en la facultad y edad en la última actividad) emergen como los predictores más influyentes en la deserción estudiantil, según los números Shapley. Estos hallazgos pueden ser valiosos para identificar a los estudiantes en riesgo y desarrollar intervenciones específicas para mejorar la retención. A continuación, se pueden los resultados para las variables de mayor importancia.

4.2. Tiempo de permanencia en los estudios

Los datos presentados muestran la relación entre el tiempo que un estudiante ha permanecido en la facultad (TiempoFacultad) y la influencia de esta variable en la probabilidad de abandono académico (ShapvalueAbandono) (Ver Figura 1). Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores menores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

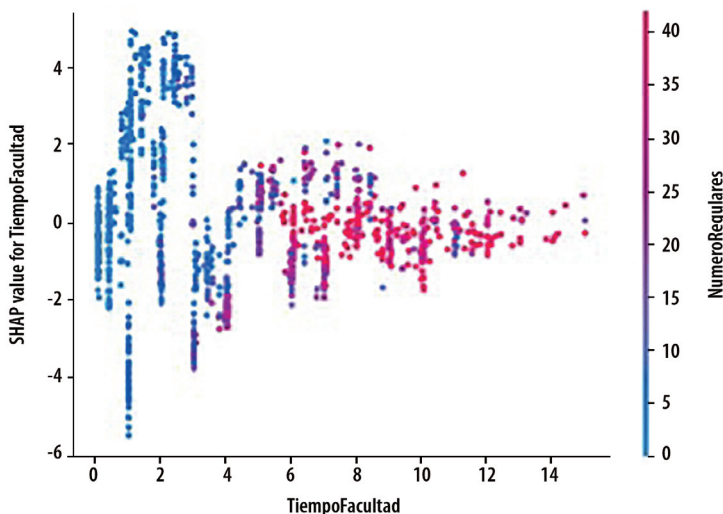
- **Tiempos Iniciales (0 a 1 año):** Para estudiantes que recién ingresan, los valores de Shap_value_Abandono fluctúan entre valores positivos y negativos, pero la tendencia general es hacia valores negativos. Esto sugiere que, en este primer período, el tiempo en la facultad es un factor que contribuye positivamente a la probabilidad de permanencia, aunque la influencia puede no ser tan fuerte.
- **Tiempos Iniciales (más de 1 año a 3 años)** Para estudiantes que han estado en la facultad durante este periodo, los valores de Shap_value_Abandono son principalmente positivos. Esto indica que este período se asocia con una mayor probabilidad de abandono. Los valores positivos sugieren que esta variable contribuye negativamente a la probabilidad de permanencia en este rango de tiempo. Este factor se acentúa claramente desde el año y medio hasta los 3 años.

- **Tiempos Intermedios (3 a 6 años):** En el rango de 3 a 6 años de permanencia en la facultad, los valores de *Shap_value_Abandono* fluctúan, pero la tendencia general es hacia valores negativos. Esto sugiere que, en este período intermedio, el tiempo en la facultad sigue siendo un factor que contribuye positivamente a la probabilidad de permanencia, aunque la influencia puede no ser tan fuerte como en los primeros años.
- **Tiempos Avanzados (6 años en adelante):** Para estudiantes que han estado en la facultad durante más de 6 años, los valores de *Shap_value_Abandono* tienden a ser negativos, con facetas en donde se torna positivo entre los 7 y 9 años. Posteriormente desciende nuevamente a valores negativos presentando una leve tendencia ascendente a medida que transcurren los años.

El análisis de los valores de Shapley para la variable «TiempoFacultad» sugiere un patrón relevante. Los estudiantes presentan un claro patrón de abandono en un período que va desde el año y medio hasta los tres años. Superado ese tiempo, los estudiantes que permanecen muestran un descenso en la probabilidad de abandono, la que nuevamente se incrementa cuando se llega a los 7 años de permanencia. Posteriormente se pasa nuevamente a una probabilidad de abandono, aunque menos marcada que la anterior. Se puede observar que existen períodos de mayor probabilidad de abandono que se alternan con otros de menor probabilidad. Si se compara con el número de asignaturas regulares, vemos que a mayor cantidad de esta variable existe una mayor probabilidad de permanencia, aunque este efecto positivo disminuye a medida que transcurre el tiempo. Para tiempos de permanencia mayores a 12 años se observa una mayor tendencia al abandono, aun a pesar de tener una cantidad alta de asignaturas regulares.

4.3. Edad del estudiante en su última actividad registrada

Los datos presentados muestran la relación entre la edad en la que un estudiante realizó su última actividad académica (*Edad_ultima_actividad*) y la influencia de esta variable en la probabilidad de abandono académico (*Shap_value_Abandono*) (Ver Figura 2). Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores me-

Figura 1. Valores de Número de Shapley para TiempoFacultad

Fuente: Elaboración propia.

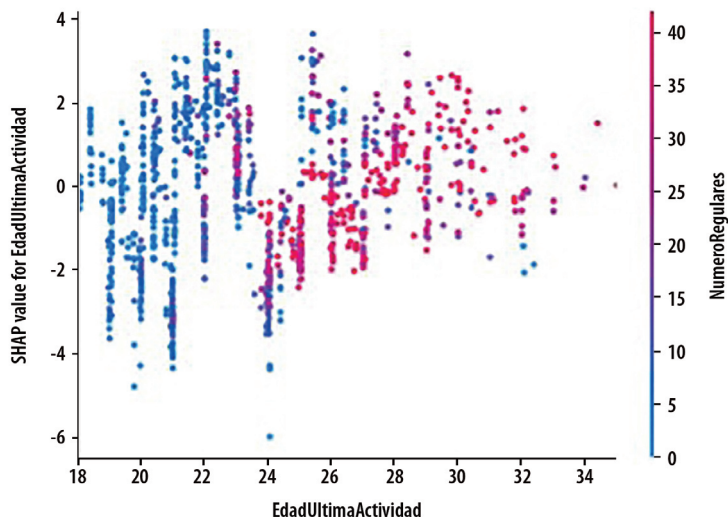
nores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

- **Edades Iniciales (18 a 19 años):** Para estudiantes que realizaron su última actividad académica a edades tempranas (entre 18 y 19 años), los valores de `Shap_value_Abandono` son variados. Algunas edades dentro de este rango tienen valores positivos, lo que indica una mayor probabilidad de abandono, mientras que otras tienen valores negativos, lo que sugiere una mayor probabilidad de permanencia. Esto podría deberse a otras variables no consideradas en este análisis que influyen en la toma de decisiones de los estudiantes.
- **Edades Intermedias (20 a 26 años):** En el rango de edades de 20 a 26 años, los valores de `Shap_value_Abandono` son predominantemente positivos. Esto indica que, en general, a medida que los estudiantes envejecen en este rango, tienen una mayor probabilidad de abandono. La influencia positiva en la probabilidad de abandono se vuelve más pronunciada a medida que la edad aumenta dentro de este rango. Se puede observar además que a la edad de entre 24 y 25 años se tienen valores de probabilidad de abandono marcadamente negativos.

- **Edades Avanzadas (27 años en adelante):** Para estudiantes que realizaron su última actividad académica a edades avanzadas (27 años en adelante), los valores de `Shap_value_Abandono` son principalmente positivos y tienden a aumentar. Esto sugiere que, después de cierta edad, un mayor "envejecimiento" se asocia con una mayor probabilidad de abandono.

El análisis de los valores de Shapley para la variable «`Edad_ultima_actividad`» indica que la edad en la que un estudiante realizó su última actividad académica tiene una influencia en la probabilidad de abandono académico. Las edades más jóvenes (18 a 19 años) y las avanzadas (27 años en adelante) se asocian con mayor probabilidad de abandono, mientras que las intermedias (20 a 26 años) influyen positivamente en la probabilidad de abandono a medida que aumenta la edad.

Figura 2. Valores de Número de Shapley para `EdadUltimaActividad`



Fuente: Elaboración propia.

4.4. Asignaturas cursadas regularizadas

Los datos presentados muestran la relación entre el número de cursos regulares aprobados (`NumeroRegulares`) y la influencia de esta variable en la probabilidad de abandono académico (`Shap_value_Abandono`) (Ver Figura 3).

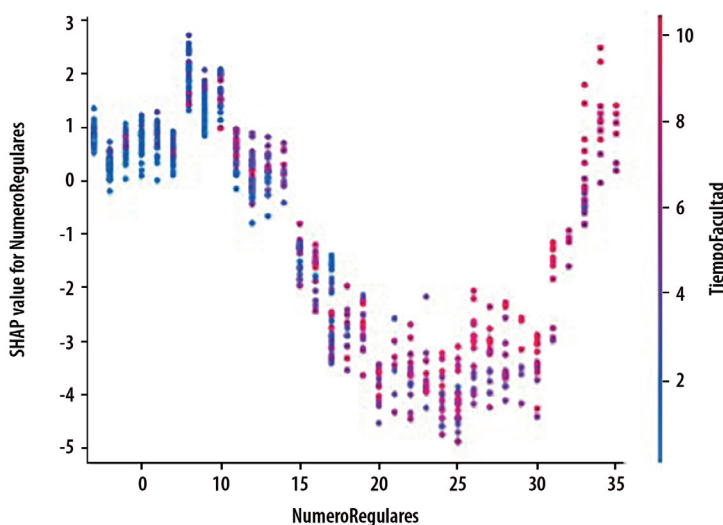
Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores menores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

- **Número Bajo de Cursos Regulares (0 a 5):** Para estudiantes que han aprobado un número bajo de cursos regulares (0 a 5), los valores de Shap_value_Abandono son positivos, lo que indica una mayor probabilidad de abandono. Esto sugiere que los estudiantes que han aprobado pocos cursos regulares tienen una mayor propensión a abandonar sus estudios. A medida que el número de cursos regulares aprobados aumenta dentro de este rango, la influencia positiva disminuye, pero sigue siendo significativa.
- **Número Moderado de Cursos Regulares (6 a 13):** En el rango de 6 a 13 cursos regulares aprobados, los valores de Shap_value_Abandono fluctúan, pero tienden a ser positivos en general. Esto sugiere que, incluso cuando los estudiantes aprueban una cantidad moderada de cursos regulares, todavía existe una influencia positiva en la probabilidad de abandono, aunque esta influencia se vuelve menos pronunciada a medida que se aprueban más cursos.
- **Número Alto de Cursos Regulares (14 a 30):** Para estudiantes que han aprobado un número significativamente alto de cursos regulares, los valores de Shap_value_Abandono son negativos. Esto indica que un mayor número de cursos regulares aprobados está asociado con una menor probabilidad de abandono. A medida que los estudiantes completan más cursos regulares, la influencia negativa en la probabilidad de abandono aumenta significativamente.
- **Número Alto de Cursos Regulares (30 en adelante):** Sin embargo, a partir de las 30 asignaturas regularizadas, se muestra un cambio en la tendencia de permanencia. A partir de las 32 asignaturas regularizadas se puede observar que los valores de Shap pasan a ser de signo positivo, mostrando una tendencia al abandono asociado a una mayor permanencia en la facultad. Aunque la influencia positiva en la permanencia de haber regularizado una gran cantidad de asignatura deja de ser efectiva porque el estudiante permanece más de ocho años en la facultad. Esto nos mostraría un punto de inflexión en la relación entre las asignaturas regulares y la permanencia, en la que el tiempo de permanencia en la

facultad comienza a tener mayor relevancia y afecta significativamente el abandono de los estudiantes.

El análisis de los valores de Shapley para la variable «NúmeroRegulares» revela patrones interesantes. Los estudiantes que han aprobado un número bajo o moderado de cursos regulares tienen una mayor probabilidad de abandono, mientras que aquellos que han aprobado un número significativamente alto de cursos regulares, al menos hasta las 30 asignaturas, tienen una menor probabilidad de abandono.

Figura 3. Valores de Número de Shapley para NúmeroRegulares



Fuente: Elaboración propia.

4.5. Número de asignaturas aprobadas

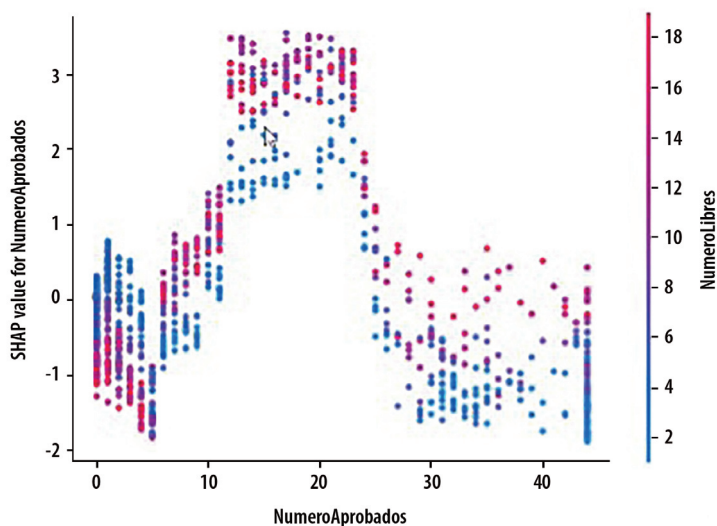
Los datos presentados muestran la relación entre el tiempo que un estudiante ha permanecido en la facultad (NúmeroAprobadas) y la influencia de esta variable en la probabilidad de abandono académico (Shap_value_Abandono). Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores menores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

- **Número Bajo de Asignaturas Aprobadas (0 a 11):** Para estudiantes que han aprobado un número bajo de asignaturas (0 a 11), los valores de `Shap_value_Abandono` son principalmente negativos. Esto indica que, en general, un menor número de asignaturas aprobadas está asociado con una mayor probabilidad de abandono. La influencia negativa en la probabilidad de abandono es más pronunciada en el rango de 0 a 7 asignaturas aprobadas.
- **Número Moderado de Asignaturas Aprobadas (12 a 27):** En el rango de 12 a 27 asignaturas aprobadas, los valores de `Shap_value_Abandono` fluctúan, pero en su mayoría tienden a ser positivos. Esto sugiere que, a medida que los estudiantes aprueban un número moderado de asignaturas, la influencia positiva en la probabilidad de abandono se vuelve más prominente. Los estudiantes que han aprobado entre 12 y 27 asignaturas a menudo tienen una mayor probabilidad de abandono. Esto está correlacionado con mayores valores de asignaturas en las que han quedado libres; es decir que deben recurrir a las mismas para poder seguir en la carrera.
- **Número Alto de Asignaturas Aprobadas (más de 27):** Para estudiantes que han aprobado un número significativamente alto de asignaturas (más de 27), los valores de `Shap_value_Abandono` son predominantemente negativos, con valores positivos para aquellos estudiantes que presentan un mayor número de asignaturas en las que quedaron libres. Esto indica que un mayor número de asignaturas aprobadas está fuertemente asociado con una mayor probabilidad de abandono, aunque condicionada por la cantidad de asignaturas en las que quedaron libres. La influencia positiva en la probabilidad de abandono se vuelve más pronunciada a medida que se aumenta el número de asignaturas aprobadas.

En general, los valores de Shap sugieren que promocionar más asignaturas está relacionado con una menor probabilidad de abandono, y a medida que el número de asignaturas promocionadas aumenta, este efecto es más pronunciado.

4.6. Número de asignaturas promociones

Los datos presentados muestran la relación entre el número de asignaturas aprobadas (`Número_asignaturas_aprobadas`) y la influencia de esta variable

Figura 4. Valores de Número de Shapley para Número Aprobados

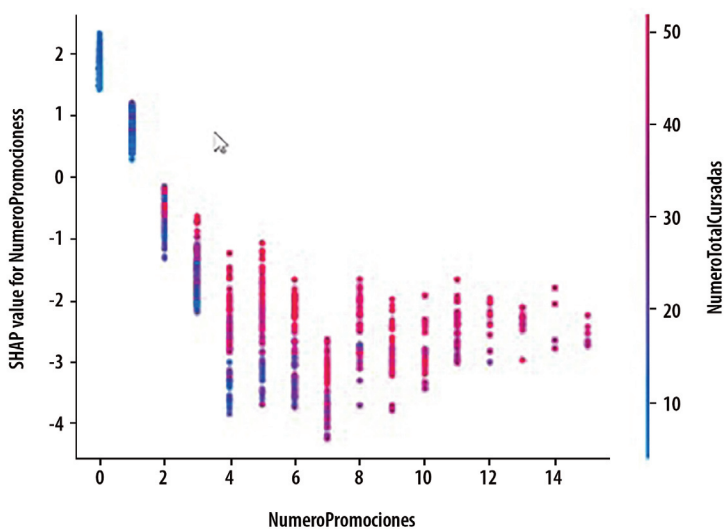
Fuente: Elaboración propia.

en la probabilidad de abandono académico (Shap_value_Abandono) (Ver Figura 5). Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores menores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

- Cuando el número de asignaturas promocionadas es 0, el valor de Shap es 2, lo que sugiere que tener 0 asignaturas promocionadas está asociado con una mayor probabilidad de abandono.
- A medida que aumenta el número de asignaturas promocionadas, el valor de Shap disminuye gradualmente, llegando a ser negativo a partir de 2. Esto significa que a medida que promocionas más asignaturas, la probabilidad de abandono disminuye.
- A partir de un valor de 7 en adelante, los valores de Shap son cada vez más negativos, lo que indica que promocionar un mayor número de asignaturas tiene un impacto aún más positivo en la reducción de la probabilidad de abandono.

El análisis de los valores de Shapley para la variable «Número_asignaturas_aprobadas» sugiere que el progreso académico y el número de asignaturas aprobadas tienen una influencia significativa en la probabilidad de abandono académico. En particular, los estudiantes que han aprobado un número bajo o moderado de asignaturas tienden a mostrar una mayor probabilidad de abandono, mientras que aquellos que han aprobado un número significativamente alto de asignaturas también tienen una mayor probabilidad de abandono.

Figura 5. Valores de Número de Shapley para NúmeroPromociones



Fuente: Elaboración propia.

4.7. Número de asignaturas reprobadas

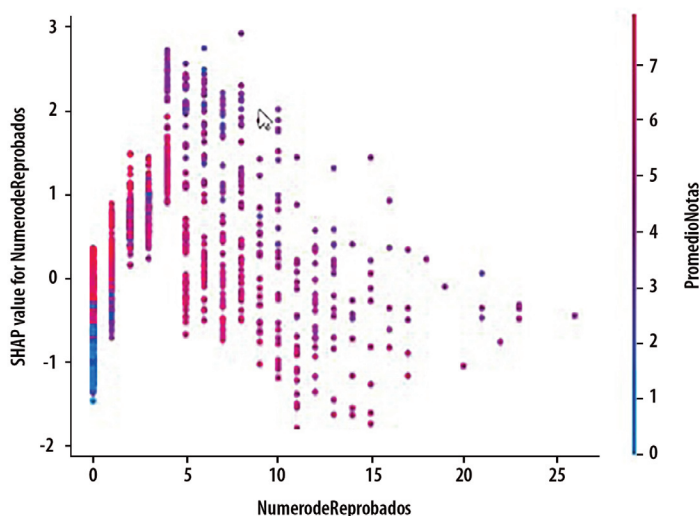
Los datos presentados muestran la relación entre el número de asignaturas reprobadas (Número_asignaturas_reprobadas) y la influencia de esta variable en la probabilidad de abandono académico (Shap_value_Abandono) (Ver Figura 6). Dado que los valores de Shapley mayores a 0 indican una mayor probabilidad de abandono, mientras que los valores menores a 0 sugieren una mayor probabilidad de permanencia, podemos realizar las siguientes observaciones:

- **Número Bajo de Asignaturas Reprobadas (0 a 5):** Para estudiantes que tienen un número bajo de asignaturas reprobadas (0 a 5), los valores de Shap_value_Abandono son principalmente negativos, lo que indica una menor probabilidad de abandono. Esto sugiere que los estudiantes que han reprobado pocas asignaturas tienden a tener una mayor probabilidad de permanencia en la facultad. Los valores negativos son más pronunciados en el rango de 0 a 3 asignaturas reprobadas.
- **Número Moderado de Asignaturas Reprobadas (6 a 16):** En el rango de 6 a 16 asignaturas reprobadas, los valores de Shap_value_Abandono fluctúan, pero tienden a ser positivos en general. Esto indica que, a medida que el número de asignaturas reprobadas aumenta dentro de este rango, la influencia positiva en la probabilidad de abandono se vuelve más prominente. Los estudiantes que han reprobado entre 6 y 16 asignaturas tienden a mostrar una mayor probabilidad de abandono.
- **Número Alto de Asignaturas Reprobadas (más de 16):** Para estudiantes que tienen un número significativamente alto de asignaturas reprobadas (más de 16), los valores de Shap_value_Abandono son levemente positivos. Esto indica que un mayor número de asignaturas reprobadas está asociado con una mayor probabilidad de abandono. La influencia positiva en la probabilidad de abandono muestra un descenso a medida que aumenta el número de asignaturas reprobadas, sin embargo, también se observa que la cantidad de estudiantes en este rango de valores desciende drásticamente, lo que muestra que aquellos con un gran número de reprobados abandonaron la carrera en instancias anteriores.

El análisis de los valores de Shapley para la variable «Número_asignaturas_reprobadas» sugiere que el número de asignaturas reprobadas tiene una influencia significativa en la probabilidad de abandono académico. Los estudiantes con pocas asignaturas reprobadas tienden a mostrar menor probabilidad de abandono, mientras que los que han reprobado un número moderado o alto de asignaturas suelen tener una mayor probabilidad de abandono.

5. DISCUSIÓN DE LOS RESULTADOS

Los resultados de este estudio que analizan el abandono universitario, examinados principalmente a través de variables como el rendimiento académico,

Figura 6. Valores de Número de Shapley para NúmeroReprobados

Fuente: Elaboración propia.

el tiempo de permanencia en la facultad y la edad, corroboran hallazgos de investigaciones previas, proporcionando un panorama más detallado y actualizado de los factores que influyen en la deserción estudiantil. Entendemos además que la importancia del estudio se centra además en la cuantificación del efecto de estas variables, ya que se discrimina como influyen estas variables con relación a su valor específico.

Estos hallazgos están en consonancia con investigaciones previas que destacan la importancia del rendimiento académico en la retención estudiantil. Estudios recientes han confirmado que el bajo rendimiento académico, medido a través de la cantidad de cursos reprobados, es uno de los principales factores que contribuyen a la deserción universitaria. Además, investigaciones realizadas por Alban y Mauricio (2019) señalan que los factores académicos, junto con económicos y sociofamiliares, son determinantes críticos en la decisión de abandonar los estudios universitarios.

Estudios recientes han identificado que un bajo rendimiento académico es un predictor fuerte de la deserción. Por ejemplo, una investigación encontró

que los estudiantes con rendimiento académico bajos tienen una mayor probabilidad de abandonar sus estudios debido a la acumulación de dificultades académicas y la pérdida de motivación (Realinho et al. 2022, Won et al. 2023). Similarmente, la utilización de modelos de aprendizaje automático para predecir el rendimiento académico y la deserción ha mostrado que los estudiantes con dificultades académicas tempranas presentan un riesgo significativamente mayor de no completar sus estudios (Realinho et al. 2022).

Un estudio brasileño de ingenieros que desertaron de la ciencia encontró que el 61% de los desertores universitarios entre 2013 y 2014 en la Universidad Tecnológica Federal de Paraná lo hicieron por problemas académicos. Asimismo, el 29,5% de los que habían dejado la carrera de ingeniería eligieron otras facultades en la nueva universidad (Soistak Christo, Martins de Resende, y Galan Kuhn 2018).

En cuanto al tiempo de permanencia en la facultad y la edad, los estudios indican que los estudiantes mayores tienden a abandonar con más frecuencia que sus contrapartes más jóvenes. Esto puede deberse a las responsabilidades adicionales que los estudiantes adultos enfrentan, como el trabajo y la familia, que pueden interferir con su capacidad para mantenerse en sus estudios (Realinho et al. 2022, Won et al. 2023). Además, el tiempo prolongado en la facultad sin progresar significativamente hacia la graduación incrementa la probabilidad de deserción debido a la fatiga académica y la presión financiera acumulada (Realinho et al. 2022). Por su parte, en otro estudio se ha observado que los estudiantes que no completan sus estudios en el tiempo estándar tienen una mayor probabilidad de abandonar, ya que los retrasos en la graduación incrementan las tasas de deserción (Rumberger 2020).

Estos resultados también concuerdan con investigaciones que utilizan modelos predictivos basados en datos demográficos y académicos para identificar a los estudiantes en riesgo de deserción temprana. La integración de múltiples factores, incluyendo el rendimiento académico y las características demográficas, permite a las instituciones educativas desarrollar estrategias de intervención más efectivas y personalizadas para reducir la tasa de deserción (Won et al. 2023).

Nuestros hallazgos reafirman que tanto el rendimiento académico como el tiempo de permanencia en la facultad son indicadores cruciales de la deserción estudiantil. La implementación de sistemas de alerta temprana y programas

de apoyo dirigidos específicamente a estos factores podría ser una estrategia eficaz para mejorar la retención estudiantil y apoyar el éxito académico.

5.1. Consideraciones sobre limitaciones y sesgos

Al considerar los resultados previamente analizados, es esencial tener en cuenta algunas limitaciones y posibles sesgos en nuestra investigación. En primer lugar, aunque hemos identificado variables significativas que influyen en la probabilidad de abandono académico, este estudio se basa en datos históricos de estudiantes de ingeniería civil de una institución específica, lo que puede limitar la generalización de nuestros hallazgos a otros contextos académicos. Además, no hemos considerado ciertas variables socioeconómicas o personales que podrían influir en las decisiones de abandono de los estudiantes. A pesar de estas limitaciones, nuestros resultados proporcionan una visión valiosa de las variables clave que deben ser consideradas por las instituciones educativas al desarrollar estrategias de retención y apoyo a los estudiantes. Futuras investigaciones pueden abordar estas limitaciones y explorar aún más las interacciones complejas que influyen en las decisiones de abandono académico.

Es fundamental destacar que, si bien realizamos un análisis exhaustivo utilizando el método SHAP (Shapley Additive Explanations), los resultados obtenidos no establecen una relación de causalidad entre las variables analizadas y el abandono académico. En otras palabras, no podemos inferir que las variables identificadas y sus respectivos resultados sean las causas directas del abandono estudiantil. Lo que estos resultados proporcionan son pistas valiosas y significativas sobre las influencias y patrones asociados a la probabilidad de abandono. La retención académica es un fenómeno multidimensional y complejo, y las decisiones de los estudiantes pueden estar influenciadas por factores interrelacionados, algunos de los cuales pueden no haber sido considerados en este análisis. Por lo tanto, es esencial que futuras investigaciones exploren más a fondo estas relaciones y consideren posibles variables adicionales para una comprensión más completa de los determinantes del abandono académico.

6. CONCLUSIONES

Este estudio ha arrojado luz sobre los factores que influyen en la probabilidad de abandono académico en estudiantes de ingeniería civil de la FACET-UNT.

Hemos identificado variables significativas, como el número de asignaturas reprobadas, el número de asignaturas aprobadas, la edad en la que se realizó la última actividad académica y el tiempo de permanencia en la facultad, asociadas, todas estas con la probabilidad de abandono.

Sin embargo, es primordial enfatizar que estos resultados no establecen relaciones causales, sino que ofrecen pistas valiosas para comprender mejor este fenómeno complejo. Además, hemos utilizado el método SHAP para analizar la importancia relativa de estas variables en la toma de decisiones de abandono, destacando la necesidad de brindar un apoyo más personalizado a los estudiantes que enfrentan dificultades académicas. Si bien este estudio contribuye a nuestra comprensión de la retención estudiantil, es fundamental reconocer las limitaciones inherentes y la falta de causalidad en nuestros hallazgos. Las instituciones educativas pueden utilizar esta información como punto de partida para desarrollar estrategias de retención más efectivas y brindar apoyo a los estudiantes de manera más precisa.

Las implicaciones prácticas derivadas de los hallazgos de este estudio son significativas para las instituciones educativas que deseen mejorar sus estrategias de retención estudiantil. En primer lugar, dado que variables como el número de asignaturas reprobadas, el número de asignaturas aprobadas, la edad en la que se realizó la última actividad académica y el tiempo de permanencia en la facultad tienen un impacto en la probabilidad de abandono, las instituciones pueden utilizar esta información para identificar a los estudiantes en riesgo y brindarles apoyo personalizado. Esto podría incluir asesoramiento, tutorías específicas para áreas problemáticas y programas de seguimiento para monitorear el progreso académico de cerca.

Además, la aplicación del método SHAP para analizar la importancia relativa de estas variables resalta la necesidad de considerar enfoques más avanzados de análisis de datos y explicación de modelos en la toma de decisiones de retención. Las instituciones pueden aprovechar técnicas similares para comprender mejor los factores que influyen en el abandono y diseñar intervenciones más efectivas.

Para futuras investigaciones, se ampliará el alcance de este estudio considerando la inclusión de variables adicionales, como factores socioeconómicos. Esto permitirá una comprensión más completa de los determinantes del abandono académico. Además, indagaciones longitudinales que sigan a los estu-

diantes a lo largo de su trayectoria académica podrían proporcionar información valiosa sobre cómo evolucionan estos factores con el tiempo.

Para finalizar, consideramos que este estudio subraya la importancia de abordar la retención estudiantil como un desafío multidimensional que requiere enfoques holísticos y personalizados. Las recomendaciones basadas en evidencia y la aplicación de técnicas avanzadas de análisis de datos pueden contribuir significativamente a la mejora de las tasas de retención y, en última instancia, al éxito académico de los estudiantes.

REFERENCIAS

- Abdala Leiva, S., Castiglione, A.M. y Infante, L.A. (2008). La deserción universitaria. Una asignatura pendiente para la gestión institucional. *Cuadernos de La Facultad de Humanidades y Ciencias Sociales*, Universidad Nacional de Jujuy, 34, 173-91.
- Alban, M. y David M. (2019). Predicting university dropout through data mining: A systematic literature. *Indian Journal of Science and Technology* 12(4), 1-12.
- Arancibia, S., Rodríguez, G., Fritis, R. Tenorio, N. y Poblete, H. (2013) Representaciones sociales en torno a equidad, acceso y adaptación en educación universitaria. *Psicoperspectivas* 12 (1), 116-38.
<http://dx.doi.org/10.5027/psicoperspectivas-Vol12-Issue1-fulltext-236>
- Araque, F., Roldán, C. y Salguero, A. (2009). Factors influencing university drop out rates. *Computers & Education* 53(3), 563-74.
- Arce, D. M., Gallo, L., Guiller, C., Iotti, A., & Ungaro, A. M. (2017). Una aproximación a las experiencias estudiantiles que se configuran en el proceso de afiliación académica en la universidad: Su lectura desde la mirada de los estudiantes en sus atravesamientos por el escenario cultural actual. Montevideo.
<https://bit.ly/3VZ3J2Q>
- Armatas, C., & Spratt, C. F. (2019). Applying learning analytics to program curriculum review. *The International Journal of Information and Learning Technology*, 36(3), 243-253. <https://doi.org/10.1108/IJILT-11-2018-0133>
- Barbier, J., Dia, M., Macris, N., Krzakala, F., Lesieur, T., & Zdeborová, L. (2016). Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 29). Curran Associates, Inc. <https://bit.ly/3xVecEE>

- Bernardo, A., Esteban, M., Fernández, E., Cervero, A., Tuero, E., & Solano, P. (2016). Comparison of personal, social and academic variables related to university drop-out and persistence. *Frontiers in Psychology*, 7, 1610.
- Cabrera, L., Bethencourt, J. T., Alvarez Pérez, P., & González Afonso, M. (2006). The problem of University dropout. *RELIEVE - Revista Electrónica de Investigación y Evaluación Educativa*, 12(2), 171-203. <https://doi.org/10.7203/relieve.12.2.4226>
- Castillo-Sánchez, M., Gamboa-Araya, R., & Hidalgo-Mora, R. (2020). Factores que influyen en la deserción y reprobación de estudiantes de un curso universitario de matemáticas. *Uniciencia*, 34(1), 219-245.
- Chacín, A. J. P., Inciarte González, A., & Walles Peñaloza, D. (2020). Educación superior e investigación en Latinoamérica: Transición al uso de tecnologías digitales por Covid-19. *Revista de Ciencias Sociales*, 26(3), 98-117.
- Da Re, L., & Clerici, R. (2017). Abandono, rendimiento académico y tutoría: una investigación de la Universidad de Padua. *Educatio Siglo XXI*, 35(2 Jul-Oct), 139-160.
- De Coninck, D., Matthijs, K., & Luyten, P. (2019). Subjective well-being among first-year university students: A two-wave prospective study in Flanders, Belgium. *Student Success*, 10(1), 33-46.
- Fonseca, G., & García, F. (2016). Permanence and dropout rates among university students: An organizational theory analysis. *Revista de la Educación Superior*, 45(179), 25-39.
- Gagliardi, J., Parnell, A., & Carpenter-Hubin, J. (2018). The Analytics Revolution in Higher Education. *Change: The Magazine of Higher Learning*, 50(March), 22-29. <https://doi.org/10.1080/00091383.2018.1483174>
- González Fiegehen, L. E. (2007). *Repitencia y deserción universitaria en América Latina. En Informe sobre la Educación Superior en América Latina y el Caribe 2000-2005. La Metamorfosis de la educación superior* (pp. 156-170), IESALC.
- Gonzalez-Ramirez, T., & Pedraza-Navarro, I. (2017). Social and families variables associated with university drop-out. *Educatio Siglo XXI*, 35(2), 365-387.
- Hilliger, I., Miranda, C., Celis, S., & Pérez-Sanagustín, M. (2019). Evaluating Usage of an Analytics Tool to Support Continuous Curriculum Improvement. In M. Scheffel, J. Broisin, Pammer-Schindler, A. Ioannou, & J. Schneider (Eds.), 14th European Conference on Technology Enhanced Learning EC-TEL 2019 (pp. 114). Springer. <https://bit.ly/3xlAoln>
- Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. <https://bit.ly/3LpB6Xy>

- Kori, K., Pedaste, M., & Must, O. (2018). The academic, social, and professional integration profiles of information technology students. *ACM Transactions on Computing Education (TOCE)*, 18(4), 1-19.
- Lang, C., Siemens, G., Wise, A., & Gašević, D. (Eds.). (2017). *Handbook of Learning Analytics (1st ed.)*. Society for Learning Analytics Research.
<https://doi.org/10.18608/hla17>
- Lázaro Álvarez, N. (2020). Acciones tutoriales con TIC atendiendo a factores predictivos de la deserción estudiantil en carreras de Ingeniería Informática. Universidad de Granada. <https://bit.ly/4bB1usd>
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2(1), 56-67. <https://doi.org/10.1038/s42256-019-0138-9>
- Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates. <https://bit.ly/4czw0Jq>
- Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., Liston, D. E., King-Wai Low, D., Newman, S.F., Kim, J. & Lee, S.I. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering*, 2(10), 749-760.
<https://doi.org/10.1038/s41551-018-0304-0>
- Mitic, M., Woodcock, K. A., Amering, M., Krammer, I., Stiehl, K. A. M., Zehetmayer, S., & Schrank, B. (2021). Toward an integrated model of supportive peer relationships in early adolescence: A systematic review and exploratory meta-analysis. *Frontiers in Psychology*, 12, 589403.
- Namoun, A., & Alshantqiti, A. (2021). Predicting Student Performance Using Data Mining and Learning Analytics Techniques: A Systematic Literature Review. *Applied Sciences*, 11(1). <https://doi.org/10.3390/app11010237>
- Niitsoo, M., Paales, M., Pedaste, M., Siiman, L., & Tõnisson, E. (2014). Predictors of informatics students progress and graduation in university studies. In International Technology, Education and Development Conference (pp. 380-392). <https://bit.ly/3zEdYIG>
- Ozdemir, D., Opseth, H.M. & Taylor, H. (2020). Leveraging learning analytics for student reflection and course evaluation. *Journal of Applied Research in Higher Education*, 12(1), 27-37. <https://doi.org/10.1108/JARHE-11-2018-0253>
- Paz, H. R. (2022). Hacia un nuevo paradigma en el diseño curricular. *Revista Argentina de Investigación Educativa*, 2(4), 139-160.

- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Edouard Duchesnay. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825-2830. <https://bit.ly/3LkJPKU>
- Ponce, J. P. D. A. (2021). Articulación de variables de medición de satisfacción estudiantil y desempeño docente en entornos virtuales en escuelas de negocios. *Palermo Business Review*, 23, 103-120.
- Realinho, V., Machado, J., Baptista, L., & Martins, M. V. (2022). Predicting Student Dropout and Academic Success. *Data*, 7(11). <https://doi.org/10.3390/data7110146>
- Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *WIREs Data Mining and Knowledge Discovery*, 10(3), e1355. <https://doi.org/10.1002/widm.1355>
- Rumberger, R. W. (2020). Chapter 12 - The economics of high school dropouts. In S. Bradley & C. Green (Eds.), *The Economics of Education* (2nd ed., pp. 149-158). Academic Press. <https://doi.org/10.1016/B978-0-12-815391-8.00012-4>
- Shi, Y., Ke, G., Chen, Z., Zheng, S., & Liu, T.-Y. (2022). Quantized Training of Gradient Boosting Decision Trees. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 18822-18833). Curran Associates, Inc. <https://bit.ly/4bzwwwM>
- Soistak Christo, M. M., de Resende, L. M. M., & do Carmo Galan Kuhn, T. (2018). Why engineering students give up in their formation: A case study. *Nuances-Estudos sobre Educacao*, 29(1), 154-168.
- Tinto, V. (1975). Dropout from higher education: A theoretical synthesis of recent research. *Review of Educational Research*, 45(1), 89-125. <https://doi.org/10.3102/00346543045001089>
- Tomás, J.-M., & Gutiérrez, M. (2019). Aportaciones de la teoría de la autodeterminación a la predicción de la satisfacción escolar en estudiantes universitarios. *Revista de Investigación Educativa*, 37(2), 471-485.
- Volungeviciene, A., Duart, J. M., Naujokaitiene, J., Tamoliune, G., & Misiuliene, R. (2019). Learning Analytics: Learning to Think and Make Decisions. *Journal of Educators Online*, 16(2), n2.
- Won, H.-S., Kim, M.-J., Kim, D., Kim, H.-S., & Kim, K.-M. (2023). University student dropout prediction using pretrained language models. *Applied Sciences*, 13(12). <https://doi.org/10.3390/app13127073>
- Xenos, M., Pierrakeas, C., & Pintelas, P. (2002). A survey on student dropout rates and dropout causes concerning the students in the Course of Informatics of the Hellenic Open University. *Computers & Education*, 39(4), 361-377.