ISBN: 978-628-96613-1-6. ISSN: 2414-6390. Digital Object Identifier: https://dx.doi.org/10.18687/LACCEI2025.1.1.1998

Web Platform for the Predictive Analysis of Student Dropout in a Higher Education Institution in Latin America

Blanco López, Jaime[1]; Galeano Ospino, Saray[2]; Niño Manrique, Jhon Fredy [3]; Ramírez Chiquito, Alexander[4]; Gonzalez Santamaría, Andrés Esteban[5]

[1,2,3,4,5]Corporación Universitaria Adventista (CO), Colombia, *jaime.blanco@unac.edu.co, sagaleano@unac.edu.co, jfnino@unac.edu.co, alexander.ramirezc@unac.edu.co, andres.gonzalezs@unac.edu.co*

Abstract– ***Student dropout in Colombian higher education is a complex problem that affects educational quality and the socioeconomic development of the country. This research focused on the development and implementation of a web platform to analyze and predict student dropout in a denominational higher education institution in Colombia, using a logistic regression model. The research was developed in four phases: 1) literature review and expert consultation to identify input variables related to dropout, 2) requirements definition, system architecture design and technology selection, 3) project development using Kanban methodology, and 4) implementation of the solution, execution of a pilot test and satisfaction evaluation through a survey. The predictive algorithm, with an accuracy of over 90%, allows the institution's professionals to detect early on students at risk of dropping out. The web platform, developed using Python provides an intuitive interface to visualize the current dropout status of faculties and programs, predict student dropout risk and understand the influencing factors. Validation through a pilot test and a satisfaction survey confirmed the effectiveness and usability of the platform, although areas of improvement for future implementations were also identified.***

Keywords-- Student desertion, Higher Education Institutions, predictive model, web platform.

Plataforma Web para el Análisis Predictivo de la Deserción Estudiantil en una Institución de Educación Superior en América Latina

Blanco López, Jaime[1]; Galeano Ospino, Saray[2]; Niño Manrique, Jhon Fredy [3]; Ramírez Chiquito, Alexander[4]; Gonzalez Santamaría, Andrés Esteban[5]
[1,2,3,4,5]Corporación Universitaria Adventista (CO), Colombia, *jaime.blanco@unac.edu.co, sagaleano@unac.edu.co, jfnino@unac.edu.co, alexander.ramirezc@unac.edu.co, andres.gonzalezs@unac.edu.co*

Resumen– ***La deserción estudiantil en la educación superior colombiana es un problema complejo que afecta la calidad educativa y el desarrollo socioeconómico del país. Esta investigación se centró en el desarrollo e implementación de una plataforma web para analizar y predecir la deserción estudiantil en una institución de educación superior confesional en Colombia, utilizando un modelo de regresión logística. La investigación se desarrolló en cuatro fases: 1) revisión de literatura y consulta a expertos para identificar variables entradas relacionadas con la deserción, 2) definición de requisitos, diseño de la arquitectura del sistema y selección de tecnologías, 3) desarrollo del proyecto utilizando la metodología Kanban, y 4) implementación de la solución, ejecución de una prueba piloto y evaluación de la satisfacción mediante una encuesta. El algoritmo predictivo, con una precisión superior al 90%, permite a los profesionales de la institución detectar tempranamente a los estudiantes en riesgo de deserción. La plataforma web, desarrollada utilizando Python ofrece una interfaz intuitiva para visualizar el estado actual de la deserción de las facultades y programas, predecir el riesgo de deserción del estudiante y comprender los factores influyentes. La validación mediante una prueba piloto y una encuesta de satisfacción confirmó la efectividad y usabilidad de la plataforma, aunque también se identificaron áreas de mejora para futuras implementaciones.***

Palabras clave-- ***Deserción estudiantil, Instituciones de Educación Superior, modelo predictivo, plataforma web.***

I. Introducción

La deserción estudiantil representa una problemática de alcance nacional e internacional que inquieta a la comunidad académica, generando graves repercusiones económicas y sociales [1]. Cerca del 50% de estudiantes abandonan sus estudios, este fenómeno se configura como un factor preocupante que exige la implementación de estrategias para su identificación y seguimiento [2]. En Colombia, según el Sistema de Prevención y Análisis de la Deserción en las Instituciones de Educación Superior (SPADIES), cuando un estudiante no matricula dos semestres consecutivos se considera un desertor [3]. Además, cerca del 15% de los estudiantes universitarios deserta al realizar el primer semestre, este porcentaje aumenta en el segundo semestre a un 20% y luego en los semestres siguientes la deserción acumulada es cerca del 45% [3].

Diversas y complejas variables inciden en la deserción estudiantil tales como: la elección de carrera, adaptación a la vida universitaria, consumo de sustancias psicoactivas, edad, género, estado civil, expectativas personales insatisfechas, relaciones intrafamiliares, motivación, expectativas personales, problemas de salud, temperamento, apatía, tendencia a la depresión, ausencia de perspectiva del futuro e incompatibilidad de los valores personales con los valores institucionales [4]. Sin embargo, estas variables son propias a los estudiantes y cambiantes según el entorno en el que se desarrolla la institución.

Por lo tanto, esta investigación tiene como objetivo desarrollar una plataforma web que permita analizar las variables que están relacionadas en la deserción estudiantil en una Institución de Educación Superior confesional en América Latina, mediante un modelo predictivo, con regresión logística.

La investigación se llevó a cabo en cuatro fases. En la primera fase se revisó la literatura y se consultó a expertos para identificar las variables que intervienen en la deserción. Seguidamente, en la fase dos se definieron los requisitos mediante historias de usuarios, se diseñó la arquitectura del sistema y se seleccionaron las tecnologías a utilizar. Luego, en la fase tres se desarrolló el proyecto utilizando la metodología Kanban para un flujo de trabajo eficiente y adaptable. Finalmente, en la fase cuatro se implementó la solución, se ejecutó una prueba piloto y se midió la satisfacción de los interesados mediante una encuesta.

El algoritmo desarrollado muestra una alta precisión al identificar estudiantes en riesgo de deserción, con una tasa de acierto superior al 90%. Dicho algoritmo integrado en la plataforma permite a los profesionales de la institución detectar tempranamente a aquellos estudiantes con alta probabilidad de abandonar sus estudios. Esta información facilita la toma de decisiones y la implementación de estrategias de apoyo, con el propósito de reducir significativamente las tasas de deserción y promover la permanencia estudiantil.

Este trabajo está estructurado de la siguiente forma: la sección 1 muestra los antecedentes sobre la práctica del análisis predictivo en la deserción estudiantil en una IES. Luego, en la sección 2 se describen los pasos y actividades a seguir para la construcción del modelo y el desarrollo de la plataforma web para analizar la deserción estudiantil. Seguidamente, en la sección 3 se muestran los resultados

obtenidos y en la sección 4 las discusiones. Finalmente, en la sección 5 se presentan las conclusiones y el trabajo futuro.

II. ANTECEDENTES

En esta sección se describen los trabajos relacionados con las propuestas que abordan el análisis predictivo de la deserción estudiantil en instituciones de educación superior en América Latina.

López en su estudio presentó un modelo matemático basado en regresión logística para predecir la deserción estudiantil en un Instituto de Educación Superior (IES), analizando datos de 849 estudiantes entre 2018 y 2020. El modelo final alcanzó una precisión del 83% en datos de entrenamiento y 79% en datos de prueba, mostrando que los estudiantes que han reprobado alguna asignatura tienen un 74% más de probabilidad de desertar [5].

Llontop et al., desarrollaron una plataforma web donde se propone un modelo de análisis predictivo basado en *Machine Learning* para monitorear la deserción estudiantil en universidades peruanas. Para esto, utilizaron el algoritmo *Random Forest*, evaluando 12 variables relacionadas con datos demográficos, formación preuniversitaria, entorno familiar, integración social, desempeño académico y aspectos cognitivos y emocionales del estudiante. Los resultados sugieren que las universidades pueden utilizar este tipo de herramientas para detectar tempranamente a estudiantes en riesgo y así implementar estrategias de retención más efectivas [6].

Polo Romero en su estudio presentó un modelo predictivo basado en el algoritmo *Naive Bayes* para identificar la deserción estudiantil en instituciones de educación tecnológica pública en la región La Libertad, Perú. Dicho algoritmo alcanzó una confiabilidad del 93% en sus predicciones utilizando un conjunto de datos que incluye registros de matrículas, notas y características socioeconómicas de aproximadamente 500 estudiantes desde 2010. El estudio sugiere el potencial del uso de técnicas de aprendizaje automático para anticipar y abordar la deserción estudiantil de manera efectiva en IES [7].

Valero et al., en su investigación buscaba determinar el algoritmo de *Machine Learning* que tiene mejor desempeño para detectar la deserción universitaria. Para el desarrollo de los modelos de clasificación utilizó el lenguaje de programación Python a través de sus distintas librerías. Los resultados mostraron que el algoritmo *K-Nearest-Neighbor* es el más eficaz para predecir la deserción universitaria en los primeros semestres de estudios. También, el modelo envía una notificación de alerta con los estudiantes que están en riesgo de deserción a la oficina de bienestar [8].

Bustamante y García, investigaron sobre las causas detrás de la deserción utilizando el modelo de árboles de decisión para analizar el comportamiento y rendimiento académico de los estudiantes. En dicho estudio consideran diversas variables asociadas a factores socioeconómicos y sociales. La investigación se realizó con datos proporcionados por la Facultad de Ingeniería de la Universidad Jorge Tadeo Lozano. Los resultados muestran que los factores identificados permiten tomar decisiones oportunas dentro de la gestión académica para mejorar el apoyo a los estudiantes [9].

En general, los estudios presentados evidencian la tendencia del uso de modelos predictivos para abordar la deserción estudiantil en IES de América Latina. Se observa el uso de diversas técnicas de aprendizaje automático, como la regresión logística, *Random Forest*, *Naive Bayes* y *K-Nearest-Neighbor*, con resultados prometedores en la predicción de la deserción. Además, las variables que se seleccionan para el análisis varían según el contexto y las características de la población estudiantil generando efectos combinados que pueden aumentar o disminuir el riesgo de deserción.

III. METODOLOGÍA

Se han propuesto cuatro etapas para el desarrollo de este trabajo, a saber: Fase 1: Revisión y Planeación; Fase 2: Análisis y Diseño; Fase 3: Desarrollo del Producto y Fase 4: Ejecución e Implementación. Ver Fig. 1.

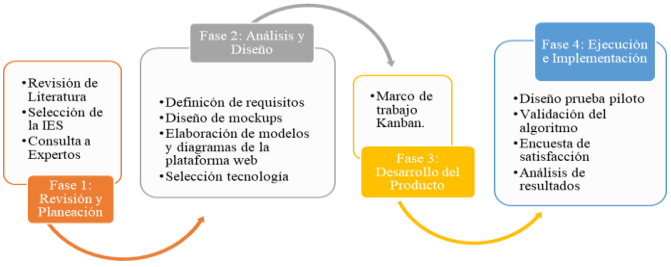

Fig. 1 Metodología.

A. *Fase 1 Revisión y Planeación*

Esta fase se centra en identificar las variables de entrada que se han utilizado en estudios similares para realizar análisis predictivos de la deserción estudiantil en una IES en América Latina. Seguidamente, se selecciona la IES en la cual se realizará la prueba piloto. Luego, docentes y administrativos de dicha institución seleccionan las variables que tienen mayor influencia en la deserción estudiantil, a partir del conocimiento que han adquirido en sus áreas de trabajo.

B. *Fase 2 Análisis y Diseño*

En esta fase se realizó una revisión de la base de datos de la IES seleccionada para determinar cuáles variables de las identificadas en la primera fase estaban disponibles y accesibles para el análisis. Luego, cuando se definen las variables entradas, según su disponibilidad y accesibilidad, se preparan los datos para la extracción, transformación y limpieza. En esta fase también se definen los requisitos funcionales y no funcionales para el desarrollo de la plataforma web. Dichos requisitos se convierten en historias de usuarios que luego serán utilizadas en la fase tres para el desarrollo del producto. Además, se diseñan los mockups utilizando la herramienta Figma [10], se elabora el diagrama de arquitectura utilizando UML (*Unified Modeling Language*), y los diagramas de secuencia. También, se elabora el diagrama

de flujo de datos para representar cómo la información se mueve a través de la plataforma web.

C. Fase 3 Desarrollo del producto

Para el desarrollo de la plataforma web se utilizó el método Kanban. Dicho marco de trabajo permite llevar una gestión flexible y visual de las tareas, asegurando que todo el flujo de trabajo sea adaptable a cambios en el proceso [11]. Se definen los roles que utilizarán los miembros del equipo y se asignará a éstos las historias de usuario definidas en la fase 2 utilizando la plataforma de gestión de proyectos *Azure Boards* [12].

D. Fase 4 Ejecución e Implementación

En esta fase se diseña la prueba piloto para evaluar el comportamiento de la plataforma web y del algoritmo de predicción. También, se valida el algoritmo para conocer la cantidad de aciertos positivos y negativos de la predicción del modelo sobre los datos de prueba. Además, se define un cuestionario para evaluar la satisfacción de los participantes frente a la experiencia con la plataforma.

IV. RESULTADOS

A. Fase 1 Revisión y Planeación

Se realizó una revisión sistemática de literatura (RSL) para identificar los estudios que han realizado análisis predictivo de la deserción estudiantil en IES en América Latina. De dichos estudios se identificó el año (últimos 5), el modelo estadístico utilizado para el análisis de los datos, la tecnología que apoyó dicho proceso y las variables utilizadas. Como resultado de este proceso se identificaron 10 estudios que se detallan en la Tabla 1.

TABLA I
Artículos seleccionados

No	Nombre	Año	Modelo Estadístico	Tecnología	Cantidad Variables	País
1	Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior tecnológica [5]	2021	Regresión Logarítmica	R	7	Ecuador
2	Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando machine learning en la educación superior universitaria del Perú [6]	2024	Regresión Lineal	Python	12	Perú
3	Modelamiento de la deserción universitaria en la Universidad Cooperativa de Colombia sede Villavicencio mediante algoritmos de machine learning [13]	2024	Random Forest	Python	43	Colombia
4	Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena [14]	2020	Regresión logística	SPSS	17	Chile
5	Modelo predictivo basado en Naive Bayes a través de machine learning supervised y la deserción estudiantil, en centros de Educación Tecnológicos públicos de la región La Libertad [7]	2024	Naive Bayes	Python	13	Perú
6	Un análisis multinomial y predictivo de los factores asociados a la deserción universitaria[15]	2019	Modelo de Regresión Multinomial	SPSS	20	Costa Rica
7	Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica [16]	2021	Arboles de decisión KNN	Python	16	Perú
8	Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción [8]	2022	K-Nearest-Neighbor	Python	8	Perú
9	Modelo predictivo de rendimiento académico para el apoyo, prevención y disminución de la tasa de deserción universitaria [9]	2021	Clasificador Bayesiano Árboles de decisión Regresión logística Random Forest	Python	16	Colombia
10	Modelo de regresión logística de la deserción estudiantil en un instituto tecnológico en el Cantón Daule [17]	2020	Regresión Logística	SPSS	10	Ecuador

Para la predicción de la deserción estudiantil, las investigaciones revisadas en promedio utilizaron 16 variables. Un 60% de las investigaciones optaron por el lenguaje de programación Python como herramienta para el desarrollo de modelos predictivos. Dentro de las técnicas estadísticas empleadas, la regresión logística fue la más utilizada, representando el 40% de los estudios. En cuanto a la distribución geográfica de los estudios, Perú concentra el mayor porcentaje de investigaciones (40%), seguido por Ecuador y Colombia, ambos con un 20%, Chile y Costa Rica representan un 10% cada uno.

A partir de la revisión de dichos estudios, se identificaron 162 variables relevantes (Tabla 1). Mediante un proceso de análisis y agrupación, que consideró la repetición de variables, similitud en su definición y valores que almacenaban, se logró reducir este conjunto a 58 variables. Estas variables conformaron la base de un cuestionario diseñado para

docentes y administradores de la institución seleccionada para el estudio. El cuestionario solicitaba a los participantes seleccionar, desde su experiencia profesional, aquellas variables que consideraban ellos más influyentes en la deserción estudiantil. Adicionalmente, se incluyó una pregunta abierta para que estos pudieran adicionar otras variables que pudieran ser relevantes y no estuvieran contempladas en el listado inicial.

Tras la identificación de variables relevantes en el cuestionario, se realizó una consulta al Departamento de Sistemas e Informática de la institución seleccionada para determinar la disponibilidad de dichas variables. A partir de este proceso, se definió un conjunto de 33 variables para el análisis de datos.

Las variables de entrada para el modelo de predicción seleccionadas se describen en la Tabla 2.

TABLA II

Variables seleccionadas

Variable	Descripción
semestre académico	Semestre actual. Año {1, 2}
facultad	Nombre de la facultad. Existen 5 facultades
programa	Nombre del programa que cursa. Son alrededor de 33 programas
fecha de nacimiento	Fecha de nacimiento.
edad	Edad. Numérico
género	0=Masculino, 1= Femenino. 0, 1
estado civil	Estado civil. Categórico
religión	Religión que profesa. Categórico
municipio de procedencia	Municipio. Categórico
tipo de población que proviene	Tipo de población. Categórico
estrato socioeconómico	Estrato. entre 1 y 6
vulnerable	Vulnerable. si, no
grupo étnico	Nombre del grupo. Alrededor de 30 grupos
víctima conflicto	Conflicto. si, no
necesidades educativas	Necesidades. si, no
discapacidad	Discapacidad. Alrededor de 7 discapacidades
capacidad excepcional	Excepcional. Alrededor de 6
país Frontera	Frontera. País con el cual limita
Sisbén	Sistema que permite clasificar a la población de acuerdo con sus condiciones de vida e ingresos. A, B, C, D
bachillerato	Validó bachillerato. si, no
cantidad integrantes del grupo familiar	Grupo familiar. Numérico
aportante familiar	Aportante familiar. si, no
nivel educativo del padre	Nivel padre. Niveles educativos
nivel educativo de la Madre	Nivel madre. Niveles educativos
número de hermanos	Hermanos. Rango
número de personas a cargo	Cantidad de personas a cargo. Numérico
trabajaba actualmente	Trabaja. si, no
residente interno	Interno. si, no
promedio de notas en materias del programa académico	Promedio notas programa. [0-5] con un decimal
promedio de notas en matemáticas	Promedio notas en asignaturas de matemáticas básicas y aplicadas. [0-5] con un decimal
promedio de notas en materias de religión	Promedio notas en asignaturas de religión. [0-5] con un decimal
promedio general	Promedio general. [0-5] con un decimal
desertor	Variable dependiente 0, 1. 0=no desertó, 1=desertó

B. Fase 2 Análisis y Diseño

Se realizó un análisis de los requisitos del sistema, identificando las necesidades de los usuarios (decanos, coordinadores y personal administrativo) y los objetivos del proyecto. En la Tabla III se detallan los requisitos funcionales y no funcionales para el desarrollo de la plataforma web. Dichos requerimientos posteriormente se convirtieron en historias de usuarios para la asignación de actividades que debían desarrollar los integrantes del equipo de desarrollo.

TABLA III

Requisitos principales del sistema

Tipo de Requerimientos	Descripción
Funcionales	• La plataforma debe acceder y procesar la información de los estudiantes almacenada en el Sistema de Información Organizacional Universitario, incluyendo datos personales, académicos y socioeconómicos. • La plataforma debe implementar un modelo predictivo que calcule un índice de riesgo de deserción para cada estudiante, utilizando variables personales, académicas y socioeconómicas. • La plataforma debe permitir la visualización de gráficas que muestren el riesgo de deserción por facultad y por variable, facilitando el análisis de la información. • La plataforma debe ser capaz de realizar análisis predictivo utilizando la información disponible para identificar patrones de deserción y predecir el riesgo de deserción de los estudiantes. • La plataforma debe permitir la exportación de reportes en formatos como Excel, para facilitar el análisis y la difusión de la información. • La plataforma debe permitir a los decanos y coordinadores académicos analizar la información básica de los estudiantes en riesgo de deserción, con el fin de implementar estrategias de intervención. • La plataforma debe mostrar la información del estudiante, incluyendo su código estudiantil, para facilitar su identificación y localización.
No funcionales	• El sistema debe ser capaz de manejar al menos un millón de datos de estudiantes simultáneamente sin que el rendimiento se vea afectado. • La interfaz de usuario debe ser intuitiva y fácil de entender para los usuarios. • El sistema debe ofrecer una versión responsiva que se adapte a dispositivos móviles. • El sistema debe ser compatible con los principales navegadores web. • El sistema debe ser compatible con la infraestructura de los servidores de la institución seleccionada. • El sistema debe soportar la instalación de entornos y bases de datos aprobados por Sistema de Información de la institución. • El sistema debe cumplir con las normativas de seguridad establecidas para la protección de datos de los estudiantes.

También, se diseñaron los mockups de la interfaz de usuario en Figma (Fig. 2), teniendo en cuenta los requisitos identificados, la usabilidad y la accesibilidad a la plataforma. Se definieron los elementos visuales, la distribución de la información y la navegación entre las diferentes secciones del sistema.

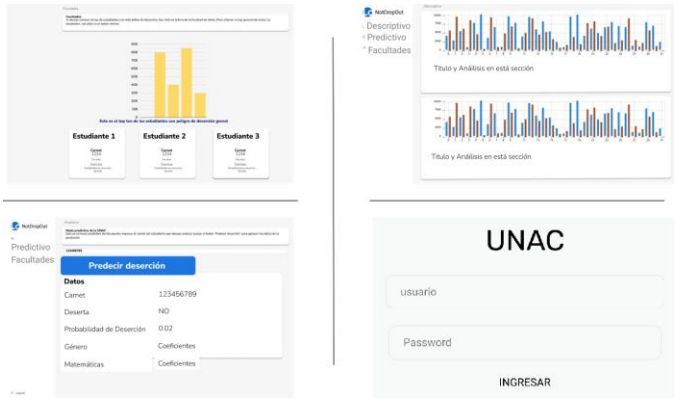

Fig. 2 *Mockups* de la plataforma.

Además, se elaboraron los diagramas de casos de uso y el diagrama de arquitectura. El diagrama de casos de uso (Fig. 3) ilustra las interacciones entre los actores principales (Coordinador, Decano y Administrador del sistema) y el sistema de predicción de deserción estudiantil. En este sentido, el Coordinador puede ver las predicciones de deserción de los estudiantes del programa académico que dirige. El Decano de la facultad puede visualizar las predicciones de deserción de todos los programas que hacen parte de la facultad que administra. Por su parte, el Administrador del sistema, además de las funciones del Coordinador y Decano, puede descargar

datos, lo que sugiere la posibilidad de realizar análisis externos o generar informes personalizados.

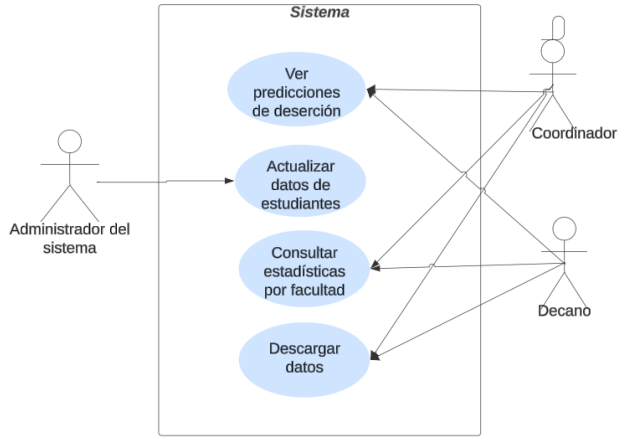

Fig. 3 Diagrama de casos de uso.

La Fig. 4 ilustra el diagrama de arquitectura el cual ilustra la interacción entre el usuario, la aplicación y la base de datos, así como los distintos módulos y funcionalidades del sistema. El usuario interactúa a través de la aplicación, desarrollada con *React*, la cual se comunica con un servidor que alberga la lógica de negocio y los modelos predictivos. Tras la autenticación del usuario, el sistema dirige la solicitud según la funcionalidad requerida, ya sea descriptiva, predictiva o de consulta por facultad.

Cada módulo del sistema accede a la base de datos MySQL para recuperar información relevante. Los resultados son presentados al usuario a través de la interfaz de la aplicación, permitiéndole visualizar la información de interés, como datos top de estudiantes por facultad y predicciones de deserción personalizadas. Adicionalmente, se contempla la descarga de datos en formato Excel para un análisis más profundo.

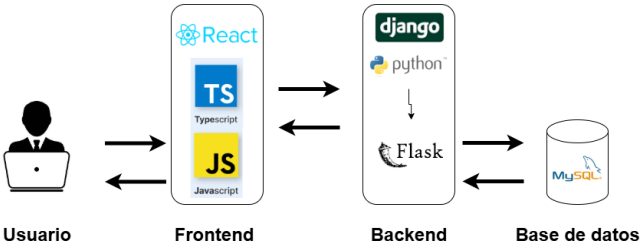

Fig. 4 Diagrama de arquitectura del sistema.

También, se implementó el patrón de diseño MVC (Modelo-Vista-Controlador) para separar las responsabilidades de cada componente del sistema [18]. En la capa modelo, se definió la estructura para el almacenamiento de los datos en el motor MySQL [19]. La capa vista, fue construida con *React* y JavaScript [20], con el propósito de obtener una interfaz de usuario dinámica e interactiva. Además, en la capa controlador se utilizó Python y *Flask* [21] para actualizar el modelo y la vista según la entrada del usuario.

Adicionalmente, el equipo de desarrollo se organizó en dos roles principales: *Front-end*, a cargo del diseño e implementación de la interfaz de usuario utilizando *React* y *TypeScript*, y *Back-end*, responsable de la lógica de negocio, gestión de la base de datos y creación de la API utilizando Python y *Django*.

C. *Fase 3: Desarrollo del producto*

De acuerdo con los roles definidos en la fase anterior el equipo de desarrollo gestionó las tareas del proyecto utilizando el método Kanban para facilitar la gestión del proyecto y el seguimiento de las incidencias, se utilizó la plataforma *Azure Boards*. En dicha plataforma, se registraron y categorizaron las incidencias, que abarcaban tanto el diseño del modelo predictivo como el desarrollo de la plataforma web.

En este sentido, inicialmente el esfuerzo de los miembros del equipo se enfocó en el desarrollo del modelo predictivo que tiene como objetivo predecir si un estudiante desertará o no. Durante el proceso de construcción, se diseñó un diagrama que muestra el flujo de datos en la plataforma. Dicho flujo incluye la preparación de los datos, la selección de variables y la construcción del modelo (Fig. 5).

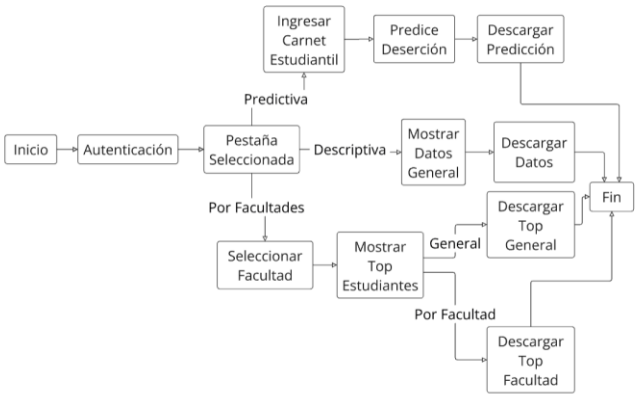

Fig. 5 Flujo de datos en la plataforma.

Los datos fueron entregados por el departamento de sistemas en varios archivos de formato CSV por sus siglas en inglés (*Comma Separated Values*). A partir de estos archivos, se definieron las tablas necesarias para la inserción de los datos, estableciendo columnas y el proceso de arquitectura de la base de datos. Dado que se requería la depuración de más de 270,000 registros, se desarrollaron varios scripts en Python para facilitar el filtrado y limpieza de los datos utilizando *Google Colab* [22].

El proceso anterior iniciaba con la carga del archivo y la eliminación de registros con valores nulos, descartando aquellos campos incompletos. A continuación, se agregaban nuevas columnas al *dataframe* donde se calcularon los valores para las variables promedio de notas en materias del programa académico, el promedio de notas en matemáticas, el promedio de notas en materias de religión y el promedio general. Dichos scripts, ayudaron a la depuración y organización de datos más eficiente y rápida.

Además, las variables numéricas se estandarizaron para que todas estuvieran en la misma escala. Esto con el propósito de mejorar el rendimiento del modelo de regresión logística. Se utilizó la función *train_test_split* de *sklearn* en Python para dividir el conjunto de datos en un 80% de entrenamiento y 20% de prueba. Seguidamente, se construyó el modelo estadístico utilizando una función de regresión logística. Dicho modelo fue seleccionado por su capacidad de interpretación y su ajuste natural a problemas de clasificación binaria.

La Ecuación (1) muestra el modelo de regresión logística desarrollado para predecir la probabilidad de deserción estudiantil. En dicha ecuación, se observa que las variables predictoras influyentes en la deserción estudiantil son: el género, la clasificación según el Sisbén, el promedio general, el promedio de notas en materias de matemáticas del programa, el promedio de notas en materias de religión y el promedio de notas en materias del programa académico. Los valores de las variables género, Sisbén y el promedio general influyen positivamente en la predicción de la deserción estudiantil. Además, los valores del promedio de las notas del estudiante en las materias de matemáticas, religión y del programa influyen negativamente la probabilidad de deserción de un estudiante. Es decir, cuando las notas en los promedios de estas últimas variables aumentan, disminuye el riesgo de deserción.

$$p(Y=1)=\frac{e^{0.74+0.16(genero)+0.03(sisben)+0.01(promgen)-0.30(prommate)-0.33(promrelig)-0.66(promprog)}}{1+e^{0.74+0.16(genero)+0.03(sisben)+0.01(promgen)-0.30(prommate)-0.33(promrelig)-0.66(promprog)}} \quad (1)$$

Después de preparar los datos, el modelo se entrenó, con el conjunto de datos de entrenamiento, utilizando las funciones *X_TRAIN* y *Y_TRAIN* de la biblioteca *sklearn*. Seguidamente, se evaluó el modelo con el conjunto de datos para las pruebas utilizando las funciones *X_TEST* y *Y_TEST*. La matriz de confusión obtenida en la evaluación mostró un resultado positivo del 97%, lo que indica una alta capacidad del modelo para clasificar correctamente las instancias.

Con el fin de integrar el modelo predictivo en la plataforma web éste se almacenó en un archivo donde se guarda un objeto serializado en Python (*pkl*), utilizando la biblioteca *pickle*. También, se dispuso a través de una API (Interfaz de Programación de Aplicaciones), para facilitar la integración de éste con las otras funcionalidades del sistema. Esto significa, que a través de una solicitud se pasan los datos de un estudiante específico y se obtiene la predicción de deserción y los coeficientes de las variables influyentes de dicho estudiante.

De acuerdo con los requisitos definidos, los diagramas diseñados y la información analizada en la fase dos se desarrolló la plataforma web. Dicha plataforma contiene dos módulos principales: módulo descriptivo y módulo predictivo (Fig. 6).

Fig. 6 Módulos principales del sistema.

En el módulo descriptivo se muestra a través de diversas gráficas la deserción de los estudiantes por género, Sisbén, nivel educativo de los padres, etnia, país de procedencia y el año (Fig. 7).

Fig. 7 Módulo descriptivo.

La Fig. 8 muestra la interfaz del módulo predictivo, diseñado para visualizar el estado actual de un estudiante en relación con el riesgo de deserción. A través de la consulta por número de carnet estudiantil, el sistema presenta una predicción sobre la probabilidad de que éste abandone sus estudios.

Adicionalmente, se muestra una tabla con las variables que influyeron en la predicción, junto con sus respectivos coeficientes, lo que permite comprender los factores de riesgo que inciden en la predicción.

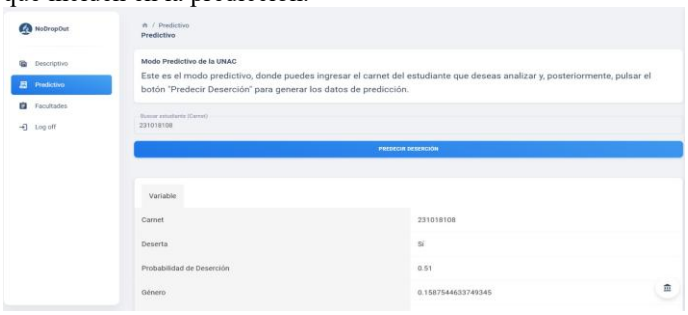

Fig. 8 Módulo predictivo.

La Fig. 9 presenta la interfaz para la sección llamada Facultades, la cual hace parte del módulo Predictivo. Esta sección permite visualizar la cantidad de estudiantes con riesgo de deserción por facultad. Además, la plataforma muestra los primeros 10 estudiantes con mayor índice de deserción por facultad (Fig. 10).

Fig. 9 cantidad de estudiantes en riesgo por facultad.

Fig. 10 top 10 en riesgo de deserción por facultad.

D. Fase 4: Ejecución e Implementación

Se llevó a cabo una prueba piloto con el objetivo de detectar los posibles fallos o problemas en la plataforma web. Aunque un estudio piloto no puede eliminar todos los errores sistemáticos o problemas inesperados, se reduce la probabilidad de errores que harían de la investigación una pérdida de esfuerzo, tiempo y dinero [23].

La plataforma web se implementó en un servidor web Apache dispuesto por la institución, el cual opera bajo el sistema operativo Linux Ubuntu. En dicho servidor, se creó la base de datos para asegurar la correcta carga de la información. Luego, se ejecutaron una serie de scripts que automatizaron el proceso de inserción y verificación de los datos en la base de datos. Adicionalmente, se instalaron las herramientas y librerías necesarias para la ejecución del front-end (*React* y *TypeScript*) y el *back-end* (Python y *Flask*) del sistema, para garantizar así la correcta comunicación e interacción entre la interfaz de usuario y la base de datos.

Tras la implementación de la plataforma y puesta en marcha, se realizó una encuesta de satisfacción a los interesados (4 decanos y 6 coordinadores) con el objetivo de recopilar información sobre la experiencia de usuario frente al uso de la plataforma e identificar áreas de mejora. La consistencia interna de las respuestas, evaluada mediante el Alfa de Cronbach, fue de 0.94, indicando una alta confiabilidad de la encuesta.

Los resultados de la encuesta la cual contenía 11 preguntas con una escala de respuesta: positivo, neutro y negativo. En términos generales la plataforma obtuvo una buena aceptación, ya que muestra un porcentaje alto en las respuestas con tendencia a positivo.

Los resultados de las preguntas asociadas al diseño visual de la plataforma muestran que el 60% de los participantes tienen una percepción positiva del diseño de la plataforma. En cuanto al rendimiento de la plataforma el 90% de los encuestados calificaron los tiempos de carga positivamente. Además, el 80% de los participantes indicó que fue fácil encontrar la información relacionada con la deserción de los estudiantes en su facultad y que las estadísticas y análisis disponibles en la plataforma son útiles para la toma de decisiones.

También, el 70% de los encuestados encontraron las gráficas intuitivas, mientras que un 30% las encontró neutrales. La totalidad de los encuestados estuvo de acuerdo en que la plataforma permite comprender el estado actual de la deserción en la institución. Así como, el 70% de los encuestados consideró que el contenido cubre adecuadamente los aspectos que esperaban sobre la deserción estudiantil. Sin embargo, sólo un 30% de los encuestados reportaron haber experimentado errores o problemas técnicos mientras navegaban en la plataforma.

V. DISCUSIONES

Los estudios de deserción estudiantil en América Latina muestran la tendencia en el uso de modelos predictivos y técnicas de aprendizaje automático para comprender mejor cómo se comporta este fenómeno y tomar decisiones más acertadas para disminuir la deserción estudiantil. Dichas investigaciones emplean diversos métodos como la regresión logística, *Random Forest*, *Naive Bayes* y *K-Nearest-Neighbor*, para analizar las variables influyentes y predecir la probabilidad de deserción.

En este sentido, los estudios analizados resaltan la importancia de considerar un enfoque multifactorial, que incluya variables demográficas, académicas, socioeconómicas y psicosociales,

para obtener modelos más precisos y completos. Si bien los estudios muestran resultados prometedores en la predicción de la deserción, es importante tener en cuenta las limitaciones de cada modelo y la necesidad de adaptarlos a las particularidades de cada contexto.

Al igual que en el estudio de Llontop et al. [6] y Valero et al., [8] esta investigación seleccionó Python como la herramienta adecuada para el desarrollo del modelo predictivo y el desarrollo de la plataforma web. La elección de la regresión logística como técnica de modelado coincide con la utilizada por Lopez y Morocho Valarezo, lo que se justifica por su interpretabilidad y capacidad para problemas de clasificación binaria.

La selección de 33 variables para el modelo de predicción de la deserción estudiantil resultó en un alto rendimiento, con una precisión del 97% en la clasificación de los datos de prueba. Este resultado sugiere que las variables seleccionadas capturan los datos relevantes para la predicción de la deserción, y que el modelo es capaz de predecir con exactitud la probabilidad de deserción de un estudiante. Sin embargo, se sugiere que se podría explorar la inclusión de variables adicionales para mejorar la precisión del modelo.

Se observar que los coeficientes para las variables de promedio de notas en matemáticas, religión y materias del programa son negativos. Esto indica que un mayor promedio en estas áreas se asocia con una menor probabilidad de deserción. Este hallazgo resalta la importancia del rendimiento académico en áreas clave para la permanencia estudiantil. Por otro lado, el coeficiente positivo para la variable "género" sugiere que, en este contexto, ser hombre aumenta la probabilidad de deserción en comparación con ser mujer. Este resultado indica que se debe investigar las causas de esta diferencia y diseñar estrategias de apoyo específicas para los estudiantes hombre. Además, el coeficiente positivo de la variable Sisbén, aunque pequeño, indica que un mayor nivel de vulnerabilidad socioeconómica se asocia con una mayor probabilidad de deserción. Este hallazgo refuerza la importancia de implementar políticas de apoyo a estudiantes de bajos recursos para garantizar la equidad en el acceso y la permanencia en la educación superior

Al igual que en el estudio de Llontop et al., se desarrolló una plataforma web para visualizar y analizar la deserción estudiantil. Esto facilita el acceso a la información y promueve la toma de decisiones a los diferentes interesados. La plataforma web incluye funcionalidades similares a las reportadas en otros estudios, como la visualización de estadísticas y la predicción de la deserción para estudiantes individuales. Sin embargo, se podría explorar la incorporación de nuevas funcionalidades, como la generación de alertas tempranas o la recomendación de estrategias de intervención personalizadas.

Al igual que en esta investigación, otros estudios han validado sus modelos y plataformas a través de pruebas piloto y encuestas de satisfacción. Esto permite evaluar la efectividad y la usabilidad de las herramientas desarrollada. Los resultados muestran que la plataforma web desarrollada facilitó la visualización, el análisis y la predicción de la deserción en la institución y apoyó la toma de decisiones y la implementación de estrategias de retención.

También, en la evaluación de la satisfacción de la plataforma web los tiempos de carga de los datos fueron valorados positivamente al igual que la facilidad para encontrar información. La mayoría de los encuestados encontró la información útil y consideró que la plataforma ayuda a comprender el estado actual de la deserción en la universidad. Sin embargo, se sugiere la necesidad de optimizar el sistema para garantizar una mejor experiencia para todos los usuarios especialmente en la interpretación de los resultados de las variables influyentes del modelo.

CONCLUSIONES

La deserción estudiantil es un problema persistente en las instituciones de educación superior de América Latina, que afecta la eficiencia del sistema educativo. Diversas variables inciden en la deserción estudiantil, asociadas a factores demográficos, académicos, socioeconómicos y psicosociales. La influencia de dichas variables puede variar según el contexto sociocultural, económico y académico de cada institución. Sin embargo, comprender la interacción de estas variables en cada contexto específico es fundamental para diseñar estrategias efectivas de prevención y retención estudiantil.

La revisión de la literatura sobre deserción estudiantil en IES en América Latina muestra el creciente uso de modelos predictivos y técnicas de aprendizaje automático para identificar estudiantes en riesgo y predecir la probabilidad de deserción. La regresión logística se identifica como el modelo estadístico más utilizado para predecir la deserción estudiantil. Las variables de entrada que mayormente se utilizaron en los estudios fueron la edad, el género, el estado civil, el promedio académico y los ingresos económicos personales o del grupo familiar al que pertenece el estudiante. También, se logró identificar que un mayor número de variables de entrada en el modelo predictivo aumenta la probabilidad de acierto, y una mayor cantidad de registros mejora el entrenamiento del modelo.

La integración del modelo predictivo a una plataforma web permite a los decanos, coordinadores y administrativos interactuar mejor con la información y tomar decisiones oportunas con relación a la deserción estudiantil de sus estudiantes. La visualización de los gráficos en dicha plataforma ayuda a resumir la información de la deserción en la institución, lo que facilita la comprensión del fenómeno y la identificación de áreas de intervención. En cuanto a la satisfacción con la plataforma, los resultados de la encuesta revelaron que la mayoría de los encuestados estaban satisfechos con el diseño visual de la plataforma y encontraron las visualizaciones claras y útiles. Sin embargo, se sugieren

mejoras en la presentación de la información, la inclusión de variables adicionales y la resolución de problemas técnicos.

Como trabajo futuro, se propone un análisis más exhaustivo que incluya la depuración de variables irrelevantes y la incorporación de variables adicionales, como el nivel académico del estudiante y los aspectos psicológicos de los estudiantes, para refinar el modelo predictivo y obtener una comprensión más completa de los factores que influyen en la deserción estudiantil.

REFERENCES

[1] A. Amaya-Amaya, F. Huerta-Castro, y C. O. Flores-Rodríguez, "Big data, a strategy to prevent academic dropout in heis", Revista Iberoamericana de Educacion Superior, vol. 11, núm. 31, 2020, doi: 10.22201/iisue.20072872e.2020.31.712.

[2] F. L. Oswald, "Developing a Biodata Measure and Situational Judgment Inventory as Predictors of College Student Performance", 2004. doi: 10.1037/0021-9010.89.2.187.

[3] SPADIES, "El Ministerio de Educación Nacional presenta la nueva versión del Sistema para la Prevención y Análisis de la Deserción en las Instituciones de Educación Superior -SPADIES 3.0", Men, pp. 1–5, 2023.

[4] D. Barragán Diaz y L. Patiño Garzón, "Elementos para la comprensión del fenómeno de la deserción universitaria en Colombia. Más allá de las mediciones", Cuadernos Latinoamericanos de Administración, vol. 9, núm. 16, 2016, doi: 10.18270/cuaderlam.v9i16.1248.

[5] C. N. Vinueza Lopez y E. F. Loza Aguirre, "Diseño de un modelo matemático para estimar la deserción estudiantil mediante técnicas de análisis multivariado en una institución de educación superior Tecnológica", 2021.

[6] A. A. Jesús LLontop, O. A. Jimenez Ramirez, y R. Pérez Pichos, "Modelo de análisis predictivo para el monitoreo de la deserción estudiantil aplicando machine learning en la educación superior universitaria del Perú", 2024.

[7] V. J. Polo Romero, "Predictive model based on Naive Bayes through Supervised Machine Learning and student dropout, in public Technological Education centers in the La Libertad region", Revista Ciencia y Tecnología, vol. 20, núm. 4, pp. 59–71, dic. 2024, doi: 10.17268/rev.cyt.2024.04.05.

[8] J. E. Valero, Á. F. Navarro, y A. C. Larios, "Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción", Como citar APA, vol. XXVIII, núm. 3, pp. 362–375, 2022.

[9] D. L. Bustamante Peña y O. García Bedoya, "Modelo predictivo de rendimiento académico para el apoyo, prevención y disminución de la tasa de deserción", 2021.

[10] Figma Design, "Figma: the collaborative interface design tool", 2017.

[11] M. O. Ahmad, D. Dennehy, K. Conboy, y M. Oivo, "Kanban in software engineering: A systematic mapping study", Journal of Systems and Software, vol. 137, 2018, doi: 10.1016/j.jss.2017.11.045.

[12] J. Cool, "Introducing Azure DevOps", Microsoft Azure, 2018.

[13] M. A. Galvis Martínez, "Modelamiento de la deserción universitaria en la Universidad Cooperativa de Colombia sede Villavicencio mediante algoritmos de machine learning", 2024.

[14] N. Henriquez Cabezas y D. Vargas Escobar, "Modelos predictivos de rendimiento y deserción académica en estudiantes de primer año de una universidad pública chilena", Revista de Estudios y Experiencias en Educación, vol. 21, núm. 45, pp. 299–316, abr. 2022, doi: 10.21703/0718-5162.v21.n45.2022.015.

[15] T. Fernández-Martín, M. Solís-Salazar, M. T. María Teresa, y T. E. Moreira-Mora, "A multinomial and predictive analysis of factors associated with university Dropout", Revista Electronica Educare, vol. 23, núm. 1, oct. 2019, doi: 10.15359/ree.23-1.5.

[16] K. Rivera Vergaray, "Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica Predictive model for the early detection of students with high risk of academic dropout", Revista Innovación y Software, vol. 2, núm. 2, 2021.

[17] K. M. Morocho Valarezo, "Modelo de regresión logística de la deserción estudiantil en instituto tecnológico en el Cantón Daule", 2020.

[18] Sadika, "The MVC Architecture", 2023.

[19] Interfell, "7 razones para elegir MySQL como gestor de base de datos", 2023.

[20] E. Britanica, "Qué es React y para qué sirve: ventajas y desventajas, casos de uso, características, quién lo utiliza", 2023.

[21] MDN contributors, "Introducción a Django - Aprende sobre desarrollo web | MDN", 2020.

[22] W. Vallejo, C. Díaz-Uribe, y C. Fajardo, "Google Colab and Virtual Simulations: Practical e-Learning Tools to Support the Teaching of Thermodynamics and to Introduce Coding to Students", ACS Omega, vol. 7, núm. 8, 2022, doi: 10.1021/acsomega.2c00362.

[23] M. González Mares, "Hernández-Sampieri, R. & Mendoza, C (2018). Metodología de la investigación. Las rutas cuantitativa, cualitativa y mixta", Revista Universitaria Digital de Ciencias Sociales (RUDICS), vol. 10, núm. 18, 2019, doi: 10.22201/fesc.20072236e.2019.10.18.6.