Machine learning para identificar cuáles son las variables relevantes que causan la deserción estudiantil.

José Luis Espinoza Melgarejo, Universidad Nacional del Santa

Implementación de modelos de aprendizaje automático para predecir la deserción estudiantil en Tecsup, 2024

Implementation of machine learning models to predict student dropout at Tecsup, 2024

RESUMEN

Este trabajo tiene como objetivo principal pronosticar si un estudiante desertará o no en el 2024 en Tecsup implementando para ello distintos modelos de clasificación de *machine learning* y elegir el mejor, además de identificar cuáles son las variables relevantes que causan la deserción estudiantil. La justificación para este estudio es que, de acuerdo con la revisión de literatura, la deserción es un problema que sigue aquejando a las instituciones educativas peruanas y, por esta razón, se quieren tomar medidas preventivas para evitar que un estudiante abandone sus estudios en Tecsup.

El alcance de este estudio es descriptivo; el diseño es no experimental, transversal y descriptivo. La población está conformada por 38 835 registros de estudiantes en el periodo 2019-2022 con información de tipo personal, académica y financiera, entre las más importantes. No se llevó a cabo un muestreo para contar con la mayor cantidad de datos posible y obtener mayor precisión en la predicción. Asimismo, se usaron técnicas estadísticas como mapa de calor, histograma, gráfico de distribución, gráfica de cajas, gráfico de barras, gráfico de barras dobles y tablas; se implementaron ocho distintos modelos de clasificación mediante *Python* a través de *Jupyter Notebook* para su procesamiento.

Por otra parte, dentro de los resultados más destacados tenemos la alta correlación existente (0,92) entre las variables cantidad de cursos cursados y cantidad de cursos aprobados, por lo cual se procedió a eliminar la primera debido a que es la suma de la cantidad de cursos aprobados y cursos desaprobados. Se llevó a cabo un proceso de discretización para las variables cantidad de cursos aprobados, cantidad de cursos desaprobados, edad y estado de pago de pensión a tiempo, quedando al final con 4, 4, 9 y 2 categorías, respectivamente. Del total de 50 variables numéricas que se obtuvieron luego de un proceso de dummificación, se eligieron 36 de estas como las más relevantes en la deserción. De los ocho modelos de clasificación propuestos (regresión logística, k-NN, árbol de decisión, *random forest*, XGBoost, LightGBM, CatBoost y red neuronal multicapa), finalmente se eligió LightGBM con un valor de exactitud en el conjunto de entrenamiento de 0,9512 y un valor de exactitud en el conjunto de prueba de 0,8892.

Consecuentemente, se puede considerar al modelo LightGBM como uno adecuado para pronosticar la deserción debido a su alta capacidad de generalización por su elevado valor de exactitud en el conjunto de prueba y la ausencia de sobreajuste por su mínima diferencia entre los valores de exactitud en el conjunto de entrenamiento y prueba (0,0619). Además, este modelo posee ventajas como mayor velocidad de entrenamiento, menor uso de memoria y mayor exactitud en comparación con otros modelos de clasificación.

ABSTRACT

The main objective of this study is to predict whether a student will drop out at Tecsup in 2024 by implementing different machine learning classification models and selecting the best one, while also identifying the relevant variables that cause student dropout. The justification for this study is that, according to the literature review, dropout remains a persistent challenge for Peruvian educational institutions, and for this reason, preventive measures are needed to prevent students from abandoning their studies at Tecsup.

The scope of this study is descriptive; the design is non-experimental, cross-sectional, and descriptive. The population consists of 38,835 student records from the 2019-2022 period,

Palabras Claves
Deserción estudiantil, preprocesamiento de datos, modelos de clasificación, predicción, exactitud, aprendizaje automático, minería de datos

Key words
Student dropout, data preprocessing, classification models, prediction, accuracy, machine learning, data mining.

comprising personal, academic, and financial data. No sampling was performed in order to utilize the entire dataset and maximize prediction accuracy. Additionally, exploratory data analysis employed heat maps, histograms, distribution graphs, box plots, bar charts, double bar charts, and tables; eight different classification models were implemented using Python and Jupyter Notebook for processing.

Notably, a high correlation (0.92) was found between the variables "number of courses taken" and "number of courses passed." Therefore, the former was eliminated because it is the sum of the number of courses passed and failed. A discretization process was carried out for the variables "number of courses passed," "number of courses failed," "age," and "on-time tuition payment status," resulting in 4, 4, 9, and 2 categories, respectively. Of the total of 50 numerical variables obtained after generating dummy variables, 36 were selected as the most relevant to dropout rates. Of the eight proposed classification models (logistic regression, k-NN, decision tree, random forest, XGBoost, LightGBM, CatBoost, and multilayer perceptron), LightGBM was ultimately chosen with an accuracy of 0.9512 on the training set and 0.8892 on the test set.

Consequently, the LightGBM model can be considered suitable for predicting dropout due to its high generalization capacity—evidenced by its high accuracy on the test set—and the absence of overfitting, indicated by the minimal difference between the accuracy values on the training and test sets (0.0619). Furthermore, this model has advantages such as faster training speed, lower memory usage, and higher accuracy compared to other classification models.

INTRODUCCIÓN

La interrupción de estudios es un problema que afecta no solo al sistema educativo básico, sino también al superior a nivel mundial. De acuerdo con el Ministerio de Educación (Minedu) [18], en un informe de la Superintendencia Nacional de Educación Superior Universitaria (Sunedu) de 2022 y con información de la Encuesta Nacional de Hogares (Enaho) del mismo año, 17,6 % de los estudiantes abandonaron sus estudios debido a que sus padres no contaban con estudios superiores y 14 % cuyos padres tenían estudios superiores completos. Además, 21,9 % desertaron porque contaban con un miembro dependiente del hogar y 13,8 % que no contaban con miembros dependientes, es decir, niños o adultos mayores que requieren cuidados por parte de los estudiantes. Por último, no hubo diferencias en la deserción con respecto al género y los gastos de estos.

Algo más grave que el problema anterior es lo que se conoce como deserción estudiantil, existiendo varias definiciones para esta. Tinto (citado en [26]) afirma que la deserción es una situación en la que un estudiante no logra concretar sus proyectos educativos. Además, Tinto (citado en [26]) considera que no hay una definición exacta y sostiene que los investigadores utilizan aquella que resulte más apropiada para su estudio. Según Díaz (citado en Viale, 2014), a pesar de que no existe una definición precisa, sí hay consenso en caracterizarla mediante distintas categorías de variables como socioeconómicas, individuales, institucionales y académicas.

Source (citado en [10]) afirma que la cantidad de estudiantes que cursan educación superior se ha incrementado de 196 millones a 250 millones en el 2021, pero muchos de ellos no logran culminar sus estudios y terminan por abandonarlos. El porcentaje de deserción en Estados Unidos llegó a ser 40 %, mientras que en la Unión Europea alcanzó el 24 % para estudiantes de pregrado según Rudín en el 2019 (citado en [10]). Por otro lado, en América Latina, los porcentajes de deserción en Colombia, Perú y Chile fueron, respectivamente, 31 %, 23 % y 23 % en el 2019 según Behr (citado en [10]).

El primer año de estudios ha sido identificado como crítico en términos de deserción y un tercio de los estudiantes abandona la universidad de acuerdo con Feldman (citado en [11]). Por otra parte, Reason (citado en [11]) afirma que esta cifra llega al 62 %. Braxton (citado en [11]) concluyó que las características preuniversitarias influyen de forma directa en que un estudiante abandone la universidad.

Según Sineace (citado en [19]), de los más de 400 000 estudiantes que cursan estudios en institutos de educación superior, en Lima, con alrededor de 198 196 estudiantes, aproximadamente el 36 % no logró graduarse y cada año 90 000 abandonan sus estudios, de los cuales el 70 % pertenece a instituciones privadas y el resto a instituciones del Estado.

El objetivo de este trabajo es determinar si un estudiante desertará o no en el instituto tecnológico Tecsup en el 2024. Para ello, se implementarán distintos modelos de clasificación para realizar dicha predicción, eligiendo el mejor modelo, además de efectuar un análisis de cuáles son las características más importantes que provocan la deserción estudiantil utilizando datos de las diversas áreas involucradas dentro de la institución.

FUNDAMENTOS

En el trabajo titulado «Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018-2023», Villegas y Núñez [28] tienen como objetivo realizar una revisión de la literatura para determinar los factores más importantes que influyen en la deserción estudiantil y así reducirla, efectuando una búsqueda en el periodo 2018-2023 en WoS (52 %), Scopus (22 %) y Redalyc (26 %), cuyos resultados mostraron que dichos factores se pueden dividir en aspectos sociológicos, psicológicos y económicos. Además, Pedroza, Chindoy y Rosado [9], en su trabajo titulado «Review of techniques, tools, algorithms and attributes for data mining used in student desertion», tienen como objetivo identificar factores comunes mediante una búsqueda literaria sobre deserción estudiantil, siendo los criterios seguidos las técnicas, algoritmos, herramientas y atributos de la publicación, donde la mayoría de investigaciones se relacionan con el aprendizaje supervisado y la técnica más utilizada es la clasificación, mientras que el algoritmo predominante es J48.

Fernández y Silva [11], en su trabajo titulado «Deserción estudiantil universitaria en el primer semestre. El caso de una institución de educación superior ecuatoriana», tienen como objetivo identificar relaciones entre variables preuniversitarias y la deserción estudiantil en el primer semestre de estudio en una universidad de Ecuador, con base en 1276 registros de estudiantes censados en un diseño transversal. Concluyen que los estudiantes mayores de 20 años y que llevan más de cinco cursos desertan 2,6 veces más que los menores, y también que las variables nota de colegio y nota de examen de admisión no influyen en la deserción. Por otra parte, Aleans [2], en su trabajo titulado «Determinantes de la deserción estudiantil universitaria por niveles de formación en

instituciones de educación superior de la ciudad de Medellín», tiene como objetivo identificar los determinantes influyentes en la deserción estudiantil entre distintos grupos poblacionales en instituciones de educación superior en Medellín, utilizando datos de estudiantes que ingresaron al sistema educativo superior en el año 2006 para construir un modelo probit que identifique las variables significativas en cada nivel de formación. Concluye que las variables socioeconómicas, individuales, tipos de apoyo y aquellas relacionadas con áreas de conocimiento tienen efectos diferentes en los niveles universitario, tecnológico y técnico.

En el trabajo de Rivera [20], titulado «Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica», el objetivo es comparar cuatro modelos de minería de datos (regresión logística, árboles de decisión, k-NN y red neuronal) para predecir la deserción estudiantil de estudiantes de la Universidad Nacional Intercultural de la Amazonía, utilizando datos socioeconómicos y de rendimiento académico. Los resultados mostraron que los cuatro modelos alcanzaron una exactitud superior al 80 % y se concluyó que su aplicación es altamente beneficiosa para detectar tempranamente a posibles desertores. Asimismo, en el trabajo titulado «Factores de deserción estudiantil: Un estudio exploratorio desde Perú», Viera *et al.* [27] tienen como objetivo identificar cuáles son los factores que causan la deserción estudiantil en la Escuela de Ingeniería Industrial de la UNSA, tomando una muestra de 220 estudiantes y utilizando para ello una metodología causal basada en un modelo de regresión logística. Concluyen que los factores individuales (que incluyen las variables educación del padre y de la madre) y académicos influyen de manera directa en la deserción estudiantil; por el contrario, los factores socioeconómicos e institucionales no resultaron influyentes. Además, estos factores fueron generados mediante la técnica de análisis factorial a partir de 18 variables.
En la tesis de Camargo [7], titulada «Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos», cuyo objetivo es crear un modelo para predecir la deserción estudiantil de la Universidad de La Costa mediante distintas fases utilizando datos demográficos, culturales, sociales, familiares, educativos, entre otros, de estudiantes del periodo 2013-2018, con un total de 1606 registros, los resultados mostraron que el modelo *random forest* obtuvo un mejor desempeño con una exactitud de 84,8 %. Por otra parte, González y Arismendi [12], en su trabajo titulado «Deserción estudiantil en la educación superior técnico-profesional: Explorando los factores que inciden en alumnos de primer año», tienen como objetivo determinar las variables más influyentes en la deserción estudiantil en una institución técnico-profesional en el periodo 2014-2016. Se trata de un estudio de alcance explicativo y no experimental con una muestra de 1876 estudiantes, utilizando un modelo lineal generalizado con distribución de errores binomial y función de vínculo logit, donde se probó que las variables más importantes resultaron ser el género, el año de egreso de la enseñanza media y la jornada de estudio, aunque el modelo varía según la escuela de formación.

Alania [1], en su trabajo titulado «Aplicación de técnicas de minería de datos para predecir la deserción estudiantil de la Facultad de Ingeniería de la Universidad Nacional Daniel Alcides Carrión», tiene como objetivo aplicar datos académicos de una universidad para predecir la deserción estudiantil mediante modelos de minería de datos, siendo el modelo de árbol de decisión C4.5 (J48) el que se aplicó sobre los datos de notas finales para determinar si un estudiante deserta o no. Asimismo, se realizó una comparación con un modelo de *random forest*, logrando este último una mayor precisión en la predicción. Además, en el trabajo titulado «Modelos predictivos de la deserción estudiantil en una universidad privada peruana», Sifuentes [22] tiene como objetivo determinar que el uso de modelos predictivos en asignaturas críticas ayuda a identificar a estudiantes con intención de desertar, utilizando para ello siete modelos mediante la aplicación de la metodología para ciencia de datos CRISP-DM y datos históricos de estudiantes en siete cursos, cuyos resultados mostraron que los modelos propuestos redujeron la deserción entre 25 % y 40 %. Las variables más influyentes fueron la vocación, el número de veces que se matriculan en una asignatura y la nota obtenida en quinto de secundaria.

En el trabajo de Cuji *et al*. [8], titulado «Modelo predictivo de deserción estudiantil basado en árboles de decisión», cuyo objetivo es construir un modelo de deserción estudiantil basado en árboles de decisión para pronosticar la probabilidad de que un estudiante abandone sus estudios utilizando como metodología el Knowledge Discovery in Databases (KDD), se concluyó que el algoritmo Classification and Regression Tree (CART), con cuatro niveles de profundidad y las mismas reglas, fue el más eficiente para la predicción, donde las variables nivel y notas resultaron ser las más influyentes. Por otra parte, Vásquez [25], en su trabajo titulado «Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior», tiene como objetivo construir un modelo de predicción para detectar desertores entre estudiantes de Ingeniería en Información y Control de Gestión de la Facultad de Economía y Negocios de la Universidad de Chile, basado en la metodología KDD y tomando como referencia el estado académico registrado por la FEN a inicios del semestre 2016. Concluye que las variables más influyentes fueron el rendimiento en la PSU, el número de padres vivos, la evaluación que los alumnos realizan a los profesores de manera semestral y el rendimiento académico universitario, aunque los predictores no son los mismos en cada semestre. Además, se indica que los estudiantes deben agruparse por antecedentes familiares y rendimiento por semestre mediante técnicas de clusterización, siendo cinco de los seis modelos propuestos basados en estas.

Amaya *et al.* [3], en su trabajo titulado «Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos», tienen por objetivo construir un modelo predictivo de deserción estudiantil de estudiantes de la Universidad Simón Bolívar utilizando modelos basados en árboles de decisión como C4.5 e ID3 bajo distintos contextos y comparando la precisión de un modelo respecto del otro. Con una muestra de 201 registros y 40 variables, concluyeron que el modelo ID3 es más preciso para predecir la deserción, dado que cuenta con una mayor cantidad de reglas que el otro modelo. Además, Sposito *et al.* [23], en su trabajo titulado «Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil», tienen como objetivo evaluar el rendimiento académico y la deserción estudiantil de estudiantes del área de Ingeniería e Investigaciones Tecnológicas de la UNLaM mediante el proceso de descubrimiento de conocimiento (KDD) en el periodo 2003-2008, utilizando SQL Server para recopilar, integrar y almacenar datos, SPSS para depurar, seleccionar y transformar la información, y Weka para obtener un clasificador de minería de datos. Concluyeron que los árboles de decisión de tipo J48 y FT, basados en un conjunto de variables construidas en las fases anteriores, son adecuados, siendo el modelo FT superior al J48 para clasificar la deserción estudiantil en las categorías inactivo, activo y reincorporado.

Así, este trabajo busca predecir la deserción estudiantil en Tecsup mediante un modelo de clasificación utilizando datos del 2019

al 2022 y determinar cuáles son las variables más influyentes en dicha deserción mediante técnicas basadas en árboles de decisión.

Dentro de las limitaciones encontradas, no se pudieron obtener datos de los años 2023 y 2024 debido a trámites administrativos o porque la información correspondiente a dichos años no contenía todas las variables de estudio.

METODOLOGÍA

Tipo y diseño de investigación

Según Hernández *et al.* [15], la investigación es de tipo cuantitativa y de alcance descriptivo, ya que se busca especificar propiedades y características de un fenómeno. En este estudio se pretende pronosticar la deserción estudiantil mediante modelos de clasificación y se determinará cuáles son las variables que más influyen en ella.

De acuerdo con Hernández *et al.* [15], el diseño es no experimental, pues las variables independientes no fueron manipuladas de forma intencional para observar su efecto en la variable dependiente. Además, según los mismos autores, el estudio es transversal, porque los datos se tomaron en un solo instante de tiempo, y descriptivo, ya que se examinan los niveles de una o más variables en la población.

Se siguió la metodología de Hernández *et al.* [15] debido a que son algunos de los autores más importantes en materia de investigación, aunque, según Tam *et al.* [24], existen muchos tipos de clasificaciones y métodos de investigación.

El uso de algoritmos de aprendizaje automático es apropiado para este estudio, ya que permite identificar patrones multivariados entre factores académicos, socioeconómicos y de comportamiento, en comparación con los métodos tradicionales [6]. Asimismo, los modelos de aprendizaje se actualizan con nuevos datos, lo cual les permite adaptarse a los cambios y mejorar, con el tiempo, su rendimiento [14].

Según Han, Kamber y Pei [13], quienes definen la minería de datos como «el proceso de descubrir patrones y conocimiento relevantes a partir de grandes cantidades de datos», no emplearemos una metodología como tal, ya que la minería de datos forma parte del proceso Knowledge Discovery in Databases (KDD), que consta de las etapas limpieza de datos, integración de datos, selección de datos, transformación de datos, minería de datos, evaluación de patrones y presentación del conocimiento, aunque tomaremos este último como referencia.

Población y muestra

La población de estudio está conformada por 38 835 registros de estudiantes durante el periodo 2019-2022 y cuyas 26 variables se detallan en la tabla 1.

Tabla 1
Variables, descripción y tipo de dato de estudiantes de Tecsup durante el periodo 2019-2022

Variable	Descripción	Tipo de dato
SEMESTRE	Semestre que cursa el estudiante del 2019 a 2022	Alfanumérico
SEDE	Sede en la que cursa estudios el estudiante	Texto
CODIGO	Código Tecsup del estudiante	Alfanumérico
CODORACLE	Código identificador del estudiante	Alfanumérico
CICLO	Ciclo en el que se encuentra el estudiante del primero al sexto	Numérico
SEXO	Sexo del estudiante	Texto
CURSO_CURSADOS	Número de cursos matriculados del estudiante en el semestre en curso	Numérico
CURSO_APROBADOS	Número de cursos aprobados del estudiante en el semestre en curso	Numérico
CURSO_DESAPROBADOS	Número de cursos desaprobados del estudiante en el semestre en curso	Numérico
CURSO_DIFCICLO	Si el estudiante se matriculó en cursos de diferentes ciclos y toma como valores 0 o 1	Numérico
CURSO_FALLO_INASISTENCIA	Si el estudiante tuvo algún curso en el que no asistió en más del 30 % de las clases y toma como valores 0 o 1	Numérico
CURSO_INASISTENCIAS	Si el estudiante tuvo inasistencias en el semestre y toma como valores 0 o 1	Numérico
DESERTOR_ESTADO	Si el estudiante desertó o no y toma como valores 0 o 1	Numérico
PROMEDIO_GENERAL	Promedio general del estudiante en el semestre que cursa	Numérico

Variable	Descripción	Tipo de dato
APROBADO_ESTADO	Estado de aprobado del estudiante en el semestre que cursa y toma como valores 0 o 1	Numérico
DESAPROBADO_ESTADO	Estado de desaprobado del estudiante en el semestre que cursa y toma como valores 0 o 1	Numérico
CONCREDITO_ESTADO	Si el estudiante cuenta con un crédito y toma como valores 0 o 1	Numérico
EDAD_ANIOS	Edad del estudiante en años	Numérico
ESTADO_PAGO_A_TIEMPO_MAT	Si el estudiante pagó a tiempo la matrícula en el semestre que cursa y toma como valores 0 o 1	Numérico
ESTADO_PAT_DIAS_MAT	Número de días que el estudiante pagó con anticipación la matrícula en el semestre que cursa	Numérico
ESTADO_PAGO_TARDANZA_MAT	Si el estudiante no pagó a tiempo la matrícula en el semestre que cursa y toma como valores 0 o 1	Numérico
ESTADO_PTA_DIAS_MAT	Número de días que el estudiante se tardó en pagar la matrícula en el semestre que cursa	Numérico
ESTADO_PAGO_A_TIEMPO_PEN	Si el estudiante pagó a tiempo las pensiones en el semestre que cursa y toma como valores 0 o 1	Numérico
ESTADO_PAT_DIAS_PEN	Número de promedio de días que el estudiante pagó con anticipación las pensiones en el semestre que cursa	Numérico
ESTADO_PAGO_TARDANZA_PEN	Si el estudiante no pagó a tiempo las pensiones en el semestre que cursa y toma como valores 0 o 1	Numérico
ESTADO_PTA_DIAS_PEN	Número de promedio de días que el estudiante se tardó en pagar las pensiones en el semestre que cursa	Numérico

Fuente: Gestión de Performance, Tecsup, 2024.

Para poder contar con la mayor cantidad de datos, no se realizará un muestreo, siendo este no probabilístico por conveniencia y, según Arias-Gómez *et al.* [4], menos costoso que el muestreo probabilístico, aunque la muestra puede no ser representativa y, por ende, no sería recomendable realizar generalizaciones.

Técnicas de recolección de datos

La fuente de datos es secundaria, ya que incluye información publicada en agencias del gobierno, estatales y locales, según Malhotra (2008). Los datos fueron obtenidos de diversas áreas de Tecsup como Servicios Educativos, Finanzas y Tecnologías de Información, entre otras, y fueron gestionados por personal del área de Gestión de Performance para su acceso.

Debido a ello, no se utilizó un instrumento de recolección de datos; estos se encontraban en bases de datos relacionales y archivos en formato CSV o Excel, que posteriormente se integraron en varios archivos en formato CSV que contienen información sobre estudiantes de Tecsup con diferentes atributos durante el periodo 2019-2022.

Análisis de los datos

El análisis de la información se realizó mediante mapa de calor, histograma, gráfico de distribución, gráfica de cajas, gráfico de barras, gráfico de barras dobles y tablas para la parte descriptiva y el conteo de categorías. Se emplearán también tablas para realizar comparaciones entre los diferentes modelos de clasificación propuestos mediante sus métricas de exactitud obtenidas para los conjuntos de entrenamiento y prueba.

Se utilizó el *software* Python en su versión 3.11.0 por medio de la interfaz web Jupyter Notebook en su versión 7.0.8 de la distribución Anaconda, de la cual se emplearon librerías como pandas para crear y manipular data frames; matplotlib y seaborn para crear visualizaciones de datos; numpy para crear y manipular arreglos y emplear funciones matemáticas, y sklearn para realizar preprocesamiento de datos, implementar modelos de machine learning y calcular métricas de clasificación.

RESULTADOS

Se procedió a eliminar las variables CODIGO, CODORACLE, DESAPROBADO_ESTADO, ESTADO_PAT_DIAS_MAT, ESTADO_PAGO_TARDANZA_MAT, ESTADO_PAT_DIAS_PEN y ESTADO_PAGO_TARDANZA_PEN, puesto que los valores de las dos primeras son únicos para cada estudiante; DESAPROBADO_ESTADO contiene valores opuestos a la variable APROBADO_ESTADO, y las cuatro últimas se utilizaron para crear otras variables, quedando al final lo que se muestra en la Tabla 2.

Tabla 2
Variables, descripción y tipo de dato de estudiantes de Tecsup durante el periodo 2019-2022 luego de la primera depuración

Variable	Descripción	Tipo de dato
SEMESTRE	Semestre que cursa el estudiante (2019-1, 2019-2, 2020-1, 2020-2, 2021-1, 2021-2, 2022-1 o 2022-2)	Alfanumérico
SEDE	Sede en la que cursa estudios el estudiante (A, L o T)	Texto
CICLO	Ciclo en el que se encuentra el estudiante (1, 2, 3, 5, 5 o 6)	Numérico
SEXO	Sexo del estudiante (M o F)	Texto
CURSO_CURSADOS	Número de cursos matriculados del estudiante en el semestre en curso	Numérico
CURSO_APROBADOS	Número de cursos aprobados del estudiante en el semestre en curso	Numérico
CURSO_DESAPROBADOS	Número de cursos desaprobados del estudiante en el semestre en curso	Numérico
CURSO_DIFCICLO	Si el estudiante se matriculó en cursos de diferentes ciclos y toma como valores 0 o 1	Numérico
CURSO_FALLO_INASISTENCIA	Si el estudiante tuvo algún curso en el que no asistió en más del 30% de las clases y toma como valores 0 o 1	Numérico
CURSO_INASISTENCIAS	Si el estudiante tuvo inasistencias en el semestre y toma como valores 0 o 1	Numérico
PROMEDIO_GENERAL	Promedio general del estudiante en el semestre que cursa	Numérico
APROBADO_ESTADO	Estado de aprobado del estudiante en el semestre que cursa y toma como valores 0 o 1	Numérico
DESAPROBADO_ESTADO	Estado de desaprobado del estudiante en el semestre que cursa y toma como valores 0 o 1	Numérico
CONCREDITO_ESTADO	Si el estudiante cuenta con un crédito y toma como valores 0 o 1	Numérico
EDAD_ANIOS	Edad del estudiante en años	Numérico
ESTADO_PAGO_A_TIEMPO_MAT	Si el estudiante pagó a tiempo la matrícula en el semestre que cursa y toma como valores 0 o 1	Numérico
ESTADO_PAGO_A_TIEMPO_PEN	Si el estudiante pagó a tiempo las pensiones en el semestre que cursa y toma como valores 0 o 1	Numérico

Fuente: Elaboración propia.

Tomando en cuenta la definición de cada variable se dividieron en un grupo de variables categóricas, y otra en numéricas de acuerdo a la tabla 3.

Tabla 3
Variables y tipo de estudiantes de Tecsup durante el periodo 2019-2022

Variable y tipo	
Categórica	**Numérica**
SEMESTRE	
SEDE	
CICLO	
SEXO	
	CURSO_CURSADOS
	CURSO_APROBADOS
	CURSO_DESAPROBADOS
CURSO_DIFCICLO	
CURSO_FALLO_INASISTENCIA	
CURSO_INASISTENCIAS	
	PROMEDIO_GENERAL
APROBADO_ESTADO	
DESAPROBADO_ESTADO	
CONCREDITO_ESTADO	
	EDAD_ANIOS
ESTADO_PAGO_A_TIEMPO_MAT	
ESTADO_PAGO_A_TIEMPO_PEN	

Fuente: Elaboración propia.

Se procedió a realizar un análisis de la correlación entre las variables numéricas, con base a la información de la tabla 3.

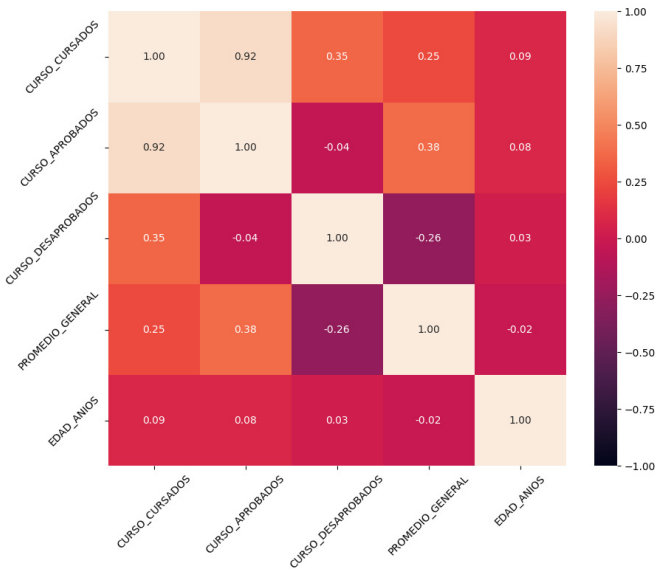

Figura 1. Mapa de calor de las variables numéricas de estudiantes de Tecsup durante el periodo 2019-2022

Fuente: Elaboración propia.

En la figura 1, podemos observar que existe una alta correlación (0,92) entre las variables CURSO_CURSADOS y CURSO_APROBADOS, por lo que procedemos a eliminar la primera debido a que esta es la suma de las variables CURSO_APROBADOS y CURSO_DESAPROBADOS.

Tabla 4
Variables discretizadas de estudiantes de Tecsup durante el periodo 2019-2022

Variable nueva	Categorías
CURSO_APROBADOS_AGRUP	0 1-3 4-6 7-9
CURSO_DESAPROBADOS_AGRUP	0 1-3 4-6 7-10
EDAD_ANIOS_AGRUP	14-19 20-24 25-29 30-34 35-39 40-44 45-49 50-54 55-59
ESTADO_PAGO_A_TIEMPO_PEN	0 1

Fuente: Elaboración propia.

Debido a la alta diferencia en las frecuencias de las variables CURSO_APROBADOS y CURSO_DESAPROBADOS, se procedió a discretizarlas en cuatro categorías para cada una. Asimismo, se construyeron nueve rangos para la variable EDAD_ANIOS y dos categorías para la variable ESTADO_PAGO_A_TIEMPO_PEN, asignándoles nuevos nombres y eliminando las variables originales, quedando como se muestra en la tabla 4.

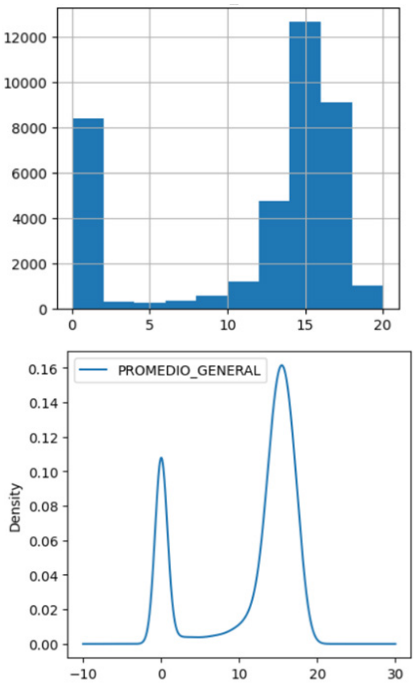

Figura 2. Histograma y gráfico de densidad de la variable PROMEDIO_GENERAL de estudiantes de Tecsup durante el periodo 2019-2022

Fuente: Elaboración propia.

En la figura 2 se observa que la distribución de datos de la única variable numérica PROMEDIO_GENERAL no es normal, puesto que la mayoría de valores se concentran a la derecha y, a la izquierda, se registran muchos valores iguales a cero (8007), lo cual corresponde no solo a estudiantes que han desertado, sino también a aquellos que se matricularon y reprobaron por inasistencias o por bajo rendimiento.

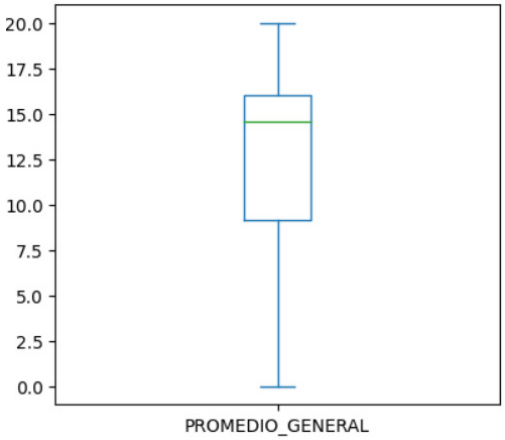

Figura 3. Diagrama de caja de la variable PROMEDIO_GENERAL de estudiantes de Tecsup durante el periodo 2019-2022

Fuente: Elaboración propia.

En la figura 3, podemos observar que no existen problemas con valores atípicos, por lo cual no será necesario utilizar alguna técnica para tratar estos.

Tabla 5
Cantidad de variables categóricas y numéricas antes y después de estudiantes de Tecsup durante el periodo 2019-2022

Tipo de variable	Cantidad	
	Antes	Después
Categórica	16	0
Numérica	1	50
Total	17	50

Fuente: Elaboración propia.

Estandarizamos la variable PROMEDIO_GENERAL y convertimos las variables categóricas en *dummies* para poder incluirlas en un modelo de clasificación. Tomando en cuenta la información de las Tablas 3 y 4, se obtuvo finalmente un total de 42 variables numéricas.

Tabla 6
Variables originales y seleccionadas mediante random forest de estudiantes de Tecsup durante el periodo 2019-2022

Tipo	Variables	Cantidad
Originales	PROMEDIO_GENERAL SEMESTRE_2019-1, SEMESTRE_2019-2, SEMESTRE_2020-1, SEMESTRE_2020-2, SEMESTRE_2021-1, SEMESTRE_2021-2, SEMESTRE_2022-1, SEMESTRE_2022-2 SEDE_A, SEDE_L, SEDE_T CICLO_1, CICLO_2, CICLO_3, CICLO_4, CICLO_5, CICLO_6 SEXO_F, SEXO_M CURSO_DIFCICLO_0, CURSO_DIFCICLO_1 CURSO_FALLO_INASISTENCIA_0, CURSO_FALLO_INASISTENCIA_1 CURSO_INASISTENCIAS_0, CURSO_INASISTENCIAS_1 APROBADO_ESTADO_0, APROBADO_ESTADO_1 CONCREDITO_ESTADO_0, CONCREDITO_ESTADO_1 EDAD_ANIOS_AGRUP_0, EDAD_ANIOS_AGRUP_1, EDAD_ANIOS_AGRUP_2, EDAD_ANIOS_AGRUP_3, EDAD_ANIOS_AGRUP_4, EDAD_ANIOS_AGRUP_5 ESTADO_PAGO_A_TIEMPO_MAT_0, ESTADO_PAGO_A_TIEMPO_MAT_1, ESTADO_PAGO_A_TIEMPO_PEN_0, ESTADO_PAGO_A_TIEMPO_PEN_1 CURSO_APROBADOS_AGRUP_0, CURSO_APROBADOS_AGRUP_1-3, CURSO_APROBADOS_AGRUP_4-6, CURSO_APROBADOS_AGRUP_7-9 CURSO_DESAPROBADOS_AGRUP_0, CURSO_DESAPROBADOS_AGRUP_1-3, CURSO_DESAPROBADOS_AGRUP_4-6, CURSO_DESAPROBADOS_AGRUP_7-10	50

Fuente: Elaboración propia.

Tipo	Variables	Cantidad
Seleccionadas	PROMEDIO_GENERAL SEMESTRE_2019-1, SEMESTRE_2019-2, SEMESTRE_2020-1, SEMESTRE_2020-2, SEMESTRE_2021-1, SEMESTRE_2021-2 SEMESTRE_2022-2 SEDE_A, SEDE_T CICLO_1, CICLO_2, CICLO_3, CICLO_4, CICLO_5, CICLO_6 SEXO_F, SEXO_M CURSO_FALLO_INASISTENCIA_0, CURSO_FALLO_INASISTENCIA_1 APROBADO_ESTADO_0, APROBADO_ESTADO_1 EDAD_ANIOS_AGRUP_0, EDAD_ANIOS_AGRUP_1, EDAD_ANIOS_AGRUP_2 ESTADO_PAGO_A_TIEMPO_MAT_0, ESTADO_PAGO_A_TIEMPO_MAT_1, ESTADO_PAGO_A_TIEMPO_PEN_0, ESTADO_PAGO_A_TIEMPO_PEN_1 CURSO_APROBADOS_AGRUP_0, CURSO_APROBADOS_AGRUP_1-3, CURSO_APROBADOS_AGRUP_4-6, CURSO_DESAPROBADOS_AGRUP_0, CURSO_DESAPROBADOS_AGRUP_1-3, CURSO_DESAPROBADOS_AGRUP_4-6, CURSO_DESAPROBADOS_AGRUP_7-10	36

Fuente: Elaboración propia.

En la tabla 6, se observan las variables más relevantes que fueron seleccionadas mediante la técnica de selección de características (o variables) basada en modelos y secuenciales [21], la cual elimina aquellas que se encuentran por debajo de un umbral (*threshold*) que debe especificarse. Además, se debe incluir un modelo para llevar a cabo este proceso, siendo el elegido *random forest*, el cual se utiliza para determinar la importancia de características con base en impurezas, resultando ideal para eliminar variables no relevantes.

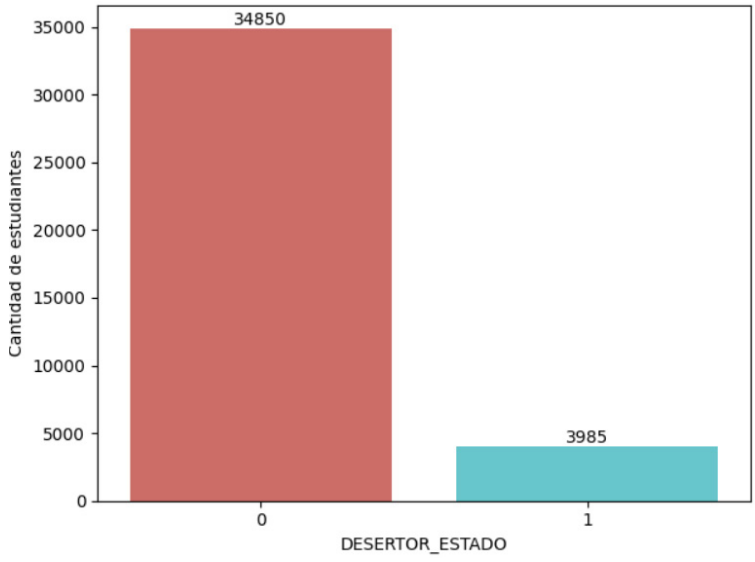

Figura 4. Diagrama de barras de las categorías 0 y 1 de la variable DESERTOR_ESTADO de estudiantes de Tecsup durante el periodo 2019-2022

Fuente: Elaboración propia.

En la figura 4, se observa que existe una mayor cantidad de estudiantes que no son desertores (34 850) frente a quienes sí lo son (3985), siendo la proporción de aproximadamente 8,74. Debido a ello, se realizó una validación cruzada con una partición de 5 (CV = 5) para compensar la disparidad entre las categorías 0 y 1 de la variable DESERTOR_ESTADO, junto con un ajuste de hiperparámetros para los ocho modelos de clasificación propuestos.

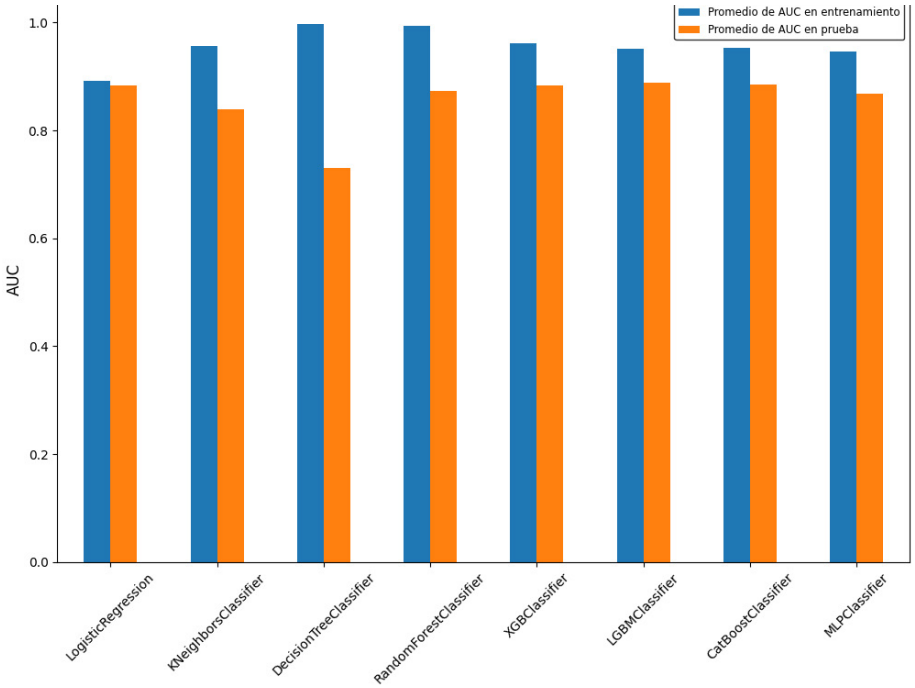

Figura 5. Diagrama de barras doble de las exactitudes de entrenamiento y prueba para los diferentes modelos de clasificación de estudiantes de Tecsup durante el periodo 2019-2022

Fuente: Elaboración propia.

En la figura 5, se observa que el modelo de regresión logística parece ser el más equilibrado, puesto que su promedio de exactitud (AUC) en el entrenamiento es mayor que en el de prueba por un margen muy reducido. Otros modelos que presentan un mejor desempeño que el anterior, aunque menos equilibrados respecto de sus promedios de exactitud en entrenamiento y prueba, son XGBoost, LightGBM y CatBoost, los cuales están basados en árboles de decisión y presentan buen rendimiento, escalabilidad y versatilidad [16].

Tabla 7

Valores promedio de las exactitudes de entrenamiento y prueba para los diferentes modelos de clasificación de estudiantes de Tecsup durante el periodo 2019-2022

Modelo	Promedio de AUC en entrenamiento	Promedio de AUC en prueba	Diferencia de promedio de AUC
Regresión logística	0,8915	0,8831	0,0084
k-NN	0,9567	0,8393	0,1173
Árbol de decisión	0,9973	0,7312	0,2660
Random forest	0,9930	0,8715	0,1215
XGBoost	0,9620	0,8825	0,0795
LightGBM	0,9512	0,8892	0,0619
CatBoost	0,9527	0,8853	0,0673
Red neuronal multicapa	0,9449	0,8715	0,0734

Fuente: Elaboración propia.

En la tabla 7, analizando los valores promedio de exactitud en el entrenamiento y en la prueba, se elegirá el mejor modelo. Tomando en cuenta el mayor valor promedio de exactitud en el conjunto de prueba (0,8892), el mejor modelo sería LightGBM y, respecto de la diferencia entre los valores promedio de exactitud en los conjuntos de entrenamiento y prueba, el modelo de regresión logística presenta la menor diferencia (0,0084), seguido por el modelo LightGBM (0,0619). Considerando su mejor capacidad de generalización debido a su mayor promedio de exactitud en el conjunto de prueba [17] y que no presenta un sobreajuste significativo, dada la segunda menor diferencia promedio de exactitud, el modelo elegido sería LightGBM, el cual también ofrece una serie de ventajas como mayor velocidad de entrenamiento y eficiencia, menor uso de memoria, mayor exactitud y capacidad para gestionar datos a gran escala, entre otras [5].

CONCLUSIONES

En una primera etapa del preprocesamiento de datos, se eliminaron siete variables de un total de 26 que había originalmente, siendo la principal causa el hecho de que algunas tenían valores únicos, pues representaban códigos, y otras eran redundantes con respecto a variables que se encontraban mejor definidas. Además, se eliminó la variable CURSO_CURSADOS debido a que estaba altamente correlacionada con la variable CURSO_APROBADOS.

En una segunda etapa, se procedió a discretizar las variables CURSO_APROBADOS y CURSO_DESAPROBADOS debido a la marcada diferencia entre sus categorías, y también se discretizó la variable EDAD_ANIOS para captar mejor la deserción a través de rangos de edad. Asimismo, al analizar la variable PROMEDIO_GENERAL, se pudo observar que existía una cantidad importante de estudiantes cuyos promedios eran cero, posiblemente reprobados por inasistencia o bajo rendimiento. Por otro lado, después de la discretización, se tuvo una mayor predominancia de variables categóricas (16) respecto de las variables numéricas (1) y, al realizar la dummificación, se obtuvieron en total 50 variables numéricas.

Utilizando una técnica de selección de mejores variables basada en modelos y secuenciales, se eligieron 36 variables del total de 50. Como variables más influyentes para pronosticar la deserción se identificaron: promedio general, semestres (menos el 2022-2), sede (menos Lima), ciclo, sexo, curso reprobado por inasistencia, estado de aprobado, edad (menos los que tienen de 30 a 59 años), estado de pago de matrícula a tiempo, estado de pago de pensión a tiempo, cursos aprobados (menos quienes aprobaron de 7 a 9) y cursos desaprobados.

Existe una mayor cantidad de estudiantes que no desertaron respecto de quienes sí lo hicieron, razón por la cual se empleó la validación cruzada con cinco particiones para compensar esta diferencia. Así, se entrenaron ocho modelos de clasificación (regresión logística, k-NN, árbol de decisión, *random forest*, XGBoost, LightGBM, CatBoost y red neuronal multicapa), de los cuales se calcularon sus respectivas exactitudes y se realizó lo mismo para los conjuntos de prueba, siendo en un principio los modelos XGBoost, LightGBM y CatBoost seleccionados como los mejores debido a sus mínimas diferencias entre las exactitudes en entrenamiento y prueba.

Tomando en cuenta el mayor valor de exactitud en el conjunto de prueba, lo cual permite una mejor capacidad de generalización para el modelo, y la segunda menor diferencia entre las exactitudes en los conjuntos de entrenamiento y prueba, lo cual evidencia que no existe un sobreajuste, el mejor modelo para predecir la deserción estudiantil sería LightGBM.

Una de las limitaciones más importantes encontradas durante el desarrollo de este estudio fue no poder contar con datos de los años más recientes, específicamente 2023 y 2024, debido principalmente a que estos datos se encuentran dispersos en diferentes áreas de Tecsup, almacenados en diversos repositorios y requieren una gestión previa para acceder a ellos. Una posible solución para evitar esta situación y realizar un futuro estudio sobre deserción que sea más preciso sería implementar un almacén de datos (*data warehouse*) en el que se integren y preprocesen todos los datos de la institución, ya sea a través de un servidor local o mediante una implementación en la nube.

A nivel institucional, la deserción estudiantil siempre generará un impacto negativo económico y/o académico, razón por la cual se pueden adoptar medidas como realizar cambios o actualizaciones en las mallas curriculares acorde con el mercado laboral, llevar a cabo un seguimiento minucioso a los estudiantes de los primeros ciclos atendiendo a las dificultades que puedan presentar, indagar por qué existen estudiantes que faltan continuamente durante el semestre académico e implementar políticas que permitan cumplir con los pagos de matrículas o pensiones a tiempo, entre las más importantes.

REFERENCIAS

[1] Alania, P. (2018). Aplicación de técnicas de minería de datos para predecir la deserción estudiantil de la facultad de ingeniería de la Universidad Nacional Daniel Alcides Carrión [Tesis para obtener el grado de magíster]. Repositorio Institucional UNDAC.

[2] Aleans, K. (2012). *Determinantes de la deserción estudiantil universitaria por niveles de formación en instituciones de educación superior de la ciudad de Medellín.* Universidad EAFIT.

[3] Amaya, Y., Barrientos, E., & Heredia, D. (2014). *Modelo predictivo de deserción estudiantil utilizando técnicas de minería de datos.* RedCLARA. https://documentos.redclara.net/bitstream/10786/759/1/124-22-3-2014-Modelo%20predictivo%20de%20deserci%C3%B3n%20estudiantil%20utilizando%20t%C3%A9cnicas%20de%20miner%C3%ADa%20de%20datos.pdf

[4] Arias-Gómez, J., Villasís-Keever, M., & Miranda, M. (2016). El protocolo de investigación III: la población de estudio. *Alergia México*, 201-206.

[5] Banerjee, P. (2020). *LightGBM classifier in Python*. Kaggle. https://www.kaggle.com/code/prashant111/lightgbm-classifier-in-python

[6] Berens, J., Schneider, K., Görtz, S., Oster, S., & Burghoff, J. (2019). Early detection of students at risk–predicting student dropouts using administrative student data and machine learning methods. *Journal of Educational Data Mining*, 1-41.

[7] Camargo, A. (2020). *Modelo para la predicción de la deserción de estudiantes de pregrado, basado en técnicas de minería de datos* [Tesis para obtener el grado de magíster]. Repositorio Universidad de La Costa.

[8] Cuji, B., Gavilanes, W., & Sánchez, R. (2017). Modelo predictivo de deserción estudiantil basado en arboles de decisión. *Revista Espacios*, 17-25.

[9] Díaz, K., Chindoy, B., & Rosado, A. (2019). Review of techniques, tools, algorithms and attributes for data. En *Journal of Physics: Conference Series* (pp. 1-6). IOP Publishing.

[10] Escalante, J., Medina, C., & Vásquez, A. (2023). La deserción universitaria: un problema no resuelto en el Perú. *Revista Hacedor*, 60-72.

[11] Fernández, X., & Silva, E. (2014). Deserción estudiantil universitaria en el primer semestre. El caso de una institución de educación superior ecuatoriana. *Cuadernos del Contrato Social por la Educación*, 34-48.

[12] González, F., & Arismendi, K. (2018). Deserción estudiantil en la educación superior técnico-profesional: Explorando los factores que inciden en alumnos de primer año. *Revista de la Educación Superior*, 109-137.

[13] Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques*. Elsevier Inc.

[14] Hellas, A. *et al.* (2018). Predicting academic performance: A systematic literature review. En *Proceedings Companion of the 23rd Annual ACM Conference on Innovation and Technology in Computer Science Education (ITiCSE '18 Companion)* (pp. 175-199).

[15] Hernández, R., Fernández, C., & Baptista, M. (2014). *Metodología de la investigación.* McGraw-Hill Education.

[16] Iljin, V. (2023, 4 de mayo). *Comparing the Titans of Machine Learning: XGBoost, CatBoost and LightGBM*. LinkedIn. https://www.linkedin.com/pulse/comparing-titans-machine-learning-xgboost-catboost-lightgbm-iljin/

[17] James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *The elements of statistical learning with applications in R.* Springer.

[18] Ministerio de Educación (Minedu). (2024). *Resolución Viceministerial N° 095-2024-MINEDU.* https://cdn.www.gob.pe/uploads/document/file/6894408/5957002-rvm_n-_095-2024-minedu.pdf

[19] Mori, J. (2021). Factores asociados al riesgo en la deserción estudiantil en un instituto de educación superior tecnológico público. *Revista de Investigación de la Universidad Norbert Wiener*, 59-72.

[20] Rivera, K. (2021). Modelo predictivo para la detección temprana de estudiantes con alto riesgo de deserción académica. *Revista Innovación y Software*, 6-13.

[21] scikit-learn. (s. f.). *Feature selection.* scikit-learn. https://scikit-learn.org/stable/modules/feature_selection.html

[22] Sifuentes, O. (2018). Modelos predictivos de la deserción estudiantil en una universidad privada peruana. *Revista Industrial Data*, 47.52.

[23] Spositto, O., Etcheverry, M., Ryckeboer, H., & Bossero, J. (2010). *Aplicación de técnicas de minería de datos para la evaluación del rendimiento académico y la deserción estudiantil.* https://repositoriocyt.unlam.edu.ar/handle/123456789/1267

[24] Tam, J., Vega, G., & Oliveros, R. (2008). Tipos. métodos y estrategias de investigación científica. *Pensamiento y Acción*, 145-154.

[25] Vásquez, J. (2016). *Modelo predictivo para estimar la deserción de estudiantes en una institución de educación superior* [Tesis para obtener el grado de magíster]. Repositorio Académico Universidad de Chile.

[26] Viale, H. (2014). Una aproximación teórica a la deserción estudiantil universitaria. *Revista Digital de Investigación en Docencia Universitaria*, 59-75.

[27] Viera, D., Flores, M., & Pachari-Vera, E. (2020). Factores de deserción estudiantil: Un estudio exploratorio desde Perú. *Interciencia*, 586-591.

[28] Villegas, B., & Núñez, L. (2024). Factores asociados a la deserción estudiantil en el ámbito universitario. Una revisión sistemática 2018-2023. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo, 14*(28).

ACERCA DEL AUTOR

José Luis Espinoza Melgarejo

Magíster en Docencia Universitaria e Investigación Pedagógica de la Universidad San Pedro de Chimbote (USP), Perú; licenciado en Matemática de la Universidad Nacional de Ingeniería (UNI), Perú. Docente de educación superior con más de 11 años de experiencia laborando en instituciones como IDAT, Tecsup, Universidad Privada del Norte, Universidad Nacional Mayor de San Marcos, Universidad Tecnológica del Perú y Universidad de Ingeniería y Tecnología. Ha dictado diversos cursos en ciencia de datos, matemática, estadística, finanzas entre otros. Especialista en estadística y ciencia de datos que cuenta con estudios de posgrado en la Universidad Nacional Mayor de San Marcos (UNMSM), Perú y actualmente estudiante de doctorado en Estadística Matemática de la Universidad Nacional del Santa (UNS), Perú.

@ jespinozame@tecsup.edu.pe

@ josespijoin@yahoo.com

Recibido: 26-04-2025
Revisado: 06-10-2025
Aceptado: 17-10-2025

(cc) BY-NC

Esta obra está bajo una Licencia Creative Commons AtribuciónNoComercial 4.0 Internacional.