Education in the Knowledge Society

journal homepage http://revistas.usal.es/index.php/eks/

Ediciones Universidad Salamanca

Educational Data Science and Machine Learning: A Case Study on University Student Dropout in Mexico

Ciencia de Datos Educativos y aprendizaje automático: un caso de estudio sobre la deserción estudiantil universitaria en México

Antonieta Kuz[a]*, Rosa Morales[b]

[a] *Facultad de Informática, Universidad Metropolitana para la Educación y el Trabajo, Buenos Aires, Argentina.*
https://orcid.org/0000-0002-8696-0859 antokuz@esgcffaa.edu.ar
[b] *Departamento de Economía, Universidad de Monterrey, Monterrey, México.*
https://orcid.org/0000-0002-7044-2600 rosa.moralesv@udem.edu

ARTICLE INFO

Palabras clave
deserción estudiantil, educación superior, ciencia de datos educativos, análisis académico de datos, aprendizaje automático

RESUMEN

Hoy en día la deserción universitaria es un fenómeno inquietante que afecta a estudiantes, instituciones educativas y el estado. Una mirada a este fenómeno desde la Ciencia de Datos Educativos y la aplicación de técnicas de aprendizaje automático permite buscar las posibilidades de permanencia de los alumnos, es por ello que el objetivo de esta investigación es predecir la deserción escolar en el primer año de estudios a nivel universitario usando dichas técnicas. Se analiza un caso de estudio práctico en el ámbito educativo con información de estudiantes de una universidad privada en México. Se evidencia en el estudio que las métricas y la visualización de la estructuración para analizar patrones permiten mostrar que las características que predicen con mejor desempeño la deserción escolar institucional en el primer año de estudios a nivel universitario son el promedio estudiantil en el primer período y el porcentaje de la beca.

Keywords
students dropout, higher education, educational data science, data academics analytics, machine learning

ABSTRACT

Nowadays, university dropout is a disturbing phenomenon that affects students, educational institutions, and the state. A look at this phenomenon from Educational Data Science and the application of Machine Learning techniques allows us to search for the potential permanence of the students, which is why this research aims to predict school dropout in the first year of studies. university level using these techniques. A practical case study is analyzed in the educational field using a private university student database in Mexico. It is shown in the study that the metrics and the visualization of the structuring to analyze patterns allow to determine that the characteristics that best predict institutional dropout in the first year of studies at the university level are the average of the student in the first period and the percentage of the scholarship.

(*) Autor de correspondencia / Corresponding author

Recibido, 09/08/2022. Revisado, 22/01/2023. Aceptado, 07/02/2023. Publicado, 26/06/2024.
e-ISSN: 2444-8729

1. Introducción

Hoy en día el ritmo de producción de datos crece exponencialmente debido a la digitalización de la sociedad. La toma de decisiones basadas en datos a partir de la analítica tiene un rol central en las organizaciones. Los datos se recopilan y procesan en tiempo real a una escala sin precedentes haciendo uso de la aplicación de las Ciencias de Datos. Este campo de estudio se basa en el análisis y la comprensión adecuada de los datos para ayudar al entendimiento de distintas problemáticas que afectan a la sociedad actual, como el fenómeno de la deserción educativa en particular en el contexto de América Latina.

Si bien los indicadores educativos en América Latina mejoraron en términos de matriculación, logro educativo y deserción de estudiantes hasta 2019, una vez que comenzó la pandemia, la deserción, en particular, aumentó en la región latinoamericana, lo que provocó una gran crisis educativa en toda América Latina. La deserción estudiantil se ha mencionado como uno de los mayores problemas a abordar durante este período de tiempo. La deserción escolar en cualquier nivel representa una enorme pérdida de capital humano. En particular, la deserción estudiantil a nivel universitario puede ser muy costosa. Las personas que no completan estudios universitarios tienen alta probabilidad de tener menores ingresos (Luckman & Harvey, 2019), deteniendo así su movilidad social. Se ha documentado que, al menos para economías avanzadas, existe un mayor salario cuando se completa la educación universitaria (Fortin, 2006; Stoops, 2004). De igual forma, el abandono de la educación universitaria podría significar no conseguir las habilidades necesarias para conseguir empleos (Kirk, 2018), en especial los empleos del futuro, lo que redundaría en un mayor desempleo e inequidad social.

La deserción, retención y atracción de estudiantes son indicadores clave, formando parte de la sustentabilidad de las instituciones educativas mediante un enfoque basado en la toma de decisiones de análisis de datos. En cuanto al tema de la deserción y retención, este ha sido estudiado a nivel de educación superior bien sea para la deserción en el primer año de la carrera o para el abandono antes de completar la carrera universitaria (Cardona et al., 2023). González Fiegehen y Espinoza Díaz (2020) consideran que la deserción se puede definir como el proceso de abandono, voluntario o forzoso, de la carrera en la que se matricula un estudiante, por la influencia positiva o negativa de circunstancias internas o externas a él o ella. Se ha documentado que, en América Latina, el mayor porcentaje de la deserción a nivel universitario ocurre en el primer año de la carrera, alcanzando niveles de hasta 35% (Ferreyra et al., 2017). Adicionalmente, en los últimos años, la matricula del sistema educativo mexicano a nivel superior ha ido en retroceso más específicamente en el sector femenino que fue donde se registró la baja, actualmente el total se ubica en 4,030 millones (INEE, 2019).

Para hacer sentido de las predicciones de la deserción estudiantil en universidades privadas mexicanas hay que entender el contexto (Morales Salas & Rodríguez Pavón, 2022). A nivel de educación secundaria superior, el Instituto Nacional para la Evaluación de la Educación de México – INEE (2022) define que las causas de la deserción escolar están asociadas a una confluencia de factores, entre los que se pueden mencionar: prácticas pedagógicas inadecuadas, formación docente limitada y condiciones laborales precarias, infraestructura y equipamiento insuficiente, incompatibilidad entre la cultura juvenil y escolar, currículo poco pertinente, gestión escolar deficiente, y participación limitada de padres y estudiantes en la escuela. No todos estos factores están relacionados con la deserción estudiantil universitaria privada a nivel institucional, en donde elementos como infraestructura y equipamiento insuficiente, formación docente limitada, entre otras características mencionadas por el INEE ya estarían solventadas.

En el caso de las instituciones universitarias privadas, los costos de la matrícula en conjunto con otros factores relacionados con el mismo proceso educativo, así como los gastos en vivienda, transporte y alimentación podrían ser relevantes en la decisión de iniciar los estudios de educación superior, al igual que la continuación de los estudios a través del tiempo. La matrícula en una institución privada en México a nivel superior, por ejemplo, podría estar entre 4000-6000 dólares americanos el semestre. Este costo podría ser relativamente bajo en comparación con los costos de educación superior en economías de ingresos altos; sin embargo, en un país cuyo PIB per cápita para el 2020 estaba alrededor de los 8909 dólares americanos (*World Bank Database*) y el ingreso promedio trimestral de la población era aproximadamente 2500 US para ese mismo año (Instituto Nacional de Estadística y Geografía de México), realizar estudios universitarios en una institución privada podría ser muy costoso para el promedio de la población, de allí la importancia de estudiar aspectos económicos y financieros, entre otros factores.

Dada la problemática planteada, el propósito de esta investigación es determinar las variables que predicen con mayor exactitud la deserción escolar universitaria en el primer año de estudios en una institución universitaria privada en México. Para América Latina, la literatura se ha enfocado en analítica académica, minería y analítica de datos, o modelos *logit/probit*, entre otros (Bonaldo & Pereira, 2016; Melguizo et al., 2011;

Santos et al., 2020; Von Hippel & Hoflinger, 2021). En México, Márquez-Vera et al. (2016) indagan la deserción estudiantil con aprendizaje automático. Los estudios antes mencionados incorporan variables relacionadas a: (i) información sociodemográfica, (ii) información académica y (iii) información económica y financiera. En específico, promedios estudiantiles, becas, apoyos financieros, género, edad, contexto familiar y económico, entre otras.

Este estudio contribuye a la literatura creciente en deserción estudiantil universitaria al usar técnicas de Aprendizaje Automático (AA) que permiten clasificar y jerarquizar las categorías que predicen el abandono estudiantil institucional a nivel de educación superior. Las causas que originan una deserción escolar obedecen a razones multifactoriales que van desde aspectos personales, familiares, académicos, económicos, hasta políticos, culturales e institucionales (Donoso & Schiefelbein, 2007; Tinto, 1982). Es por ello que nos preguntamos, de los factores antes citados, cuáles podrían ser los más importantes en la predicción de la deserción escolar institucional en el primer año de estudios a nivel universitario usando técnicas de AA. Dentro del sistema educativo, interpretar los datos correctamente permite conocer el comportamiento de los distintos colectivos de alumnos, cada uno con diferentes realidades, gracias a las predicciones arrojadas por los modelos predictivos de Inteligencia Artificial (IA), se pueden diseñar estrategias académicas y de servicio a esos alumnos. Al usar técnicas sencillas de implementar, la predicción realizada en esta investigación asegura un adecuado nivel de exactitud comparable con técnicas más complicadas.

A nivel institucional, el uso del análisis basado en datos permite crear estrategias de retención e integrarlas a la gerencia educativa. Nieuwoudt y Pedler (2021) señalan que los incentivos de las instituciones universitarias para estudiar y predecir la deserción escolar así como crear estrategias de retención son claros. Los aspectos económicos y de reputación están relacionados con ellos. En particular en las instituciones privadas, un beneficio es el ingreso proveniente de las matrículas estudiantiles (Burke, 2019; Simpson, 2005), así como la disminución de los costos de reclutamiento (Simpson, 2005). En lo que respecta a la reputación, Aljohani (2016) señala que instituciones universitarias con menor deserción estudiantil gozan de mayor reputación. La lógica detrás de esto es que instituciones educativas con bajas tasas de deserción son más responsables en su estabilidad financiera y muestran mayor efectividad institucional (Fike & Fike, 2008).

El registro global de datos que surja del contexto educativo y su posterior procesamiento son factores que pueden fortalecer las prácticas educativas y potenciar a cada institución. Así mismo es posible gestionar los datos de manera estratégica, eficiente, no aleatoria e interpretativa, y consecuentemente estructurar, optimizar, y organizar los procesos de la institución educativa. Es por lo anterior que en esta investigación se usan técnicas de AA, con énfasis en el XGBoost, así como el análisis de datos académicos sobre un conjunto de datos proporcionados por una universidad privada de México buscando identificar a los alumnos que estén en riesgo de desertar.

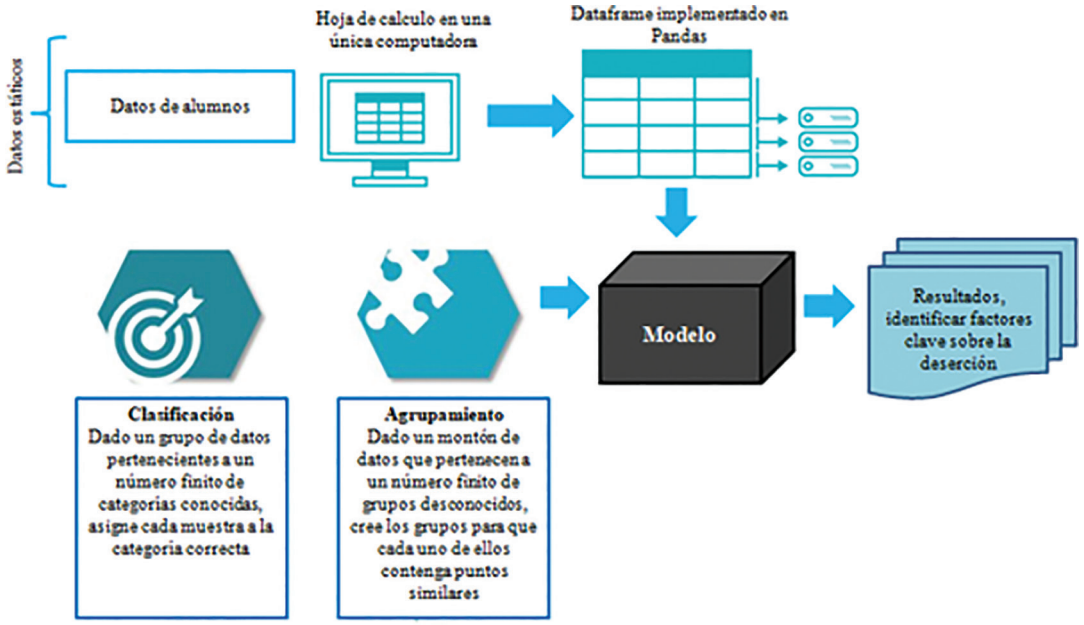
2. Metodología

2.1. Caso de Estudio

La metodología seleccionada es de un caso de estudio, la cual busca tomar un problema inmediato a resolver y su propósito fundamental se centra en aportar información respecto al contexto particular (Álvarez Álvarez & San Fabián Maroto, 2015). El objetivo principal del trabajo es detectar los factores que permitan predecir si un alumno de universidad privada en México sea retenido o abandone. Como se observa en la Figura 1 planteamos un esquema de arquitectura de solución del caso. El esquema propuesto busca curar, visualizar y extraer información del set datos, además de entrenar tres modelos predictivos de AA para luego abordar la toma de decisión basada en datos.

Para llevar adelante este caso de estudio se han tenido en cuenta los algoritmos de AA aplicando una clasificación. Con el conjunto seleccionado se busca poder generar clasificaciones en base a un entrenamiento sobre información pasada, seguida de una validación de las predicciones generadas, haciendo uso y probando de distintos algoritmos de redes neuronales, árbol de decisión y regresión logística, los cuales hacen uso en distinta manera los datos (en particular, los atributos elegidos). Por este motivo, para el desarrollo de este estudio se llevaron a cabo las siguientes fases: caracterización de los datos, selección de la tecnología para implementar el modelo, selección de las técnicas a utilizar en el desarrollo de la implementación, construcción y prueba del modelo de deserción en la educación superior y validación del modelo de deserción y análisis de los resultados arrojados en su despliegue.

Figura 1. Arquitectura de un modelo de solución mediante AA.



2.2. Caracterización de los datos

En este trabajo se consideró una colección de datos educativos denominada "*Student dropout dataset*" proporcionada por un instituto de investigación de una universidad privada en México. Esta colección de datos está conformada por características personales, socioeconómicas y académicas de los y las estudiantes. El conjunto de datos incluye información anonimizada relacionada con estudiantes de pregrado y de preparatoria que se hayan matriculado y asistido al menos un semestre de 2014 a 2020. Este conjunto de datos tiene 143326 registros con 50 características. Cada registro se corresponde con un o una estudiante. Se seleccionaron solo los registros clasificados como estudiantes de pregrado (77517 observaciones). Dado que los datos son esenciales para el AA, es relevante que esté garantizada la calidad de los datos para evitar que los resultados de los análisis estén distorsionados o sesgados. La premisa fue obtener datos de calidad con lo cual lo que se busco fue que sean precisos, completos, coherentes, uniformes y trazables, a través de los siguientes los procesos recomendados de limpieza, transformación, selección y filtrado de los datos.

Al realizar la limpieza de los datos, se detectaron y eliminaron duplicados, se identificaron 5,1% de datos faltantes sin ninguna leyenda, registros incompletos, es decir, en los que había inconsistencias y con poca información de características, en donde se leía "sin información" o "no aplica" (más del 50% de los registros contenían esa leyenda), además de los valores atípicos. Luego de la limpieza, la cantidad de registros se redujo a 1847. Para poder manejar los datos faltantes, dentro de AA existen técnicas que se dividen en dos grandes grupos que son el descarte de datos y la imputación (Batista & Monard, 2010). El descarte consiste simplemente en eliminar los registros que contengan datos faltantes, comúnmente si estos contienen más del 20% de los datos faltantes (Hernández & Rodríguez, 2008), mientras que en la imputación lo que se busca es estimar el valor del dato faltante usando la información de los registros vecinos o la información presente en otras variables (o columnas) del conjunto de datos. En nuestra investigación utilizamos el descarte para eso trabajamos con *pairwise deletion* (eliminación por pares), ya que es un método menos agresivo y que permite preservar los datos conocidos. Dado que con la imputación se busca mirar el comportamiento de los datos vecinos para poder estimar el valor del dato faltante se podía correr el riesgo de cambiar la distribución de los datos o distribuciones marginales de las variables que en su mayoría son categóricas conllevando a un aumento del sesgo. Es por esto por lo que, en el caso de nuestra investigación, se optó por la eliminación de registros y no por la imputación.

Con respecto a las variables seleccionadas, estas tienen una correspondencia con la revisión de la literatura mencionada en la introducción, así como con el contexto de universidad privada multi-campus en México, en

donde aspectos geográficos y de compromiso estudiantil y familiar podrían ser relevantes. De igual forma, se tomaron en cuenta solo aquellas categorías que aplicasen a estudios de pregrado. La Tabla 1 muestra las variables a considerar, la definición de las variables está basada en la transformación que se hizo de las mismas y en la información proporcionada por el diccionario original de la base de datos "*Student dropout dataset*", el mismo se encuentra en el siguiente link: https://doi.org/10.57687/FK2/PWJRSJ

Tabla 1. Lista de variables.

Aspectos	**Variables**	**Descripción**
	Retención	Es una variable dicotómica. Toma el valor de 1 cuando estudiante es retenido y 0 cuando no lo es.
Sociodemográficos	Género	Es una variable dicotómica. Toma el valor de 1 cuando estudiante es mujer, 0 si es hombre.
	Edad	Corresponde a la edad de una persona inscrita en la universidad en estudio.
	Estudiante Foráneo	Es una variable categórica. 1=local; 2= foráneo, 3= estudiante internacional.
	Máximo grado de estudios alcanzado por los padres	Es una variable categórica. La variable toma los siguientes valores: 1=sin grado alcanzado, 2=licenciatura, 3=maestría, 4=doctorado
	Descripción del nivel máximo de estudios alcanzado por el padre	Es una variable categórica. La variable toma los siguientes valores: 1=fue a la universidad, pero no se graduó; 2=graduado de primaria; 3=graduado de secundaria; 4= sin grado alcanzado;5= master; 6=doctorado; 7= grado técnico o comercial=7; 8=licenciatura
	Máximo grado alcanzado por el padre	Es una variable categórica. La variable toma los siguientes valores: 1=sin grado alcanzado, 2=licenciatura, 3=maestría, 4=doctorado
	Descripción del nivel máximo de estudios alcanzado por la madre	Es una variable categórica. La variable toma los siguientes valores: 1=fue a la universidad pero no se graduó; 2=graduado de primaria; 3=graduado de secundaria; 4= sin grado alcanzado;5= master; 6=doctorado; 7= grado técnico o comercial=7; 8=licenciatura
	Máximo grado alcanzado por la madre	Es una variable categórica. La variable toma los siguientes valores: 1=sin grado alcanzado, 2=licenciatura, 3=maestría, 4=doctorado
	Indice de Brecha Social	Es una variable categórica. 1= Brecha baja, 2= Brecha media, 3=Brecha alta
Académicos	Promedio semestre anterior	Es una variable continua que corresponde al promedio del semestre anterior.
	Promedio del primer período	Es una variable continua que corresponde al promedio del primer semestre de estudios.
Financieros-económicos	Porcentaje de Beca	Es una variable continua expresada en tanto por uno. 0 denota 0% de beca, 1 denota 100% de beca.
	Tipo de beca	Es una variable categórica que toma distintos valores de acuerdo al tipo de beca. 1=talento académico, 2=Beca de la armada/naval, 3=Beca hijo de profesor/ empleado/director, 4=beca de contingencia, 5= talento cultural, 6= talento empresarial, 7=beca lideres del mañana,8=talento de liderazgo, 9=No tiene beca,10= talento deportivo, 11=beca tradicional.
	Porcentaje del préstamo	Es una variable continua expresada en tanto por uno.
	Porcentaje total de ayudas financieras (beca + crédito)	Es una variable continua expresada en tanto por uno.
	Costo de la matrícula de la escuela de origen	Es una variable categórica que toma distintos valores de acuerdo a los niveles de costos de la escuela de origen. 1= Escuela pública=1, 2= Escuela de bajo costo, 3= Escuela de costo medio, 4=Escuela de costo promedio, 5= escuela de alto costo.
	Nivel socioeconómico	Es una variable categórica que toma distintos valores de acuerdo al nivel socioeconómico.1=nivel 1, 2= nivel 2, 3=nivel 3, 4=nivel 4,5= Nivel 5, 6= Nivel 6, L 7= Nivel 7

(continúo)

Tabla 1. Lista de variables. *(continuación)*

Aspectos	Variables	Descripción
Compromiso estudiantil/familiar	Padres son exalumnos	Es una variable dicotómica. Toma el valor de 1 cuando padres de el/la estudiante fueron a la universidad en estudio, 0 si no fueron.
	Padre ex-alumno	Es una variable dicotómica. Toma el valor de 1 cuando padre del estudiante fue a la universidad en estudio, 0 si no fue.
	Madre ex-alumna	Es una variable dicotómica. Toma el valor de 1 cuando madre del estudiante fue a la universidad en estudio, 0 si no fue.
	Estudiante de Prepa Tec	Es una variable dicotómica. Toma el valor de 1 cuando estudiante fue a la preparatoria de la universidad en estudio, 0 si no fue
	Total de actividades que el estudiante realiza en campus	Es una variable representada en números enteros y corresponde al total de actividades en las que un o una estudiante se inscribió en el semestre.
Geográficos	Tipo de Zona	Es una variable categórica que corresponde al tipo de zona donde se localiza la dirección con la que un o una estudiante se inscribió. 1=zona Rural, 2=zona Semiurbana, 3=zona urbana
	Región del Campus	Es una variable categórica que corresponde a la región del campus donde el o la estudiante se inscribió. 1 = Región Monterrey, 2 = Región Oeste=2, 3= Región Ciudad de Mexico, 4= Región Sur/Centro, 5= Desarrollo Regional

Para comprender la distribución de los datos en la Tabla 2 se muestran los estadísticos descriptivos.

En un análisis exploratorio del set, para XGBoost se definió un layout que tiene 10 columnas (9 predictores, uno de cada categoría, y una variable objetivo que indica si se retiene o continua el alumno o abandona o no) (Ver Tabla 3). La selección de características en un modelo es de suma relevancia por eso, siguiendo a Nair y Baghat (2019), utilizamos aquellas características en nuestros datos que más contribuyen a la variable objetivo con base al uso de las bibliotecas Scikit-learn y en particular selectKBest. Esta biblioteca se utilizó para extraer las mejores características, seleccionándolas de acuerdo con la puntuación más alta, siendo de esta manera los mejores predictores para la variable objetivo. Además, esta biblioteca al ser un desarrollo estable reduce el sobreajuste, mejora la precisión y reduce el tiempo de entrenamiento.

2.3. Selección de la tecnología para implementar el modelo

Se usa Jupyter Notebook de Anaconda ya que contribuye a modularizar e incorporar librerías como Sklearn en distintos notebooks de Jupyter que luego se pueden importar entre sí (Bobadilla, 2021).

2.4. Selección de las técnicas a utilizar en el desarrollo de la implementación

Las técnicas seleccionadas son: XGBoost, Regresión Logística, Red Neuronal y Árboles de decisión. A continuación, detallamos cada uno de estos:

a. XGBoost: es un algoritmo de clasificación de AA (Espinosa-Zúñiga, 2020) predictivo supervisado que utiliza el principio de boosting el cual genera un modelo de predicción, a partir de modelos secuencialmente débiles, empleando un algoritmo de optimización, de descenso de gradiente. Ha sido usado por Huo et al. (2023) para predecir la deserción de estudiantes no tradicionales a nivel universitario, pero de acuerdo a nuestro conocimiento es de relativo poco uso para muestras de estudiantes de distintos backgrounds ni para el caso Latinoamericano. Según Chen y Guestrin (2016), este algoritmo presenta mejores resultados frente a los devueltos por modelos más complejos computacionalmente, en particular para problemas con datos heterogéneos, debido a que utiliza un mínimo de recursos de computación en períodos cortos. El algoritmo XGBoost trabaja de la siguiente forma, parte del árbol de decisiones para clasificar o pronosticar sobre una variable objetivo (y) (que en el caso de esta investigación sería la variable retención) potenciando

Tabla 2. Resumen de los estadísticos descriptivos.

Variable	Observaciones	Promedio	Desviación Estándar	Mínimo	Máximo
Retención	1,847	0,94261	0,23265	0	1
Género	1,847	0,491067	0,5000556	0	1
Edad	1,847	17,93611	0,845119	16	26
Estudiante Foráneo(a)	1,847	1,231727	0,4220499	1	2
Máximo grado de estudios alcanzado por los padres	1,847	2,214402	0,6758073	1	4
Descripción del nivel máximo de estudios alcanzado por el padre	1,847	6,297239	2,333725	1	8
Máximo grado alcanzado por el padre	1,847	2,032485	0,6959295	1	4
Descripción del nivel máximo de estudios alcanzado por la madre	1,847	6,454792	2,282342	1	8
Máximo grado alcanzado por la madre	1,847	1,873308	0,6698725	1	4
Índice de Brecha Social	1,847	1,064429	0,254252	1	3
Promedio semestre anterior	1,847	91,22637	4,928521	0	100
Promedio del primer período	1,847	90,68799	10,51573	0	100
Porcentaje de Beca	1,847	0,307066	0,1708358	0	0.8
Tipo de beca	1,847	9,401191	3,576567	1	11
Porcentaje del préstamo	1,847	0,222935	0,0504428	0.05	0.3
Porcentaje total de ayudas financieras (beca + crédito)	1,847	0,53	0,1868044	0.1	0.9
Costo de la matrícula	1,847	3,374662	1,498778	1	5
Nivel socioeconómico	1,847	6,07634	1,1923	1	7
Padres son exalumnos	1,847	0,145642	0,3528421	0	1
Padre ex-alumno	1,847	0,108825	0,311504	0	1
Madre ex-alumna	1,847	0,07255	0,2594668	0	1
Estudiante de Prepa Tec	1,847	0,70601	0,4557109	0	1
Total de actividades que el estudiante realiza en campus	1,847	1,594478	1,112634	0	5
Tipo de Zona	1,847	2,849486	0,4938171	1	3
Region del Campus	1,847	2,653492	1,315841	1	5

Tabla 3. Selección parcial de datos y atributos.

Género	Edad	Máximo Grado Padres	Foráneo	Tipo de zona	Región	Promedio primer período	Porcentaje de beca	Total de actividades en campus	Retención
0	18	2	1	3	5	81	0,15	1	1
0	18	3	2	3	5	92	0,25	1	1
1	17	3	1	3	5	93	0,45	2	1
0	18	2	1	3	5	94	0,25	2	1
0	18	3	1	3	5	92,66	0,4	3	1

los resultados, debido al procesamiento secuencial de la data con una función de pérdida o coste, la cual, minimiza el error iteración tras iteración, haciéndolo de esta manera, un pronosticador fuerte.

b. Regresión Logística*:* es un algoritmo de aprendizaje supervisado que se utiliza cuando la variable objetivo es categórica (Cabero-Almenara et al., 2022; Peláez, 2016). La función hipotética h(x) de regresión lineal predice valores ilimitados. Pero en el caso de la regresión logística, donde la variable objetivo es categórica, y en esta investigación se representa por la variable retención, se requiere de restringir el rango de valores predichos. Siguiendo a Peláez (2016), matemáticamente, se puede formular de esta forma: $y = \sigma(z) = \sigma(WX) = \sigma(\sum(w_i\, x_i)) = \sigma$, donde los valores de x se corresponden a los distintos atributos de nuestro problema representados previamente en la Sección 2.2. Todas las entradas se combinan con una línea con los coeficientes *w*. Y luego se aplica la función logística (también llamada sigmoidea) al resultado. Dicha función es el núcleo del método se utiliza la función logística que puede tomar cualquier número de valor real y asignar a un valor entre 0 y 1, y cuya fórmula es: $\sigma(z) = \frac{1}{1+e^{-z}}$. El aprendizaje se realiza con optimización numérica, mediante el gradiente descendiente. La función de coste para optimizar los coeficientes es la siguiente: $J = -\frac{1}{m}\sum_i^m y_i \log(p_i) + (1 - y_i)\, \log(1 - p_i)$

c. Red Neuronal: la red neuronal más simple es aquella que consta de una sola neurona y se llama perceptrón la cual tiene una capa de entrada y una neurona (Balanta Viera & Palacios Medina, 2018). En esta investigación se usa una red neuronal simple con una capa utilizando Sklearn neural_network. Dado que es una red neuronal simple, no cuentan otras capas. Siguiendo a Caicedo Bravo y López Sotelo (2009), la capa de entrada actúa como dendritas y es responsable de recibir las entradas El número de nodos en la capa de entrada es igual al número de entidades en el dataset de entrada. Cada entrada se multiplica por un peso (que normalmente se inicializa con algún valor aleatorio) y los resultados se suman. Luego, la suma pasa a través de una función de activación. La función de activación de un perceptrón se asemeja al núcleo de la neurona del sistema nervioso humano. Procesa la información y produce una salida. En el caso de un perceptrón, esta salida es el resultado final.

d. Árboles de Decisión: es una técnica de AA que permite la construcción de modelos predictivos de analítica de datos basados en su clasificación según ciertas características o propiedades, o en la regresión mediante la relación entre distintas variables para predecir el valor de otra (Barros et al., 2012). En el caso de esta investigación se busca predecir el valor de la retención a través de la clasificación de las variables ya especificadas previamente. A través del árbol se analizan algunas reglas para comprobar las predicciones. El árbol de decisión es una estructura que está formada por ramas y nodos de distintos tipos. Los nodos internos representan cada una de las características o propiedades a considerar para tomar una decisión. En este caso las características se corresponden a las variables listadas en la Tabla 3. Los nodos intermedios (las ramas) representan soluciones. Los nodos finales (las hojas) nos dan la predicción y el resultado que vamos buscando.

3. Resultados

3.1. Aplicación de XGBoost

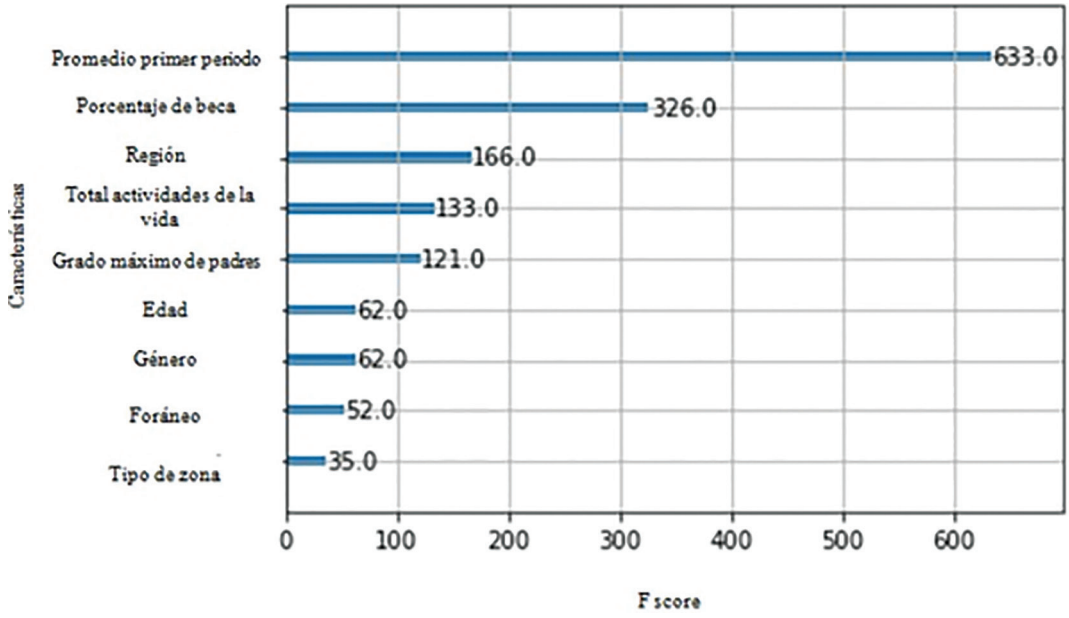
División de la base de datos: Una vez realizadas las tareas de depuración y filtrado del set previamente realizadas, se divide la base limpia en dos partes para la etapa de modelado:

- Base de entrenamiento: Con esta base se entrenara el modelo XGBoost y tendrá 70 % del total de registros elegidos aleatoriamente.
- Base de validación: en función del punto 1 se validarán los resultados de ambos modelos y contendrá el 30 % restante del total de registros.
- Para testear y entrenar el XGBoost se consideraron muestras mezcladas y se mantuvo una semilla utilizando un parámetro llamado *random_state*, para poder reproducir el mismo experimento.

Modelado

Se aplicó el algoritmo XGBoost sobre la base de entrenamiento a fin de entrenar a los correspondientes modelos, con los parámetros default de cada algoritmo en Jupyter. En la Figura 2 vemos el resultado del

Figura 2. Importancia de las características.



modelo entrenado, muestra la importancia de predictores de acuerdo con el modelo XGBoost, de las variables promedio en el primer período, porcentaje de la beca y la región, pero en particular donde se observa que promedio en el primer período es también el predictor que más pesa para determinar si el o la estudiante continúa en la universidad.

Analizamos la predicción para los valores de prueba. Para este ejemplo vemos un hombre de 18 años cuyas características son queda retenido en la universidad (ver Tabla 4).

La matriz de confusión se presenta en la Figura 3 y la exactitud del modelo XGBoost es de 0,9928. La matriz de confusión es una tabla que describe el desempeño de un modelo de predicción. Una matriz de confusión contiene los valores reales y los valores predichos y de esta manera podemos utilizar estos valores para calcular la puntuación de precisión del modelo.

3.2. Aplicación de otras técnicas de AA: Regresión Logística, Red Neuronal y Árbol de Decisión

Una vez realizadas las pruebas con XGboost se seleccionaron 2 atributos como predictores: región y promedio del primer período, ya que son los mejores atributos para evaluar la regresión logística, árbol de decisión y red neuronal. Así mismo, se escogió una variable objetivo especificada en el atributo retención, que indica si el alumno se retiene o abandona.

La aplicación de las técnicas de regresión logística, red neuronal y árbol de decisión siguen una serie de pasos como la prueba y entrenamiento de los datos, la creación del algoritmo, la predicción y evaluación de este. Una vez creado el algoritmo y las pruebas junto con algunas predicciones sobre el conjunto de datos de prueba, es necesario evaluar qué tan bien funciona el algoritmo. Para evaluar un algoritmo, las métricas más comúnmente utilizadas son una matriz de confusión, precisión, recuperación y puntuación f1. La Tabla 5 resume en detalle los pasos de la aplicación de cada una de las pruebas.

Para todas las técnicas mencionadas analizamos los resultados en función del informe de clasificación: el informe de clasificación muestra la exactitud, memoria, precisión y puntuación F1. Ver Tabla 6 para el soporte de puntuaciones para el modelo.

Para mostrar la sensibilidad en función de los falsos positivos (complementario de la especificidad) para distintos puntos de corte se muestran las curvas ROC. Estas son una representación gráfica que ilustra la relación entre la sensibilidad y la especificidad de un sistema clasificador para diferentes puntos de corte. Teniendo en cuenta la especificidad (Ratio de verdaderos negativos entre todos los negativos). Los valores calculados para los tres modelos que están detallados en la Tabla 7 y pueden verse en la Figura 4. El área bajo la curva (AUC),

Tabla 4. Ejemplo usado en la predicción con valores de prueba.

Categoría	**Valor**
Máximo grado de estudios alcanzado por los padres (MDP)	2
Estudiante Foráneo (F)	1
Tipo de Zona (Z)	3
Región (R)	5
Promedio primer período (GPA1)	81
Porcentaje de Beca (SPer.)	0,15
Número de actividades que el estudiante realiza en campus (TCLA)	1
Nota: MDP=2: ambos padres tienen licenciatura; F=1: el estudiante es nacional, Z=3: la dirección del estudiante corresponde a un área urbana, R=5: el estudiante se inscribió en una zona en desarrollo; GPA1=81: el promedio del primer período es 81; SPer=0,15: el porcentaje de la beca es de 15; TCLA=1: el número de actividades que el estudiante realiza en campus es uno.	

Figura 3. Matriz de confusión

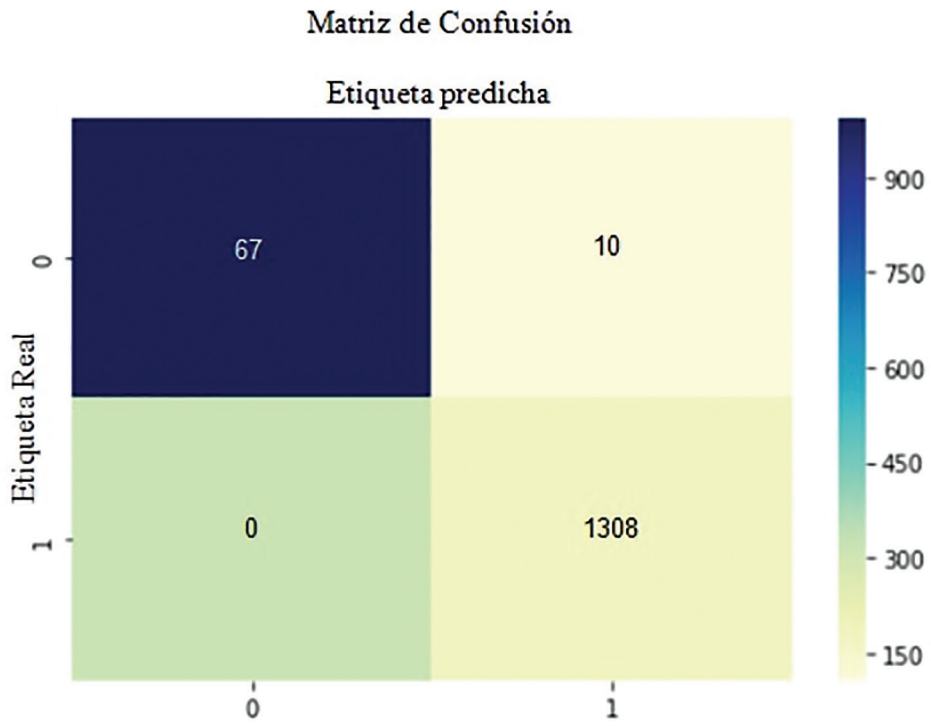


Tabla 5. Pasos de aplicación de tres técnicas de AA al set de datos.

Pasos de aplicación de técnicas	**Regresión Logística**	**Red Neuronal**	**Árbol de Decisión**
Prueba y Entrenamiento de los Datos	Datos de entrenamiento: 70%, Datos de prueba: 30%		
Algoritmo	Algoritmo de Regresión	Algoritmo voraz. El algoritmo voraz elige qué atributos y qué límites son los mejores para tomar las decisiones.	Algoritmos de Clasificación. El algoritmo hace predicciones basado en las relaciones entre las columnas de entrada del conjunto de datos.
Predicción	La función de regresión logística se implementa importando el modelo de regresión logística en el módulo sklearn. El modelo se ajusta al tren con la función de ajuste.	Se utiliza una red neuronal simple dada la cantidad de elementos de la muestra.	Se usa un árbol simple para hacer predicciones, se utiliza el método predict de la DecisionTreeClassifier clase.

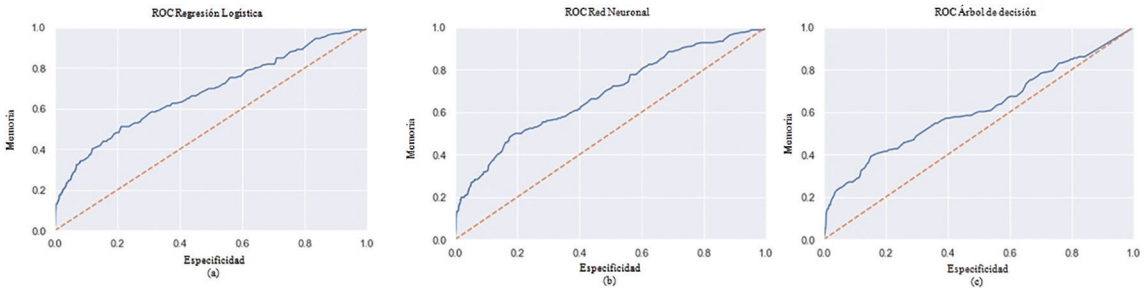
Tabla 6. Resultados obtenidos de tres técnicas de AA al set de dados en tanto por uno.

Medidas	**Regresión Logística**	**Red Neuronal**	**Árbol de Decisión**
Exactitud: mide el porcentaje de casos que el modelo ha acertado.	0,9328	0,9333	0,9336
Memoria: El resultado de predicción del modelo de clasificación se predice correctamente como la proporción de ejemplos positivos al total de ejemplos positivos.	0,9359	0,9352	0,9371
Precisión: es el porcentaje de precisión de las predicciones realizadas por el modelo. Esta puntuación significa el nivel hasta el cual la predicción hecha por el modelo es precisa.	0,9958	0,9772	0,9953
Puntuación F1: es la media armónica de precisión y recuperación.	0,9649	0,9652	0,9649

Tabla 7. AUC para los tres modelos del caso.

	Área por debajo de la curva
Regresión Logística	0,6803
Red Neuronal	0,6861
Árbol de decisión	0,6139

Figura 4. Curva ROC de las tres técnicas de AA.



ROC este puntaje nos da una buena idea de qué tan bien funciona el modelo. La curva ROC generalmente se encuentra por encima de y = x, por lo que el valor de AUC generalmente se encuentra entre 0,5 y 1. Cuanto mayor sea el AUC, mejor será el efecto de clasificación.

La Tabla 7 muestra que todos los modelos tienen la capacidad de predicción ya que todas las áreas están por encima del 0,5, lo que indica que los modelos funcionan bien en. En la figura 4 se puede observar gráficamente los AUC a través de las curvas ROC.

Al comparar las figuras 4(a), 4(b) y 4(c) se observa que, dado que las curvas azules están por encima de las líneas de 45 grados punteadas en rojo, todos los modelos tienen capacidad de discriminar entre estudiantes retenidos y no. Sin embargo, la curva ROC de la figura 4(b), correspondiente a la red neuronal, muestra un área bajo la curva con mayor distancia de la línea roja, indicando un mayor rendimiento de ese modelo sobre los representados en la figura 4(a) y 4(c).

4. Discusión y conclusiones

Dentro del dominio de la IA, las técnicas utilizadas son en realidad formas avanzadas de modelos que integran la estadística, la matemática y las ciencias de la computación para integrarlas en modelos a través del análisis inteligente de conjuntos de datos que nos proporcionan herramientas para calcular tareas que antes se pensaba que estaban reservadas para los seres humanos. A través del AA se busca que los algoritmos sean capaces de aprender por sí mismos, exponiéndolos a situaciones distintas que producen resultados diferentes. Se usó la técnica XGBoost para responder la pregunta de investigación de cuáles son los elementos que predicen la deserción

escolar en el primer año de estudios a nivel universitario. Esta técnica permitió encontrar la importancia de las características estudiadas para predecir la deserción escolar de un estudiante a nivel de educación universitaria en su primer año en México. Los resultados muestran nueve categorías más relevantes, de las cuales tres tienen una mayor ponderación: promedio estudiantil en el primer período de estudio, porcentaje de la beca otorgada y región del campus donde la o el estudiante se inscribió para estudiar respectivamente. La exactitud del XGBoost está por encima del 99%. Con el fin de validar los resultados se aplicaron la regresión logística, la red neuronal y el árbol de decisión. Para estas técnicas, la exactitud no superó el noventa y cuatro por ciento, a pesar de que se realizó un balanceo en los datos y de depurar la información para obtener una base de datos lo más limpia de ruido posible; es posible que la decantación progresiva usada tanto para el entrenamiento como para la evaluación haya incidido en esos resultados.

Ahora bien, con respecto a las implicaciones de los resultados, en cuanto al desempeño académico, estudios previos (Aulck et al., 2017; Ferreira & Andrade, 2016) han encontrado, al igual que este estudio, que el promedio de estudios del primer período predice la deserción escolar a nivel universitario, es solo que las técnicas implementadas en esta investigación son distintas y predicen con mayor exactitud y precisión. Adicionalmente, en el caso mexicano el contexto es distinto en cuanto al tipo de economía y las características del sistema educativo. En esa dirección y en lo que respecta a América Latina, estos resultados son consistentes con Von Hippel y Hofflinger (2021) quienes predicen que, en el caso de Chile, a nivel universitario el promedio en el primer período de estudios tiene alto poder predictivo. Chile posee una característica en común con México, tiene altos niveles de desigualdad económica; por tanto, el promedio del primer período estudio sirve como señal para saber si el esfuerzo económico que constituye la inversión en educación superior va a tener algún retorno. Este resultado puede orientar las políticas de gestión educativa de tal forma de extender estrategias como las planteadas por Rojas-López (2017) a otros contextos para mejorar el desempeño estudiantil en ese primer período, así como la creación apoyos al estudiantado a nivel académico que pueden ir desde los cursos remediales entre los que se puede incluir la selección de cursos de acuerdo con las habilidades iniciales.

En lo que se refiere a aspectos económicos, nuestros resultados son consistentes con la literatura previa. Herzog (2005) ha demostrado que los montos de las becas son importantes determinantes para explicar el abandono escolar. La contribución de este estudio va en esa dirección y avanza en el sentido que no solo permite explicar sino pronosticar la deserción escolar universitaria en el primer año de estudios usando técnicas como el XGBoost con la precisión antes mencionada. Si bien el estudio de Bonaldo y Pereira (2016) había encontrado las becas como determinante de deserción para Brasil, nuestro estudio contribuye en el tema de la predicción del abandono escolar universitario al predecir el abandono usando la categoría beca medida en porcentaje de la matrícula. El resultado parece solo un detalle en la forma de medición; sin embargo, en la discusión sobre el abandono escolar universitario la forma en cómo se mide y modela la ayuda económica es importante en la precisión de la predicción y explicación del abandono escolar universitario (Herzog, 2005). Los resultados vinculados con el porcentaje de la beca cobran importancia en la gestión educativa de países de ingresos medios o bajos como México en donde resulta muy costoso para la población acceder a educación universitaria privada. De tal forma que la gestión universitaria podría orientarse a intentar buscar soluciones relacionadas con las ayudas financieras.

En lo que concierne a la región del campus donde el o la estudiante se inscribe, Gitto et al. (2016) han encontrado que la localización del campus está correlacionada con la deserción estudiantil universitaria. Nuestros resultados avanzan algo más a los estudios previos, al no solamente encontrar una asociación entre aspectos geográficos y abandono escolar institucional, ni suscribirse solo al campus principal, sino también en identificar las distintas localizaciones de los campus como un predictor de la deserción estudiantil. Este resultado tiene implicaciones no solamente para la gerencia universitaria de la institución aquí estudiada, sino que podría usarse en el contexto universitario de América Latina para aquellas universidades privadas con múltiples campus. Estudiar las características particulares de los campus podría orientar las políticas en cuanto a deserción estudiantil de forma más enfocada.

Se examinaron las técnicas, métricas y la visualización de interrelaciones como herramientas útiles y potencialmente efectivas para analizar patrones de interacción. Al ser un estudio de caso, con datos particulares, los resultados del estudio no se pueden extrapolar o generalizar. Una limitación con los datos provistos fue la cantidad de vacíos y de información faltante que siguiendo procesos propios de AA como limpieza y transformación implicó una reducción considerable de estos. Estudios futuros podrían incorporar bases de datos con registros con menor información faltante, repetir N veces el número de experimentos en función de determinadas variables tales como región o tipo de zona con XGBoost, así como contrastar esta técnica contra otras distintas a las usadas en esta investigación, comparando el desempeño y resultados.

Agradecimientos

Las autoras agradecen al Living Lab & Data Hub del Institute for the Future of Education, Tecnológico de Monterrey, México, por prover la data publicada a través del Call "Bringing New Solutions to the Challenges of Predicting and Countering Student Dropout in Higher Education". Nuestros agradecimientos también van a los tres revisores anónimos por sus sugerencias. Rosa Morales agradece a Domingo Sifontes por los comentarios. Cualquier error u omisión es responsabilidad de las autoras.

Referencias

Aljohani, O. (2016). A Review of the Contemporary International Literature on Student Retention in Higher Education. *International Journal of Education and Literacy Studies*, *4*(1), 40-52. https://doi.org/10.7575/aiac.ijels.v.4n.1p.40

Álvarez Álvarez, C., & San Fabián Maroto, J. L. (2015). La elección del estudio de caso en investigación educativa. *Gazeta de Antropología, 28*(1).

Aulck, L., Velagapudi, N., Blumenstock, J., & West, J. (2017). Predicting student dropout in higher education. *arXiv*, 1606.06364v4. https://doi.org/10.48550/arXiv.1606.06364

Balanta Viera, V., & Palacios Medina, M. (2018). *Machine Learning: Redes Neuronales Artificiales*. Independently Published.

Barros, R., Basgalupp, M., Carvalho, A., & Freitas, A. (2012). A Survey of Evolutionary Algorithms for Decision-Tree Induction. *IEEE Transactions on Systems, Man and Cybernetics, Part C: Applications and Reviews, 42*(3), 291-312. https://doi.org/10.1109/TSMCC.2011.2157494

Batista, G. E. A. P. A. & Monard, M. C. (2010). An Analysis of Four Missing Data Treatment Methods for Supervised Learning. *Applied Artificial Intelligence*, *17*(5-6), 519-533. https://doi.org/10.1080/713827181

Bobadilla, J. (2021). *Machine Learning y Deep Learning: Usando Python, Scikit y Keras*. Ediciones de la U.

Burke, A. (2019). Student retention models in higher education: A literature review. *College and University*, *94*(2), 12-21.

Bonaldo, L., & Pereira, L. N. (2016). Dropout: Demographic profile of Brazilian university students. *Procedia-Social and behavioral sciences*, *228*, 138-143. https://doi.org/10.1016/j.sbspro.2016.07.020

Cabero-Almenara, J., Guillén-Gámez, F. D., Ruiz-Palmero, J., & Palacios-Rodríguez, A. (2022). Teachers' digital competence to assist students with functional diversity: Identification of factors through logistic regression methods. *British Journal of Educational Technology*, *53*(1), 41-57. https://doi.org/10.1111/bjet.13151

Caicedo Bravo, E. F., & López Sotelo, J. A. (2009). *Una aproximación práctica a las redes neuronales artificiales*. Programa Editorial Univalle.

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2023). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory& Practice*, *25*(1), 51-75. https://doi.org/10.1177/1521025120964920

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ASM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 785-794). ACM Press. https://doi.org/10.1145/2939672.2939785

Donoso, S., & Schiefelbein, E. (2007). Análisis de los modelos explicativos de retención de estudiantes en la universidad: una visión desde la desigualdad social. *Estudios Pedagógicos*, *33*(1), 7-27. https://doi.org/10.4067/S0718-07052007000100001

Espinosa-Zúñiga, J. J. (2020). Aplicación de algoritmos Random Forest y XGBoost en una base de solicitudes de tarjetas de crédito. *Ingeniería, Investigación y Tecnología*, *21*(3). https://doi.org/10.22201/fi.25940732e.2020.21.3.022

Ferreira, S. A. & Andrade, A. (2016) Academic analytics: Anatomy of an exploratory essay. *Education and Information Technology*, *21*, 229–243 https://doi.org/10.1007/s10639-014-9317-9

Ferreyra, M. M., Avitabile, C., Botero Álvarez, J., Haimovich Paz, F., & Urzúa, S. (2017). *At a crossroads: higher education in Latin America and the Caribbean*. World Bank Publications. https://doi.org/10.1596/978-1-4648-1014-5

Fike, D. S., & Fike, R. (2008). Predictors of first-year student retention in the community college. *Community College Review*, *36*(2), 68-88 https://doi.org/10.1177/0091552108320222

Fortin, N. M. (2006). Higher-education policies and the college wage premium: Cross-state evidence from the 1990s. *The American Economic Review*, *96*(4), 959-987. https://doi.org/10.1257/aer.96.4.959

Gitto, L., Minervini, L. F., & Monaco, L. (2016). University dropouts in Italy: Are supply side characteristics part of the problem? *Economic Analysis and Policy*, *49*, 108-116. https://doi.org/10.1016/j.eap.2015.12.004

González Fiegehen, L. E., & Espinoza Díaz, O. (2020). Deserción en educación superior en América Latina y el Caribe. *Paideia*, (45), 33-46.

Hernández, C., & Rodríguez, J. (2008). Preprocesamiento de datos estructurados. *Revista Vínculos, 4* (2) 27-48.

Herzog, S. (2005). Measuring determinants of student return vs. dropout/stopout vs. transfer: A first-to-second year analysis of new freshmen. *Research in Higher Education*, *46*(8), 883-928. https://doi.org/10.1007/s11162-005-6933-7

Huo, H., Cui, J., Hein, S., Padgett, Z., Ossolinski, M., Raim, R., & Zhang, J. (2023). Predicting Dropout for Nontraditional Undergraduate Students: A Machine Learning Approach. *Journal of College Student Retention: Research, Theory & Practice*, *24*(4), 1054–1077. https://doi.org/10.1177/1521025120963821

INEE. (2019). *Principales cifras nacionales. Educación básica y media superior. Inicio del ciclo escolar 2016-2017*. INNE. https://bit.ly/432IX4q

INEE. (2022). *Directrices para mejorar la permanencia escolar en la educación media superior*. INNE. https://bit.ly/3odsvPZ

Kirk, G. (2018). Retention in a Bachelor of Education (Early childhood studies) course: students say why they stay and others leave. *Higher Education Research & Development*, *37*(4), 773-787. https://doi.org/10.1080/07294360.2018.1455645

Luckman, M., & Harvey, A. (2019). The financial and educational outcomes of Bachelor degree non-completers. *Journal of Higher Education Policy and Management*, *41*(1), 3-17. https://doi.org/10.1080/1360080X.2018.1553106

Márquez-Vera, C., Cano, A., Romero, C., Noaman, A. Y. M., Mousa Fardoun, H., & Ventura, S. (2016). Early dropout prediction using data mining: a case study with high school students. *Expert Systems*, *33*(1), 107-124. https://doi.org/10.1111/exsy.12135

Melguizo, T., Torres, F. S., & Jaime, H. (2011). The association between financial aid availability and the college dropout rates in Colombia. *Higher Education*, *62*(2), 231-247. https://doi.org/10.1007/s10734-010-9385-8

Morales Salas, R. E., & Rodríguez Pavón, P. R. (2022). Retos y desafíos en la Educación Superior: una mirada desde la percepción de los docentes. *Education in the Knowledge Society, 23*, e264020. https://doi.org/10.14201/eks.26420

Nair, R., & Bhagat, A. (2019). Feature Selection Method to improve the accuracy of classification algorithm. *International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8*(6) 124-127.

Nieuwoudt, J. E., & Pedler, M. L. (2021). Student Retention in Higher Education: Why Students Choose to Remain at University. *Journal of College Student Retention: Research, Theory & Practice*. https://doi.org/10.1177/1521025120985228

Peláez, I. M. (2016). Modelos de regresión: lineal simple y regresión logística. *Revista Seden*, *14*, 195-214.

Rojas-López, A. (2017). Intervención de tres estrategias educativas para cursos de programación en educación superior *Education in the Knowledge Society*, *8*(4), 21-34. https://doi.org/10.14201/eks20171842134

Santos, A. C., Iglesias Rodríguez, A., & Pinto-Llorente, A. M. (2020). Identification of characteristics and functionalities for the design of an academic analytics model for Higher Education. *Proceeding of the TEEM'20: Eighth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 997-1003). ACM Press. https://doi.org/10.1145/3434780.3436578

Simpson, O. (2005). The costs and benefits of student retention for students, institutions, and governments. *Studies in Learning, Evaluation Innovation and Development*, *2*(3), 34-43.

Stoops, N. (2004). Educational attainment in the United States: 2003. *Current population*. https://bit.ly/3M8pN6e

Tinto, V. (1982). Limits of Theory and Practice in Student Attrition. *The Journal of Higher Education, 53*(6), 687-700. https://doi.org/10.2307/1981525

Von Hippel, P. T., & Hofflinger, A. (2021). The data revolution comes to higher education: identifying students at risk of dropout in Chile. *Journal of Higher Education Policy and Management*, *43*(1), 2-23. https://doi.org/10.1080/1360080X.2020.1739800