Depósito legal ppi 201502ZU4662
Esta publicación científica en formato
digital es continuidad de la revista impresa
Depósito Legal: pp 197402ZU789
● ISSN: 1315-9518 ● ISSN-E: 2477-9431

RCS
Revista de Ciencias Sociales

Universidad del Zulia. Revista de la Facultad de Ciencias Económicas y Sociales
Vol. XXX, Núm 1 ENERO-MARZO, 2024

Revista de Ciencias Sociales

50 Años
ANIVERSARIO

Esta publicación científica en formato
digital es continuidad de la revista impresa
Depósito Legal: pp 197402ZU789
ISSN: 1315-9518

Revista de Ciencias Sociales (RCS)
Vol. XXX, No. 1, Enero - Marzo 2024. pp. 452-469
FCES - LUZ ● ISSN: 1315-9518 ● ISSN-E: 2477-9431

Como citar: Villarreal-Torres, H., Ángeles-Morales, J., Marín-Rodriguez, W., y Cano-Mejía, J. (2024). Modelo de clasificación para la deserción estudiantil en las universidades públicas del Perú. *Revista De Ciencias Sociales*, *XXX*(1), 452-469.

Modelo de clasificación para la deserción estudiantil en las universidades públicas del Perú

Villarreal-Torres, Henry[*]
Ángeles-Morales, Julio[**]
Marín-Rodriguez, William[***]
Cano-Mejía, Jenny[****]

Resumen

Las tecnologías de información y comunicación cumplen un rol relevante en los diferentes campos del conocimiento, actualmente existe mayor capacidad para identificar patrones y anomalías en los datos de una organización utilizando la inteligencia artificial; el estudio tuvo como objetivo desarrollar un modelo de clasificación para la deserción estudiantil aplicando aprendizaje automático con el método autoML del framework H2O.ai, se ha tenido en cuenta la dimensionalidad de las características socioeconómicas y académicas. La metodología empleada fue de tipo predictivo y diseño no experimental, observacional y prospectivo; para ello, se aplicó un cuestionario de 20 ítems a 237 estudiantes de la Escuela de Posgrado matriculados en los programas de maestrías en educación. La investigación tuvo como resultado un modelo de aprendizaje automático supervisado, máquina de refuerzo de gradiente, para clasificar la deserción estudiantil, logrando así identificar los principales factores asociados que influyen en la deserción, obteniendo un coeficiente Gini del 92.20%, AUC del 96.10% y un LogLoss del 24.24% representando un modelo con desempeño eficiente. Se concluye que el modelo es apropiado por sus métricas de rendimiento, ofreciendo ventajas como trabajar con datos desequilibrados, validación cruzada y realizar predicciones en tiempo real.

Palabras clave: Aprendizaje automático; deserción estudiantil; educación superior; minería de datos; H2O.ai.

[*] Doctor en Ingeniería Informática y de Sistemas. Magíster en Ingeniería Informática y de Sistemas. Ingeniero Informático y de Sistemas. Docente Principal en la Universidad San Pedro, Chimbote, Perú. E-mail: henry.villarreal@usanpedro.edu.pe ORCID: https://orcid.org/0000-0002-5989-4534

[**] Docente en la Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima, Perú. Investigador RENACYT. E-mail: wmarin@unjfsc.edu.pe ORCID: https://orcid.org/0000-0002-7470-8154

[***] Doctor en Administracion. Magister en Administración Estratégica. Ingeniero Informático. Docente en la Universidad Nacional José Faustino Sánchez Carrión, Huacho, Lima, Perú. Investigador RENACYT. E-mail: wmarin@unjfsc.edu.pe ORCID: https://orcid.org/0000-0002-0861-9663 (Autor de vorrespondencia)

[****] Doctora en Gestión y Ciencias de la Educación. Magister en Obstetricia. Licenciada en Obstetricia. Docente Principal en la Universidad San Pedro, Chimbote, Perú. Investigadora RENACYT. E-mail: jenny.cano@usanpedro.edu.pe ORCID: https://orcid.org/0000-0001-5638-972X

Recibido: 2023-09-16 · **Aceptado:** 2023-12-05

Classification model for student dropout in public universities in Peru

Abstract

Information and communication technologies play a relevant role in different fields of knowledge. Currently, there is a greater capacity to identify patterns and anomalies in an organization's data using artificial intelligence; The study aimed to develop a classification model for student dropout by applying machine learning with the autoML method of the H2O.ai framework, taking into account the dimensionality of socioeconomic and academic characteristics. The methodology used was predictive and non-experimental, observational and prospective in design; To this end, a 20-item questionnaire was applied to 237 students from the Graduate School enrolled in master's degree programs in education. The research resulted in a supervised machine learning model, gradient boosting machine, to classify student dropout, thus identifying the main associated factors that influence dropout, obtaining a Gini coefficient of 92.20%, AUC of 96.10% and a LogLoss of 24.24% representing a model with efficient performance. It is concluded that the model is appropriate for its performance metrics, offering advantages such as working with unbalanced data, cross validation and making predictions in real time.

Keywords: Machine learning; student attrition; higher education; data mining; H2O.ai.

Introducción

La educación es fundamental para el desarrollo y el bienestar de una sociedad, por tanto, los estudiantes son la razón de ser de cualquier institución educativa. El desarrollo social y económico de un país está directamente relacionado con el rendimiento académico de sus estudiantes (Mushtaq y Khan, 2012). En el 2014 la Ley Universitaria No. 30220, crea la Superintendencia de Educación Superior Universitaria (SUNEDU), organismo que implementó el modelo de licenciamiento institucional. Ante la exigencia de cumplimiento de las condiciones básicas de calidad, es una buena opción gestionar la educación con las tecnologías de información según la propuesta de Villarreal-Torres et al. (2021); y, Briñez (2021), para tener la información disponible en el momento oportuno.

La deserción universitaria, es un problema relacionado al estudiante como responsable directo, generando preocupación en sus directivos por conocer las probabilidades de no culminación de estudios, influyendo negativamente en el desarrollo académico y económico de la institución; motivo por el cual, se pretende identificar patrones de comportamiento en los estudiantes, mediante la minería de datos, analizando los factores socioeconómicos y académicos para implementar estrategias específicas que coadyuven a mantener una economía sostenible en el tiempo, evitando el alto índice de riesgo de abandono de estudios. Específicamente en el ámbito de las universidades, y particularmente en las escuelas de postgrado, resulta necesario cumplir con estándares de calidad en lo referente a la oferta del servicio educativo (Díaz-Landa, Meleán-Romero y Marín-Rodriguez, 2021).

En el Perú se ha incrementado significativamente antes y aún más después de la pandemia de Covid-19, es por ello, que las universidades públicas necesitan identificar e implementar programas para disminuir la deserción estudiantil (Valero et al. 2022). Este problema se agudizo durante el Covid-19 ocasionando un impacto negativo en la mayoría de los sectores productivos, conduciendo a algunas instituciones

educativas a implementar estrategias para revertir la situación de abandono de estudios (Moreno et al., 2021; Félix, Urrea y López, 2023; Villarreal-Torres et al., 2023). Por ello "son múltiples las aplicaciones de inteligencia artificial [que] utilizan técnicas de minería de datos para descubrir patrones importantes y obtener información útil de sistemas de información de registros académicos" (Díaz et al. 2022, p. 198).

El informe de la Organización para la Cooperación y el Desarrollo Económicos (Organisation for Economic Co-operation and Development [OECD], 2019), indica que el 39% de los estudiantes a tiempo completo que ingresan a un programa se gradúan dentro de la duración teórica; asimismo, la tasa promedio de finalización posterior a los tres años adicionales corresponde a un incremento del 67%. Por otra parte, el 12% de ingresantes a un programa a tiempo completo abandonan sus estudios antes del inicio del segundo; asimismo, muestra un incremento del 20% al final de la duración teórica y al 24% posterior a los tres años.

En el Perú, las cifras sobre la evolución de matrículas según la Superintendencia Nacional de Educación Superior (SUNEDU, 2020) en el nivel de pregrado durante el 2018, fue de 1.59 millones cifra que se ha reducido en 1.34 millones de estudiantes en el 2020, interpretado con un 15,7% de diferencia entre los periodos; en el caso, de posgrado se tiene una reducción de 27,7%, puesto que durante el 2018 se tuvo 131.9 mil y en el periodo 2020 se contó con 95.4 mil estudiantes matriculados. Según el Diario Oficial El Peruano (2021), se detalla que en las universidades licenciadas a nivel nacional indican que el porcentaje de interrupción de estudios ha decrecido en 4,7%; es decir, de un 16,2% ha disminuido a un 11,5% entre los semestres 2020-II y 2021-I.

La investigación estuvo enmarcada en la producción de un nuevo conocimiento mediante la propuesta del modelo de clasificación, además se corroboró la teoría de deserción estudiantil sostenida por Díaz (2008). El objetivo de la investigación fue desarrollar un modelo de clasificación de deserción en estudiantes de los programas de estudio de educación mediante aprendizaje automático y técnicas de minería de datos aplicando autoML de *H2O.ai*, a fin de que los estudiantes, con potencial de deserción, puedan ser identificados por las autoridades para tomar las medidas correctivas pertinentes.

1. Fundamentación teórica

1.1. Minería de datos

La minería de datos utiliza el análisis matemático y estadístico para encontrar patrones y tendencias en grandes conjuntos de datos. La exploración de datos tradicional no puede descubrir estos patrones debido a la complejidad o a las grandes cantidades de datos (Microsoft Learn, 2023). Utilizan métodos estadísticos y algoritmos de inteligencia artificial para encontrar patrones en conjuntos de datos masivos (Camborda, 2014). Sus métodos de clasificación, agrupación y predicción hacen que tenga éxito (Zárate-Valderrama et al., 2021). Dole y Rajurkar (2014), pronostican la culminación y el estado de aprobado/reprobado utilizando el algoritmo Naive Bayes y el árbol de decisión.

En definitiva, la minería de datos debe utilizarse con cuidado y responsabilidad para garantizar que se respeta el derecho a la privacidad de las personas y se obtengan conclusiones precisas y útiles. Es una técnica importante que ha transformado la forma en que las organizaciones gestionan y toman decisiones basadas en grandes cantidades de datos.

1.2. Aprendizaje automático

Kodelja (2019), sostiene que es un subconjunto de la inteligencia artificial; además, afirma que es aprendizaje y no otra cosa; mientras que otros -incluidos los filósofos- rechazan la afirmación que es un aprendizaje real, para ellos, el aprendizaje real es la forma más elevada del aprendizaje

humano. Por su parte, Xu y Li (2014), manifiestan que es un método esencial para tratar los problemas de adquisición de conocimientos; se refiere a la construcción y el estudio de sistemas que pueden aprender de los datos.

Samuel (2000), lo define como el campo de estudio donde los ordenadores tienen la capacidad de aprender, sin ser programados explícitamente. Dwi, Prasetya y Pujianto (2018), sostienen que se enfoca en desarrollar un sistema que sea capaz de aprender de sus propios patrones sin intervención humana, su aplicación se encuentra en varios campos.

El aprendizaje automático, es la capacidad de los sistemas informáticos para aprender y evolucionar de forma autónoma a partir de datos a través del tiempo; el cual se está convirtiendo en una herramienta indispensable para la adquisición de conocimientos en diversas áreas; aunque con algunas limitaciones, sus aplicaciones son innovadoras y eficientes para la solución de problemas reales.

1.3. Tipos de aprendizaje automático

Jung (2022), describe dos tipos de aprendizaje automático; el primero, como aprendizaje supervisado, que emplea un conjunto de datos etiquetados para su predicción, se divide en regresión y clasificación; el segundo, como aprendizaje no supervisado, al conjunto de datos que no necesita etiquetas; permite a los analistas descubrir patrones de comportamientos o similitudes entre las características, solo se basa en la subdivisión o el agrupamiento (Chatterjee et al., 2023). Existiendo el aprendizaje por refuerzo, similar al aprendizaje no supervisado, puede evaluar la función de perdida; en estos casos, aprende de las experiencias de prueba y error dependiendo de la retroalimentación y su factor o agente para tener un desempeño eficiente (Sharmeela et al., 2023).

Es de vital importancia conocer las múltiples formas de aprendizaje automático y sus propias características, fortalezas y debilidades de cada una de ellas; en tal sentido, es esencial la selección del tipo de aprendizaje automático para desarrollar modelos de predicción en la solución de problemas originados en diversas ramas del conocimiento.

1.4. AutoML

AutoML, es el aprendizaje automático de las maquinas, Nagarajah y Poravi (2019), lo describen como un proceso que tiene la capacidad de elaborar modelos a la medida, reduciendo de manera considerable la intervención de las personas; además, de realizar el preprocesamiento de los datos, la ingeniería de variables, la construcción de modelos, la optimización de hiperparámetros y el análisis de los resultados de las predicciones y su respectiva evaluación.

El desarrollo del aprendizaje automático de máquinas ha permitido, en gran medida, agilizar las operaciones de desarrollo del aprendizaje de máquina que requieren mucho tiempo, pretendiendo reducir la demanda de los científicos de datos y tener la capacidad de construir aplicaciones de aprendizaje automático de buen rendimiento, sin necesidad de tener amplios conocimientos de estadística y aprendizaje de máquinas (Zöller y Huber, 2021).

Mediante la implementación del autoML, se puede lograr la automatización del proceso de desarrollo del aprendizaje automático, lo que a su vez hace posible producir aplicaciones de aprendizaje automático de alto rendimiento de una manera rápida y eficiente, sin la necesidad de tener amplios conocimientos de estadística e informática. Actualmente, el número de librerías desarrolladas ha aumentado significativamente, lo que hace posible que las organizaciones desplieguen soluciones innovadoras de una manera simple y eficaz.

1.5. Plataforma H2O.ai

LeDell y Poirier (2020), expresan

que H2O es una plataforma de aprendizaje automático distribuido de código abierto, se creó para escalar a conjuntos de datos extremadamente grandes. Sus interfaces de programación de aplicaciones (API) están escritas en *R*, *Python*, *Java* y *Scala*. Los pasos para realizar el proceso de automatización mediante H2O.autoML son: La recopilación de datos, exploración de datos, preparación de datos, transformación de datos, selección del modelo, entrenamiento del modelo, ajustes de hiperparámetros y finalmente, la predicción (Ajgaonkar, 2022).

La plataforma H2O, es una herramienta que viene ganando popularidad para quienes trabajan con enormes conjuntos de datos y buscan automatizar el proceso de aprendizaje automático; además, cuenta con interfaces de programación de aplicaciones (API) haciéndola accesible para usuarios avanzados de la comunidad de aprendizaje automático.

1.6. Selección de características

Para el desarrollo de un modelo de aprendizaje automático, es necesario realizar la selección de características, tiene como finalidad identificar la interacción de las variables dependientes para tener el mejor desempeño predictivo; este proceso es relevante porque permite conocer las variables que aportan significativamente al modelo predictivo, permitiendo así, reducir el número de variables, tiempo, velocidad y despliegue; haciendo que el modelo sea menos complejo y más fácil de explicar (Haque, 2022).

Se tiene tres clases de métodos para la selección de características según Khun y Jhonson (2019): Los métodos intrínsecos, comprenden a los modelos basados en árboles y reglas, los modelos multivariados de regresión adaptativa y los modelos de regularización; los métodos de filtro, son simples y rápidos mediante un análisis supervisado determinan las características, son propensos a sobre seleccionar predictores en el modelo. Finalmente, los métodos de envoltura, que usan procedimientos de búsqueda iterativos, proporcionando subconjuntos de predictores para el modelo teniendo mayor eficacia en el rendimiento de la predicción.

El proceso de selección de características, es un paso esencial en la construcción de modelos de aprendizaje automático, donde se utilizan a menudo enfoques como las técnicas intrínsecas, de filtro y de envoltura, para identificar las variables que aportan significativamente al modelo predictivo; además, la selección de características tiene como propósito la reducción de recursos que conlleva a una adecuada comprensión e interpretación del modelo desarrollado. En grandes volúmenes de datos, la selección de características puede conllevar a resultados sesgados o incompletos.

1.7. Deserción estudiantil

Tinto (1982); y, Félix et al. (2023), definen la deserción como una situación en la que un estudiante no logra terminar su educación o se aleja de ella de manera temporal o permanente; por lo tanto, un desertor sería aquel que está inscrito en una institución de educación superior, pero no presenta actividad académica durante tres semestres académicos seguidos. González (2005), diferencia dos tipos de abandono en la educación superior universitaria; la primera, con respecto al tiempo (inicial, temprana y tardía); y la segunda, con respecto al espacio (institucional, interna y del sistema educativo).

Tinto (1989), afirma que durante el periodo de transición se producen los abandonos; específicamente, y tal como lo señalan Duche et al. (2020), la transición secundaria-universitaria, siendo los más frecuentes los abandonos voluntarios. Díaz (2008), presentó los modelos de análisis de la deserción estudiantil, con el propósito de analizar el fenómeno de la deserción inherente a la vida estudiantil universitaria, describiendo las teorías desde diversos puntos de vista:

a. Modelo psicológico: Indica los rasgos de personalidad que establecen las diferencias entre los estudiantes que culminan y abandonan sus estudios universitarios; se

fundamenta en las propuestas de Fhisbein y Ajzen (1974), quienes sostienen la Teoría de la Acción Razonada; Ethington (1990), quien se basa en el Modelo de Elección Académica sostenido por Eccles, Adler y Meece (1984), para insertar teorías sobre conductas de logro, como el rendimiento académico que afecta al estudiante. Finalmente, Bean y Eaton (2001) fundamentan los procesos psicológicos con la integración académica y social sustentados en cuatro teorías psicológicas: Teoría de Actitud y Comportamiento; Teoría del Comportamiento de Copia, la Habilidad para Entrar y Adaptarse a un Nuevo Ambiente; la Teoría de Autoeficacia; y, la Teoría de Atribución.

b. Modelo sociológico: Hace énfasis en los factores externos de los estudiantes, los cuales influencian en la deserción estudiantil; Spady (1970), manifiesta que una de las causas de la deserción, es la integración social en la universidad, generada por las influencias, expectativas y demandas dadas en el medio familiar. Asimismo, propone seis predictores para la deserción estudiantil: Integración académica, integración social, estado socioeconómico, género, calidad de carrera y el promedio de cada semestre.

c. Modelo económico: Está basado en dos modelos: El primero, Costo/Beneficio, está relacionado a los beneficios sociales y económicos que perciben los estudiantes para permanecer en la universidad; el segundo, Focalización del Subsidio, está orientado a los estudiantes con bajos recursos o limitaciones para costear sus estudios (Cabrera, Nora y Castañeda, 1992; 1993; Bernal, Cabrera y Terenzini, 2000; St. John et al., 2000).

d. Modelo organizacional: Se fundamenta en la forma cómo la organización integra a los estudiantes (Berger, 2000; 2001; Kuh 2002).

e. Modelo de interacción: Sostiene que la permanencia en la institución está en función del grado de acoplamiento del estudiante con la institución (Tinto, 1982), se complementa con el modelo de Spady (1970), en el que se incorpora la teoría de intercambio de Nye (1976).

La deserción estudiantil en el sistema universitario, es un problema complejo ocasionado por diversos factores como sociales, económicos, personales, familiares, académicos, psicológicos, entre otros, desarrollados dentro de su entorno y experiencias; los cuales deben ser analizados desde diferentes puntos de vista con el propósito de brindar una solución integral y permita a los estudiantes finalizar sus estudios. La reducción de la deserción estudiantil puede lograrse desde un análisis de la personalidad, seguido de la integración social y académica, optimización de costos y beneficios brindados por el servicio educativo, hasta el grado de articulación o acoplamiento entre el estudiante y la institución.

1.8. Dimensiones de la deserción estudiantil

Las variables consideradas, con mayor frecuencia, en los modelos teóricos relacionados a la deserción estudiantil fueron consolidadas en el estudio realizado por Díaz (2008), se consideran cuatro categorías, las individuales (edad, género, grupo familiar e integración, social); las académicas (orientación profesional, desarrollo intelectual, rendimiento académico, métodos de estudios, procesos de admisión, grados de satisfacción de la carrera y carga académica); las institucionales (normativas académicas, financiamiento estudiantil, recursos universitarios, calidad del programa o carrera y relación con los profesores y pares); y las socioeconómicas (estrato socioeconómico, situación laboral del estudiante, situación laboral de los padres y nivel educacional de los padres).

2. Metodología

La metodología utilizada estuvo basada en el enfoque cuantitativo, en virtud al análisis y procesamiento de datos numéricos para detectar patrones y relaciones entre las variables de estudio; con respecto al tipo

de investigación corresponde un estudio predictivo, cuya finalidad es desarrollar un modelo de predicción mediante las técnicas de minería de datos, aprendizaje automático y estadísticas. Así mismo, el diseño fue no experimental, observacional y prospectivo (Supo, 2020).

El conjunto de datos fue obtenido de dos fuentes de información, en primer lugar, mediante la aplicación de un cuestionario como instrumento, que contiene 20 *ítems* agrupados en cuatro dimensiones, aplicándose a 237 participantes de la Escuela de Posgrado de la Universidad Nacional José Faustino Sánchez Carrión matriculados en los programas de maestrías en educación, seleccionados mediante muestreo aleatorio simple; en segundo lugar, se recopilaron datos del registro de evaluaciones mediante la observación. A continuación, se presentan los ítems en el Cuadro 1.

Cuadro 1
Instrumento de recolección de datos para los participantes

N	Pregunta	Tipo
P01	Rendimiento académico en secundaria	Ordinal
P02	Asignaturas desaprobadas en secundaria	Ordinal
P03	Repitencia de año en secundaria	Dicotómico
P04	Rendimiento académico en pregrado	Ordinal
P05	Asignaturas desaprobadas en pregrado	Ordinal
P06	Sexo	Dicotómico
P07	Rango edad	Ordinal
P08	Estado civil	Ordinal
P09	Empleado adecuadamente	Ordinal
P10	Número de hijos	Ordinal
P11	Ingreso familiar	Ordinal
P12	Motivación para el estudio	Dicotómico
P13	Situación económica	Ordinal
P14	Financiamiento de estudios	Dicotómico
P15	Disponibilidad de tiempo de estudio	Ordinal
P16	Nivel de estrés	Ordinal
P17	Infraestructura adecuada	Ordinal
P18	Equipamiento y mobiliario adecuado	Ordinal
P19	Asignaturas pertinentes	Ordinal
P20	Nivel de docentes	Dicotómico

Fuente: Elaboración propia, 2023.

En base a la revisión de la literatura que fundamenta la deserción estudiantil, se ha considerado la teoría de Díaz (2008), quien adaptó las teorías propuestas al contexto de la realidad peruana elaboradas por Spady (1970); y, Tinto (1989), en cuatro factores, como se detalla en la Tabla 1.

Tabla 1
Descripción *ítems* según factores propuesta de Díaz (2008)

N	Factores	*Ítems*	
		Inicio	Final
01	Académicos	01	05
02	Individuales	06	12
03	Ambientales	13	16
04	Institucionales	17	20

Fuente: Elaboración propia, 2023.

Para el desarrollo del modelo, se utilizó el lenguaje *R Statistical Software* (v4.2.2; R Core Team, 2022) y con el entorno de desarrollo *R Studio* (v2022.12.0 *Build* 353; RStudio Team, 2022) ejecutado desde el sistema operativo de escritorio *Windows* 11 (x64 *Build* 22621); así mismo, se empleó la plataforma H20.ai para la generación del modelo de clasificación a través del paquete, H2O (v 3.38.0.1; Fryda et al., 2022). Para la reducción de la dimensionalidad mediante la selección de características se utilizaron los paquetes: Familiar (v1.4.1; Zwanenburg y Löck, 2021); *Information* (v0.0.9; Larsen, 2016); *Boruta* (v8.0.0; Kursa y Rudnicki, 2010); *Regularized Random Forest*, RRF (v1.9.4; Deng, 2013); y, *FSinR* (v2.0.5; Aragón-Royón et al., 2020).

3. Resultados y discusión

A continuación, se presenta el análisis descriptivo de las opiniones emitidas por los participantes a través del cuestionario, según la Tabla 2, los resultados indican variabilidad en las respuestas. Así mismo, para desarrollar estos modelos, se definieron variables independientes, que corresponde a 20 *ítems* del instrumento y como variable dependiente, la deserción estudiantil; además, se ha considerado dos aspectos de vital importancia: La selección de características y el porcentaje para la partición del conjunto de datos para entrenamiento, validación y prueba para cada uno de los modelos.

Para la selección de las características se utilizaron diferentes algoritmos, obteniendo dos conjuntos de variables en base a las coincidencias o similitudes en común; el primer conjunto, conformado por 11 variables (P01, P02, P03, P04, P09, P10, P12, P13, P14, P16, P20); y el segundo conjunto, conformado por las cinco variables (P07, P11, P17, P18, P19), haciendo un total de 16 variables.

Tabla 2
Análisis descriptivo del conjunto de datos de los participantes

N	Etiq.	Descripción	Min	Max	Mean	DE
01	P01	Rendimiento académico en secundaria	1	5	3.633	0.977
02	P02	Asignaturas desaprobadas en secundaria	1	4	1.578	0.786
03	P03	Repitencia de año en secundaria	1	2	1.932	0.251
04	P04	Rendimiento académico en pregrado	2	5	3.443	0.879
05	P05	Asignaturas desaprobadas en pregrado	1	3	1.266	0.530
06	P06	Sexo	1	2	1.624	0.485
07	P07	Rango edad	1	3	2.004	0.805
08	P08	Estado civil	1	5	1.975	0.786
09	P09	Empleado adecuadamente	1	2	1.831	0.375
10	P10	Número de hijos	1	3	1.916	0.714
11	P11	Ingreso familiar	2	5	3.013	0.773

Cont... Tabla 2

12	P12	Motivación para el estudio	1	2	1.038	0.192
13	P13	Situación económica	2	5	3.194	0.773
14	P14	Financiamiento de estudios	1	2	1.068	0.251
15	P15	Disponibilidad de tiempo de estudio	1	5	3.118	1.477
16	P16	Nivel de estrés	1	5	2.970	1.418
17	P17	Infraestructura adecuada	1	5	3.084	1.369
18	P18	Equipamiento y mobiliario adecuado	1	5	2.924	1.376
19	P19	Asignaturas pertinentes	1	5	2.911	1.419
20	P20	Nivel de docentes	1	5	3.650	1.012

Fuente: Elaboración propia, 2023.

Posteriormente, se establecieron los parámetros para la invocación del método AutoML del objeto H2O, considerando como parámetros de datos, el conjunto de las variables independientes y luego la variable objetivo o de destino, definida como la variable dependiente; el parámetro de parada o de finalización, se consideró max_models = 100; además, de la opción balance_classes = TRUE.

Con esta configuración se presentan en la Tabla 3, los resultados de las 10 ejecuciones o iteraciones realizadas según la configuración definida; se muestra en síntesis los principales modelos de aprendizaje automático con mejores métricas de entrenamiento en comparación con otros modelos ubicados en posiciones inferiores; por ejemplo, *Extremely Randomized Trees* (XRT) y *Distributed Random Forest* (DRF), *Generalized Linear Model* (GLM). A continuación, se presentan las métricas del proceso de entrenamiento de cada uno de los modelos generados automáticamente.

Tabla 3
Modelos de aprendizaje automático según el tamaño de los conjuntos de datos

N	Modelo	*Ítems*	Conjunto de Datos		
			Entrenamiento	Prueba	Validación
01	DeepLearning Grid	16	70	30	0
02	DeepLearning Grid	11	70	30	0
03	GBM Grid	16	70	15	15
04	DeepLearning Grid	11	70	15	15
05	GBM Grid	16	80	20	0
06	GBM Grid	11	80	20	0
07	GBM Grid	16	60	40	0
08	GBM Grid	11	60	40	0
09	GBM Grid	16	75	25	0
10	GBM Grid	11	75	25	0

Fuente: Elaboración propia, 2023.

Como se aprecia en la Tabla 4, las puntuaciones obtenidas en cada métrica son muy similares y significativas durante el proceso de entrenamiento y validación, se observa valores óptimos de rendimiento en cada modelo según el tamaño de los conjuntos de datos de la Tabla 3; realizándose posteriormente, las pruebas para obtener las métricas de rendimiento de cada uno de los modelos indicados.

Tabla 4
Métricas de rendimiento de los modelos de entrenamiento y validación

N	Modelo	Ítems	AUC	LOGLOS	AUCPR
01	DeepLearning Grid	16	0.981685	0.389653	0.956428
02	DeepLearning Grid	11	0.981136	0.214359	0.951164
03	GBM Grid	16	0.980220	0.183851	0.943741
04	DeepLearning Grid	11	0.982784	0.196832	0.954476
05	GBM Grid	16	0.972311	0.258593	0.923799
06	GBM Grid	11	0.972603	0.204378	0.932085
07	GBM Grid	16	0.974163	0.246842	0.915569
08	GBM Grid	11	0.972010	0.207276	0.920860
09	GBM Grid	16	0.977618	0.218077	0.925325
10	GBM Grid	11	0.972982	0.201235	0.923862

Fuente: Elaboración propia, 2023.

Los modelos de clasificación tienen una variedad de métricas de rendimiento entre las de mayor relevancia se tiene el coeficiente de Gini, el cual es empleado para medir la calidad del modelo de predicción, teniendo como interpretación, que una valoración de cero significa una igualdad perfecta, es decir, se tiene un modelo deficiente; cuanto tiene un valor cercano a la unidad, se presenta como desigualdad máxima, y se considera un clasificador perfecto.

La Tabla 5, contiene las métricas de las ejecuciones y pruebas realizadas con cada uno de los modelos generados automáticamente, como se evidencia las métricas son similares a diferencia del tercero y cuarto modelo que se encuentran sobre ajustados, debido al número de observaciones particionadas en tres conjuntos de datos. Asimismo, se muestra un mejor desempeño en las métricas de los modelos con menor número de *ítems*; en este sentido, por el principio de parsimonia, se opta por aquellos con 11 *ítems* según los algoritmos utilizados para la selección de características, permitiendo beneficios para su futura implementación. Se observa ligeramente una mejor prestación en el décimo modelo *Gradient Boosting Machine*, seguido por el segundo modelo *DeepLearning*.

Tabla 5
Métricas de rendimiento de los modelos de pruebas

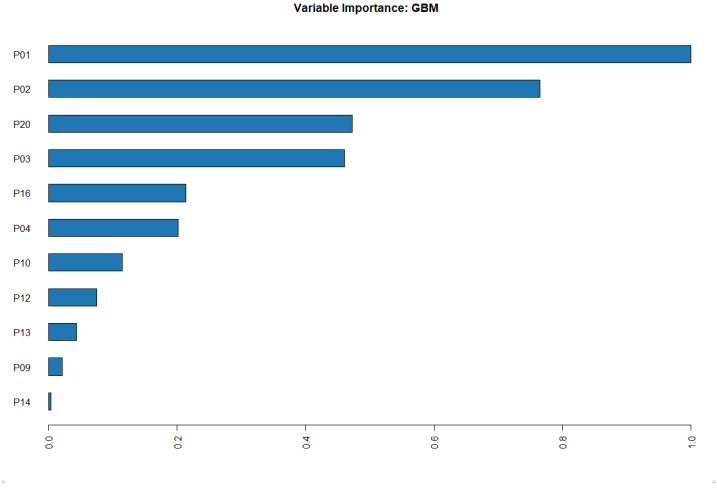
N	Modelo	Ítems	GINI	AUC	AUCPR	LOGLOSS
01	DeepLearning Grid	16	0.895981	0.947991	0.913763	0.850491
02	DeepLearning Grid	11	0.865248	0.932624	0.905851	0.546854

Cont... Tabla 5

03	GBM Grid	16	1.000000	1.000000	1.000000	0.025920
04	DeepLearning Grid	11	1.000000	1.000000	1.000000	0.044860
05	GBM Grid	16	0.915633	0.957816	0.911510	0.312979
06	GBM Grid	11	0.935484	0.967742	0.937704	0.259712
07	GBM Grid	16	0.943012	0.971506	0.919590	0.293444
08	GBM Grid	11	0.932157	0.966079	0.925879	0.217350
09	GBM Grid	16	0.912281	0.956140	0.922686	0.270146
10	GBM Grid	11	0.898246	0.949123	0.911629	0.295948

Fuente: Elaboración propia, 2023.

El Gráfico I, contempla las variables ordenadas de mayor a menor según la importancia en la predicción del modelo, en base a los valores porcentuales que se encuentran escalados al 100%. Se evidencia una influencia fuerte en la experiencia de los participantes en el nivel de secundaria: Rendimiento académico (29,65%), asignaturas reprobadas (22,67%) y repetición de año (13,65%); el desempeño de los docentes (14,03%); en menor relevancia se encuentran los aspectos relacionados a estrés de la persona (6,35%), rendimiento en pregrado (5,99%), el número de hijos (3,40%), motivación (2,23%), situación económica (1,28%), trabajo relacionado a su carrera (0,62%), y finalmente, el financiamiento de sus estudios (0,10%).



Fuente: Elaboración propia, 2023.

Gráfico I: ***Importancia de las variables en el modelo de clasificación***

La exactitud es una métrica para determinar las predicciones correctas como proporción al total de predicciones realizadas, una puntuación cercana a la unidad representa un rendimiento óptimo. De la Tabla 6, se puede obtener una precisión equivalente a un 92%, es decir, el modelo tiene una capacidad de predicción puesto que de 100 observaciones alcanza predecir 92 casos exitosamente; para la sensibilidad se tiene un 90%, indicando una predicción que, de 100 casos, 90 son exitosos para la clase positiva; finalmente, para la especificidad, identifica un 100% de los casos para predecir la clase negativa.

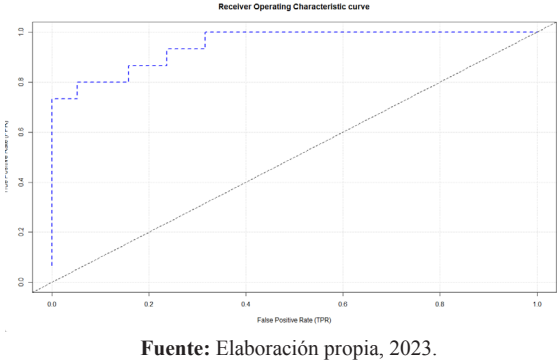
Tabla 6
Matriz de confusión del modelo GBM generado

Valores Predicción	Reales		Error	Ratio
	Positivo	Negativo		
Positivo	38	0	0.000	= 0 / 38
Negativo	4	11	0.267	= 4 / 15
Total	42	11	0.075	= 4 / 53

Fuente: Elaboración propia, 2023.

El área bajo la curva, es una métrica para evaluar la capacidad del modelo de clasificación, permitiendo diferenciar entre los verdaderos positivos y falsos positivos; un valor cercano a la unidad, se considera un modelo perfecto. A diferencia de la métrica área bajo la curva precisión - recuperación, no considera los verdaderos negativos muy utilizado en conjunto de datos desequilibrados. La métrica de pérdida logarítmica analiza la aproximación de los valores predichos de un modelo y las valoraciones del objetivo real, donde una asignación cercana a cero significa que el modelo proporciona correctamente la probabilidad.

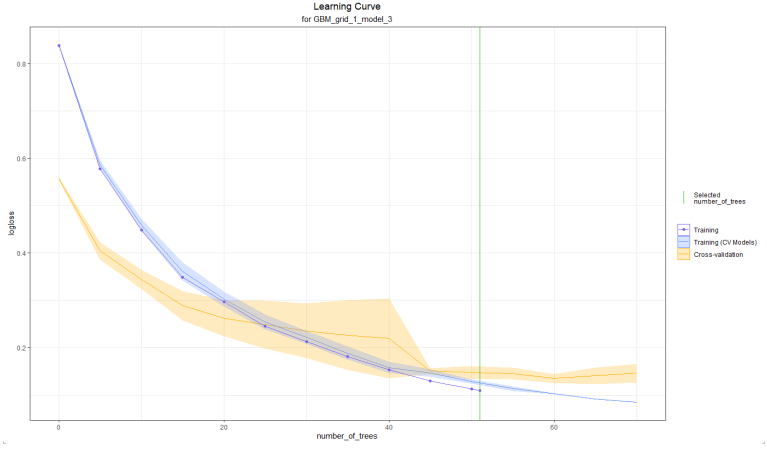
La curva ROC, es un gráfico que representa la relación entre verdaderos positivos (sensibilidad) y falsos positivos (especificidad), el Gráfico II, demuestra una curva cercana a la esquina superior izquierda, indicando así un rendimiento óptimo. Cabe precisar que, cuando la curva se aproxima a la diagonal de 45° o línea base, será menos precisa correspondiendo un desempeño deficiente. Asimismo, el lado inferior izquierdo del gráfico representa una menor tolerancia a los falsos positivos; mientras que el lado superior derecho representa una mayor tolerancia a los falsos positivos.



Fuente: Elaboración propia, 2023.

Gráfico II: **Gráfico ROC del modelo de clasificación GBM**

El Gráfico III, muestra el comportamiento del modelo de clasificación GBM mediante la curva de aprendizaje, presenta una perdida logarítmica en el conjunto de datos de entrenamiento y validación; además se visualiza que las curvas son estables al tener un número superior a los 50 árboles, es decir, al agregar más instancias al modelo, este no mejoraría mucho su desempeño o rendimiento.



Fuente: Elaboración propia, 2023.

Gráfico III: **Curva de aprendizaje del modelo de clasificación GBM**

En síntesis, el modelo GBM (*Gradient Boosting Machine*) es un método de aprendizaje automático supervisado que se utiliza para realizar la clasificación de problemas de aprendizaje automático. Está construido utilizando árboles de decisión. El modelo GBM generado consta de 51 árboles internos, con un tamaño correspondiente a 8,910 *bytes*.

El árbol tiene una profundidad mínima de 4 y una profundidad máxima de 6, con una profundidad promedio de 5.29. El número mínimo de hojas es de 7 y el número máximo es de 13, con un promedio de 9.24 hojas. Esta configuración del modelo GBM indica que los árboles de decisión internos tienen una profundidad razonable y un número moderado de hojas. Esto significa que el modelo GBM presenta una buena capacidad de ajuste y puede proporcionar una buena clasificación para los datos evidenciado por las métricas de rendimiento.

Al contrastar los resultados obtenidos con la fundamentación teórica, se puede indicar que, el modelo de clasificación ha sido posible mediante la utilización de técnicas de minería de datos para identificar patrones y tendencias que pueden ser útiles para predecir a los estudiantes con riesgo de deserción. Sin embargo, es solo una herramienta y es necesaria la intervención humana para proporcionar el apoyo emocional y académico a los estudiantes en riesgo, coincidiendo con lo indicado por Zárate-Valderrama et al. (2021); Jung (2022); y, Microsoft Learn (2023), se indica también, que Dole y Rajurkar (2014), desarrollaron un modelo de clasificación binaria mediante Naive de Bayes; en el presente estudio, fue un modelo de clasificación GBM.

El desarrollo del modelo ha conllevado los procesos de entrenamiento, validación y prueba con diversos conjuntos de datos obteniendo métricas de rendimiento eficaces concordando con el estudio de Xu y Li (2014);

además, se coincide con las investigaciones realizadas por Samuel (2000); y, Dwi et al. (2018), sobre la capacidad de los sistemas de información para aprender mediante los algoritmos AutoML y el uso de la plataforma H2O.ai expresado por LeDell y Poirier (2020).

Debido a la complejidad de la deserción estudiantil, esta fue analizada íntegramente mediante los cinco modelos propuestos por Díaz (2008), considerándose como base para la elaboración de los instrumentos de recolección de datos, consolidándose en 20 *ítems*; de los cuales, fueron utilizados sólo 11 *ítems* para el modelo de clasificación debido al proceso de selección de características (Haque, 2022), siendo los *ítems* de mayor relevancia P01, P02, P20 y P03.

Conclusiones

En vista de los resultados, se evidencia el desarrollo de un modelo GBM para la clasificación de la deserción estudiantil utilizando la plataforma H2O.ai y AutoML, se puede concluir que presenta un rendimiento eficiente debido a las métricas de precisión, sensibilidad y especificidad para identificar patrones en los estudiantes con riesgo de abandonar sus estudios; ofrece ventajas como la capacidad de trabajar con datos desbalanceados, la capacidad de mejorar los resultados mediante la sintonización de los parámetros, el uso de la validación cruzada y la capacidad de realizar predicciones en tiempo real, considerándose como herramienta útil para la toma de decisiones.

Un aspecto relevante de la investigación fue la transversalidad, en primera instancia el aprendizaje automático, tuvo la capacidad de utilizar los algoritmos para extrapolar los conocimientos adquiridos en un conjunto de datos; para el caso de la minería de datos, esta técnica ha permitido identificar patrones en los datos dentro del contexto de la educación superior universitaria, permitiendo a los usuarios compartir y reutilizar conocimientos adquiridos y mejores prácticas en otras áreas del conocimiento.

Respecto al aporte científico, la investigación es significativa y se presenta desde diferentes perspectivas; desde el punto de vista teórico, permite conocer y comprender los factores que influyen en la deserción de estudiantes contribuyendo de manera general al conocimiento en el campo de la inteligencia artificial y el aprendizaje automático; desde el punto de vista práctico, las instituciones de educación superior pueden implementar estrategias y programas de retención a los estudiantes en riesgo y evitar el abandono de los estudios.

Las limitaciones a considerar en el desarrollo de un modelo de clasificación es el tamaño de conjunto de datos, la selección de características, la discretización de las variables, datos desbalanceados, dichos factores conllevan a sesgos y predicciones inexactas; por otra parte, el modelo desarrollado funciona para un contexto especifico debido a la influencia de las variables independientes en la deserción estudiantil, las cuales pueden cambiar con el tiempo o entorno.

Las futuras líneas de investigación a desarrollar pueden incluir otros tipos de aprendizaje automático como aprendizaje profundo, ensamblajes, entre otros, así como la incorporación de conjuntos de datos no estructurados; además, se pueden incluir otras características y/o factores que influyen en la deserción estudiantil y que varían dependiendo del entorno. También se puede considerar estudios sobre la efectividad de las intervenciones basadas en las predicciones del modelo de clasificación.

Referencias bibliográficas

Ajgaonkar, S. (2022). *Practical automated machine learning using H2O.ai: Discover the power of automated machine learning, from experimentation through to deployment to production*. Packt Publishing.

Aragón-Royón, F., Jiménez-Vílchez, A.,

Arauzo-Azofra, A., y Benitez, J. (2020). FSinR: An exhaustive package for feature selection. *arXiv: 2002. 10330*. https://doi.org/10.48550/arXiv.2002.10330

Bean, J., y Eaton, S. B. (2001). The psychology underlying successful retention practices. *Journal of College Student Retention: Research, Theory & Practice, 3*(1), 73-89. https://doi.org/10.2190/6R55-4B30-28XG-L8U0

Berger, J. B. (2000). Organizational behavior in higher education and student outcomes. In J. C. Smart (Ed.), *Higher Education: Handbook of theory and research* (Vol. XV, pp. 268-338). Agathon Press.

Berger, J. B. (2001). Understanding the organizational nature of student persistence: Empirically based recommendations for practice. *Journal of College Student Retention: Research, Theory and Practice, 3*(1), 3-21. https://doi.org/10.2190/3K6A-2REC-GJU5-8280

Bernal, E. M., Cabrera, A. F., y Terenzini, P. T. (2000). The relationship between race and socioeconomic status (SES): Implications for institutional research and admissions policies. *Removing Vestiges: Research-Based Strategies to Promote Inclusion*, (3), 6-19.

Briñez, M. E. (2021). Tecnología de información: ¿Herramienta potenciadora para gestionar el capital intelectual? *Revista de Ciencias Sociales (Ve), XXVII*(1), 180-192. https://doi.org/10.31876/rcs.v27i1.35305

Cabrera, A. F., Nora, A., y Castañeda, M. B. (1992). The role of finances in the persistence process: A structural model. *Research in Higher Education, 33*(5), 571-593. https://doi.org/10.1007/BF00973759

Cabrera, A. F., Nora, A., y Castañeda, M. B. (1993). College persistence: Structural Equations modelling test of Integrated model of student retention. *Journal of Higher Education, 64*(2), 123-320. https://doi.org/10.2307/2960026

Camborda, M. G. (2014). *Aplicación de árboles de decisión para la predicción del rendimiento académico de los estudiantes de los primeros ciclos de la carrera de Ingeniería Civil de la Universidad Continental* [Tesis de maestría, Universidad Nacional del Centro del Perú]. http://repositorio.uncp.edu.pe/handle/20.500.12894/1477

Chatterjee, P., Yazdani, M., Fernández-Navarro, F., y Pérez-Rodríguez, J. (Eds.) (2023). *Machine learning algorithms and applications in engineering*. CRC Press. https://doi.org/10.1201/9781003104858

Deng, H. (2013). Guided Random Forest in the RRF Package. *ArXiv: 1306.0237*. https://doi.org/10.48550/arXiv.1306.0237

Diario Oficial del Bicentenario El Peruano (9 de noviembre de 2021). Tasa de deserción en educación universitaria. *El Peruano*. https://elperuano.pe/noticia/132960-tasa-de-desercion-en-educacion-universitaria-se-redujo-a-115

Díaz, B., Marín, W., Lioo, F., Baldeos, L., Villanueva, D., y Ausejo, J. (2022). Deserción de estudiantes, factores asociados con árboles de decisión: Caso Escuela de Postgrado de una Universidad pública en Perú. *Risti: Revista Ibérica de Sistemas e Tecnologias de Informação,* (E-53), 197-211. https://www.risti.xyz/issues/ristie53.pdf

Díaz, C. (2008). Modelo conceptual para la deserción estudiantil universitaria chilena. *Estudios*

Pedagógicos, XXXIV(2), 65-86. https://dx.doi.org/10.4067/S0718-07052008000200004

Díaz-Landa, B., Meleán-Romero, R., y Marín-Rodriguez, W. (2021). Rendimiento académico de estudiantes en Educación Superior: Predicciones de factores influyentes a partir de árboles de decisión. *Telos: Revista de Estudios Interdisciplinarios en Ciencias Sociales, 23*(3), 616-639. https://doi.org/10.36390/telos233.08

Dole, L., y Rajurkar, J. (2014). A decision support system for predicting student performance. *International Journal of Innovative Research in Computer and Communication Engineering, 2*(12), 7232-7237. https://ijircce.com/admin/main/storage/app/pdf/GE1YcjUwFse-BETuax9LEymgN04TtdLS7TGEfm-Mgr.pdf

Duche, A. B., Paredes, F. M., Gutiérrez, O. A., y Carcausto, L. C. (2020). Transición secundaria-universidad y la adaptación a la vida universitaria. *Revista de Ciencias Sociales (Ve), XXVI*(3), 244-258. https://doi.org/10.31876/rcs.v26i3.33245

Dwi, M., Prasetya, A., y Pujianto, U. (2018). Technology acceptance model of student ability and tendency classification system. *Bulletin of Social Informatics Theory and Application, 2*(2), 47-57. https://doi.org/10.31763/businta.v2i2.113

Eccles, J., Adler, T., y Meece, J. L. (1984). Sex differences in achievement: A test of alternate theories. *Journal of Personality and Social Psychology, 46*(1), 26-43. https://doi.org/10.1037/0022-3514.46.1.26

Ethington, C. A. (1990). A psychological model of student persistence. *Research in Higher Education, 31*(3), 279-293. https://doi.org/10.1007/BF00992313

Félix, A. V., Urrea, M. L., y López, S. (2023). Abandono escolar de alumnos universitarios en la carrera de Derecho y Ciencias Sociales. *Revista de Ciencias Sociales (Ve), XXIX*(2), 242-254. https://doi.org/10.31876/rcs.v29i2.39974

Fishbein, M., y Ajzen, I. (1974). Attitudes toward objects as predictors of simple and multiple behavioural criteria. *Psycological Review, 81*, 59-74. https://doi.org/10.1037/h0035872

Fryda, T., LeDell, E., Gill, N., Aiello, S., Fu, A., Candel, A., Click, C., Kraljevic, T., Nykodym, T., Aboyoun, P., Kurka, M., Malohlava, M., Poirier, S., y Wong, W. (2022). H2O: R Interface for the 'H2O' Scalable Machine Learning Platform. R package version 3.38.0.1. https://docs.h2o.ai/h2o/latest-stable/h2o-r/docs/index.html

González, L. E. (2005). *Estudio sobre la repitencia y deserción en la educación superior chilena.* Instituto Internacional para la Educación Superior en América Latina y el Caribe, IESALC – UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000140087

Haque, M. A. (2022). *Feature Engineering & Selection for Explainable Models: A second course for data scientists*. LULU Internacional.

Jung, A. (2022). *Machine Learning: The basics*. Springer. https://doi.org/10.1007/978-981-16-8193-6

Khun, M., y Jhonson, K. (2019). *Feature Engineering and Selection: A Practical Approach for Predictive Models*. Chapman and Hall/CRC. https://doi.org/10.1201/9781315108230

Kodelja, Z. (2019). Is machine learning real learning? *CEPS Journal, 9*(3), 11-23. https://doi.org/10.26529/cepsj.709

Kuh, G. D. (2002). Organizational culture

and student persistence: Prospects and puzzles. *Journal of College Student Retention: Research, Theory & Practice, 3*(1), 23-39. https://doi.org/10.2190/U1RN-C0UU-WXRV-0E3M

Kursa, M. B., y Rudnicki, W. R. (2010). Feature selection with the Boruta Package. *Journal of Statistical Software, 36*(11), 1-13. https://doi.org/10.18637/jss.v036.i11

Larsen, K. (2016). Data Exploration with Information Theory (Weight-of-Evidence and Information Value). R package version 0.0.9. https://CRAN.R-project.org/package=Information

LeDell, E., y Poirier, S. (2020). H2O AutoML: Scalable Automatic Machine Learning. *7th ICML Workshop on Automated Machine Learning*. https://www.automl.org/wp-content/uploads/2020/07/AutoML_2020_paper_61.pdf

Ley No. 30220 de 2014. Ley Universitaria. 3 de julio de 2014.

Microsoft Learn (23 de diciembre de 2023). Conceptos de minería de datos. *Microsoft Learn*. https://learn.microsoft.com/es-es/analysis-services/data-mining/data-mining-concepts?view=asallproducts-allversions

Moreno, F. O., Ochoa, F. A., Mutter, K. J., y Vargas, E. C. (2021). Estrategias pedagógicas en entornos virtuales de aprendizaje en tiempos de pandemia por Covid-19. *Revista de Ciencias Sociales (Ve), XXVII*(4), 202-213. https://doi.org/10.31876/rcs.v27i4.37250

Mushtaq, I., y Khan, S. N. (2012). Factors affecting students' academic performance. *Global Journal of Management and Business Research, 12*(9), 17-22. https://journalofbusiness.org/index.php/GJMBR/article/view/100221

Nagarajah, T., y Poravi, G. (2019). A Review on Automated Machine Learning (AutoML) Systems. *IEEE 5th International Conference for Convergence in Technology (I2CT)*, Bombay, India. https://doi.org/10.1109/i2ct45611.2019.9033810

Nye, J. S. (1976). Independence and Interdependence. *Foreign Policy,* (22), 130-161. https://doi.org/10.2307/1148075

Organisation for Economic Co-operation and Development - OECD (2019). *Education at a Glance 2019*. OECD Publishing. https://doi.org/10.1787/f8d7880d-en

R Core Team (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

RStudio Team (2022). *RStudio: Integrated Development for R*. RStudio, http://www.rstudio.com/

Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development, 44*(1.2), 211-229. https://doi.org/10.1147/rd.441.0206

Sharmeela, C., Sanjeevikumar, P., Sivaraman, P., y Joseph, M. (2023). *IoT, machine learning and blockchain technologies for renewable energy and modern hybrid power systems*. Routledge.

Spady, W. G. (1970). Dropouts from higher education: An interdisciplinary review and synthesis. *Interchange, 1*, 64-85. https://doi.org/10.1007/BF02214313

St. John, E. P., Cabrera, A. E., Nora, A., y Asker, E. H. (2000). Economic influences on persistence reconsidered:

How can finance research inform the reconceptualization of persistence models? In J. M. Braxton (Ed.), *Reworking the student departure puzzle: New theory and research on college student retention* (pp. 29-47). Vanderbilt University Press.

Superintendencia Nacional de Educación Superior Universitaria - SUNEDU (2020). *II Informe bienal sobre la realidad universitaria en el Perú*. SINEDU. https://cdn.www.gob.pe/uploads/document/file/1230044/Informe%20Bienal.pdf

Supo, J. (2020). *Metodología de la Investigación Científica: Para las Ciencias de la Salud y las Ciencias Sociales*. Independently published.

Tinto, V. (1982). Limits of theory and practice of student attrition. *Journal of Higher Education, 53*(6), 687-700. https://doi.org/10.2307/1981525

Tinto, V. (1989). Definir la deserción: Una cuestión de perspectiva. *Revista de Educación Superior,* (71), 1-9. http://publicaciones.anuies.mx/revista/71/1/3/es/definir-la-desercion-una-cuestion-de-perspectiva

Valero, J. E., Navarro, Á. F., Larios, A. C., y Julca, J. D. (2022). Deserción universitaria: Evaluación de diferentes algoritmos de Machine Learning para su predicción. *Revista de Ciencias Sociales (Ve), XXVIII*(3), 362-375. https://doi.org/10.31876/rcs.v28i3.38480

Villarreal-Torres, H., Ángeles-Morales, J., Marín-Rodriguez, W., Andrade-Girón, D., Carreño-Cisneros, E., Cano-Mejía, J., Mejía-Murillo, C., Boscán-Carroz, M. C., Flores-Reyes, G., y Cruz-Cruz, O. (2023). Development of a classification model for predicting student payment behavior using artificial intelligence and data science techniques. *EAI Endorsed Transactions on Scalable Information Systems, 10*(5). https://doi.org/10.4108/eetsis.3489

Villarreal-Torres, H. O., Marín-Rodriguez, W. J., Ángeles-Morales, J. C., y Cano-Mejía, J. E. (2021). Gestión de Tecnología de Información para universidades peruanas aplicando computación en la nube. *Revista Venezolana de Gerencia, 26*(E-6), 665-679. https://doi.org/10.52080/rvgluz.26.e6.40

Xu, W., y Li, W. (2014). Granular computing approach to two-way learning based on formal concept analysis in Fuzzy Datasets. *IEEE Transactions on Cybernetics, 46*(2), 366-379. https://doi.org/10.1109/tcyb.2014.2361772

Zárate-Valderrama, J., Bedregal-Alpaca, N., y Cornejo-Aparicio, V. (2021). Modelos de clasificación para reconocer patrones de deserción en estudiantes universitarios. *Ingeniare. Revista Chilena de Ingeniería, 29*(1), 168-177. http://dx.doi.org/10.4067/S0718-33052021000100168

Zöller, M.-A., y Huber, M. F. (2021). Benchmark and survey of automated machine learning frameworks. *Journal of Artificial Intelligence Research, 70*, 409-472. https://doi.org/10.1613/jair.1.11854

Zwanenburg, A., y Löck, S. (2021). Familiar: End-to-End Automated Machine Learning and Model Evaluation. https://cran.r-project.org/web/packages/familiar/familiar.pdf