Revista Lince de Ciencias Sociales, Humanidades y Tecnologías
Vol. 1 (1), 2025
DOI: https://doi.org/10.63622/RLI.25.1.06
https://editoriallince.uadeo.mx/index.php/LINCE

Artículo de Revisión de Literatura

# Técnicas de *Machine Learning* para Predecir el Abandono Escolar Universitario: Revisión Sistemática de la Literatura

## *Machine Learning Techniques for Predicting University Dropout: A Systematic Literature Review*

*Carlos Ramiro Ibarra Sarmiento*[1] , *Jesús Jaime Solano Noriega*[2], *Diego Alonso Gastélum Chavira*[3] y *Luis Asunción Pérez Domínguez*[4]

[1]*Departamento de Ciencias Económico–Administrativas, Universidad Autónoma de Occidente, Los Mochis, 81216, México*

[2]*Departamento de Ingeniería y Tecnología, Universidad Autónoma de Occidente, Los Mochis, 81216, México*

[3]*Departamento de Ciencias Económico-Administrativas, Universidad Autónoma de Occidente, Culiacán, 80200, México*

[4]*Departamento de Ingeniería Industrial y de Manufacturera, Universidad Autónoma de Ciudad Juárez, Chihuahua, 32315, México*

carlos.ibarras@uadeo.mx, jaime.solano@uadeo.mx, diego.chavira@uadeo.mx luis.dominguez@uacj.mx

**INFORMACIÓN**

***Historial del Artículo***

*Recibido: Abril 29, 2025*
*Aceptado: Junio 26, 2025*
*Publicado: Septiembre 22, 2025*

***Palabras Clave***

*Revisión de literatura*
*Meta-análisis*
*Inteligencia artificial*
*Abandono escolar*
*Web of science*
*SCOPUS*

**RESUMEN**

Las técnicas de *Machine Learning* (ML) se han convertido en herramientas claves en la investigación educativa, permitiendo prevenir la deserción escolar, mejorar la retención estudiantil y optimizar la gestión académica en las Instituciones de Educación Superior (IES). Este artículo presenta una revisión sistemática de la literatura para identificar los factores más relevantes en la predicción de la deserción escolar y analizar las técnicas de ML utilizadas en estudios dentro de las IES. Se consultaron las bases de datos *Web of Science* (WOS) y SCOPUS, y se empleó la metodología PRIMA *(Preferred Reporting Items for Systematic reviews and Meta-Analyses)* para reportar los hallazgos en la literatura sobre el uso de modelos de ML, balanceo de datos, métricas de evaluación, entre otros criterios sobre las publicaciones. Después de aplicar PRISMA, se seleccionaron 23 estudios publicados entre 2018 y 2022 para un análisis detallado. Los resultados muestran un análisis cuantitativo sobre los estudios encontrados, se analiza información sobre la calidad, síntesis demográfica y año de las publicaciones, las palabras claves utilizadas, atributos seleccionados y técnicas utilizadas para su selección, el tamaño de las bases de datos, las técnicas de división y balanceo de datos, los algoritmos predictivos y las métricas de evaluación.

**Cómo citar:**
Ibarra Sarmiento, C. R., Solano Noriega, J. J., Gastélum Chavira, D. A., & Pérez Domínguez, L. A. (2025). Técnicas de machine learning para predecir el abandono escolar universitario: Revisión sistemática de la literatura. *Revista Lince de Ciencias Sociales, Humanidades y Tecnologías*, 1(1), 123–183. https://https://doi.org/10.63622/RLI.25.1.06

**MANUSCRIPT INFO**

***Article History***

*Received: April 29, 2025*
*Accepted: June 26, 2025*
*Published: September 22, 2025*

***Keywords***

*Literature review*
*Meta-Analysis*
*Artificial Intelligence*
*School dropout*
*Web Of Science*
*SCOPUS*

**ABSTRACT**

Machine learning (ML) techniques have become key tools in educational research, allowing for the prevention of school dropouts, improving student retention, and optimizing academic management in Higher Education Institutions (HEIs). This article presents a systematic literature review to identify the most relevant factors in predicting school dropouts and analyze the ML techniques used in studies within HEIs. The Web of Science (WOS) and SCOPUS databases were consulted, and the PRIMA (Preferred Reporting Items for Systematic reviews and Meta-Analyses) methodology was employed to report findings in the literature on the use of ML models, data balancing, evaluation metrics, among other criteria on publications. After applying PRISMA, 23 studies published between 2018 and 2022 were selected for detailed analysis. The results show a quantitative analysis of the studies found, analyzing information on the quality, demographics and year of publication, keywords used, selected attributes and techniques used for their selection, database size, data division and balancing techniques, predictive algorithms, and evaluation metrics.

# 1. Introducción

Se puede definir a la deserción escolar, como el abandono de una institución educativa por parte de un estudiante antes de obtener las credenciales educativas o un título profesional (Delen et al., 2020; Spady, 1970). Este fenómeno se considera uno de los temas cruciales para las instituciones educativas (Fahd et al., 2022). Porque genera problemas como la disminución de la eficiencia terminal, el aumento del rezago nacional y perjudica la certificación de los programas educativos (Hernández González et al., 2016).

Para el estudio de este fenómeno se han utilizado la estadística descriptiva o también métodos estadísticos tradicionales (Delen, 2010), como ejemplo, tenemos el estudio de Sacală et al. (2021), el cual utilizó el método de regresión logística para la identificación de los factores de riesgo, otros que han utilizado este métodos para este mismo fin, son los trabajos de Onder y Batar (2021) ; Raidal et al. (2019); Ryan et al. (2021)); para conocer la relación entre los factores se han empleado métodos como el análisis del modelo de riesgo proporcional de Cox (Utami et al., 2020), el análisis de $\chi 2$ (Nguyen et al., 2022) y la correlación de Pearson o Spearman (Dancot et al., 2021), otros autores se han enfocado en la predicción de la deserción escolar y han utilizado el modelo de ecuaciones estructurales (Nikolaidis et al., 2022).

Predecir la deserción escolar en etapa temprana y propiciar la intervención oportuna por parte del personal administrativo y académico (Albreiki et al., 2021)) con la implementación de estrategias o programas (Jin et al., 2011; Lackey et al., 2003) tiene la finalidad de ayudar a los estudiantes a mejorar su desempeño académico y así elevar la tasas de retención (Ahmad et al., 2015; Umer et al., 2021) Con esta idea en mente, la predicción anticipada de este fenómeno contribuye al crecimiento de las instituciones

educativas (Balaji et al., 2021).

El ML está emergiendo hoy en día como una herramienta importante para el apoyo a muchas áreas de investigación para la toma de decisiones (Balaji et al., 2021). Chu et al. (2022) recomienda a los investigadores la utilización de las tecnologías de Inteligencia Artificial (IA) para prevenir problemas de deserción escolar y mejorar las posibilidades de retención, con este tipo de herramientas se pueden beneficiar tanto las instituciones educativas como los estudiantes (Balaji et al., 2021); mediante la utilización de conjuntos de datos los algoritmos de ML son útiles para la predicción temprana de los estudiantes en riesgo y con posibilidades de abandono (Albreiki et al., 2021), por esta razón los algoritmos de ML se utilizan para obtener mejores predicciones de los resultados (Arizmendi et al., 2022).

El creciente uso de técnicas de ML en la educación para la predicción de la deserción escolar, junto con el aumento en la frecuencia de las citas sobre el tema, refleja un gran interés de los investigadores en su estudio (Fahd et al., 2022)

Las revisiones sistemáticas, son revisiones de investigación rigurosas, las cuales siguen métodos científicos (Schlosser et al., 2013), hacer este tipo de revisiones es seguir una metodología o proceso estructurado para evaluar, comparar y sintetizar evidencia relevante sobre preguntas de investigación específicas (Brannon et al., 2014); Por un lado, en relación con los estudios individuales, las revisiones sistemáticas ocupan un lugar más alto en las jerarquías de evidencia, debido a que se sintetizan múltiples estudios, siendo en trabajo más poderoso que cualquier estudio individual (Schlosser et al., 2013).

Las revisiones sistémicas son de gran utilidad, estás pueden proporcionar una síntesis del estado del arte en un tema determinado y ayuda a identificar las necesidades de investigación en el futuro (Page et al., 2021). El presente trabajo tiene como objetivo realizar una revisión sistemática de la literatura reciente, para conocer los factores que ocasionan la deserción escolar; así como las técnicas de ML más utilizadas para la predicción de la misma.

Para lograr el objetivo en el presente trabajo es necesario contestar las siguientes preguntas de investigación (RQ): RQ1 ¿Cuáles son los factores comúnmente utilizados para la predicción de la deserción escolar? RQ2 ¿Cuáles son las técnicas de ML más utilizadas para la predicción de la deserción escolar? Las respuestas a estas preguntas de investigación proporcionarán información valiosa a los investigadores que deseen estudiar este fenómeno y su predicción con técnicas de ML.

## 2. Revisión de la literatura

El concepto de deserción o abandono escolar puede ser definido desde distintas perspectivas: i) desde la perspectiva individual, el abandono del programa educativo puede considerarse un fracaso, pero la acción de abandonar el programa educativo para cambiarse a otra institución con mejor nivel académico, puede considerarse como un acierto para el cumplimiento de una meta personal; ii) la perspectiva desde el punto de vista institucional, clasifica como desertores a todas aquellas personas que abandonan la institución educativa, este abandono crea una vacante para que pueda ser ocupada por otro estudiante; iii) y la perspectiva desde el punto de vista estatal o nacional, la cual se genera cuando el estudiante abandona sus estudios en cualquiera de las modalidades

del sistema educativo (Tinto, 1989).

La deserción escolar es uno de los principales problemas que enfrentan las instituciones educativas, puesto que esta afecta a las políticas institucionales (Otero Escobar, 2021; Spady, 1970); la pérdida de estudiantes genera problemas financieros, debido a la inestabilidad en la fuente de ingresos, cuando esta deserción se lleva a cabo en una institución privada, la institución ya no contará con esos ingresos, cuando se lleva a cabo en una institución pública, el ingreso se disminuye, debido a la disminución del presupuesto (Tinto, 1989). Otros de los problemas que se generan, son la disminución de la eficiencia terminal, aumento del rezago nacional y esto perjudica la certificación de los programas educativos (Hernández González et al., 2016), este fenómeno es importante desde el punto de vista administrativo, difícil desde el punto de vista práctico para los responsables de la toma de decisiones y los investigadores e interesante desde el punto de vista científico (Delen et al., 2020).

Desde hace varias décadas este fenómeno es estudiado para identificar los factores que la provocan, la dificultad que afrontan las universidades para definir la deserción, consiste en identificar los tipos de abandono (Tinto, 1989) o los factores que la ocasionan. Para Tinto (1975) la deserción escolar puede ser generada por diversos factores, los cuales se clasifican de la siguiente forma: trasfondo familiar, características individuales, experiencias educativas pasadas, compromiso de meta, integración académica, la integración social y calidad universitaria.

Una revisión reciente de la literatura relacionada refleja que los factores que influyen en la deserción escolar pueden ser personales, falta de motivación, relaciones interpersonales, embarazos adolescentes, problemas socioeconómicos y pedagógicos (Rochin Berumen, 2021).

Algunos estudios de investigación con la finalidad de poder predecir la deserción escolar, han utilizado los siguientes factores: raza (Arizmendi et al., 2022), género, grupo étnico (Arizmendi et al., 2022; Guarín et al., 2015), edad de ingreso, ciudad de origen (Guarín et al., 2015), clasificación socioeconómica, datos demográficos (Arizmendi et al., 2022; da Silva & Roman, 2021; Guarín et al., 2015), motivacional, psicológico, académico y personal/social (da Silva & Roman, 2021).

La investigación de Salas-Pilco y Yang (2022) demostró que la IA se utiliza principalmente para el análisis automático de contenido, el análisis de imágenes y la creación de modelos predictivos, la IA se ha desarrollado rápidamente en los últimos años, su aplicación en la educación se ha vuelto cada vez más frecuente (Huang et al., 2021).

La aplicación de IA en la educación, en sus diversas formas, ha tenido un gran impacto en el desempeño administrativo y de gestión en la educación (Chen et al., 2020). Con el ML se procesan conjuntos de datos que contienen los registros de los estudiantes, con la finalidad de obtener las características necesarias para generar un modelo predictivo; estas características pueden ser los factores académicos, demográficos, sociales y de comportamiento de los estudiantes entre otros, estas sirven de entrada a los algoritmos de ML para la identificación de estudiantes en riesgo (Kaddoura et al., 2022). La eficacia de los algoritmos se basa en la disponibilidad de una gran cantidad de datos tanto positivos como negativos (Dalipi et al., 2018).

Para Cui et al. (2019) las técnicas más utilizadas de manera exitosas son las siguientes: Decision Tree (DT), clasificador Naïve Bayes (NB), Support Vector Machine (SVM),

Artificial Neural Networks (ANN), Random Forest (RF) y Logistic Regression (LR).

En la Tabla 1 se presenta un análisis de los trabajos de investigación enfocados a una revisión sistemática de la literatura sobre los estudios que emplean técnicas de ML aplicadas a la deserción escolar.

**Tabla 1**. Análisis de trabajos de investigación estudios que aplican técnicas de predicción de ML en la educación.

| **Artículo** | **Objetivos, aportaciones y áreas de oportunidad** |
|---|---|
| *MOOC dropout prediction using machine learning techniques: Review and research challenges.* (Dalipi et al., 2018) | El objetivo de este estudio es proporcionar una descripción general sobre la aplicación de técnicas de ML para la resolución de la deserción escolar en los cursos masivos en línea, mejor conocidos como MOOC (por su acrónimo en inglés Massive Online Open Courses).<br>**Aportaciones:** Este artículo ofrece una revisión exhaustiva de las investigaciones sobre la aplicación de ML para predecir, explicar y resolver el problema del abandono estudiantil en los MOOC. Destaca tanto los factores relacionados con los estudiantes como los factores relacionados con los MOOC que conducen a un alto número de abandonos. También identifica algunos de los desafíos críticos asociados con la predicción del abandono estudiantil y ofrece recomendaciones y propuestas para ayudar a los investigadores que emplean diversas técnicas de ML a resolverlo de forma oportuna y eficiente.<br>**Área de oportunidad:** El estudio no sigue el enfoque PRISMA, además está enfocado en el entorno MOOC, no define claramente los criterios de inclusión y exclusión, no cuantifica los factores que influyen la deserción escolar, lo aborda una pequeña parte del estudio. |

| | |
|---|---|
| *Predictive analytic models of student success in higher education.* (Cui et al., 2019) | El objetivo del estudio es la revisión del 2002 al 2018 de los componentes metodológicos relacionados con los modelos predictivos que se han desarrollado o implementado en aplicaciones de análisis de aprendizaje en la educación superior.<br>**Aportaciones:** Se identifican las fortalezas y debilidades metodológicas de las aplicaciones actuales de análisis predictivo del aprendizaje y ofrecen las recomendaciones más actualizadas sobre el desarrollo, uso y evaluación de modelos predictivos.<br>**Área de oportunidad:** El estudio se enfoca en estudiar los componentes metodológicos empleados en los modelos predictivos, la revisión se sitúa hace 4 años, no especifica si utiliza el enfoque PRISMA. |
| *Data Mining and Machine Learning Retention Models in Higher Education*. (Cardona et al., 2020) | El objetivo de este estudio se centra en predecir las tasas de retención estudiantil, identificando los factores que influyen en la mismas, así como el estudio de las técnicas de predicción con mejor desempeño.<br>**Aportaciones:** Se identificaron factores importantes para los estudiantes en riesgo y también identificaron un interés considerable en crear modelos para predecir el riesgo de abandono escolar en las primeras etapas de la carrera universitaria.<br>**Área de oportunidad:** El estudio se enfoca en estudiar la predicción de la retención de los estudiantes, no especifica claramente el periodo, no identifica los factores que influyen en la deserción escolar, ni tampoco las técnicas utilizadas para predecirla, el periodo de revisión es hasta el 31 de agosto de 2018, hace 4 años, no especifica si utiliza el enfoque PRISMA. |

| | |
|---|---|
| *Predicting academic success in higher education: literature review and best practices.* (Alyahyan y Düştegör, 2020) | El objetivo de este estudio es proporcionar un conjunto de pautas paso a paso para los educadores que deseen aplicar técnicas de minería de datos para predecir el éxito de los estudiantes.<br>**Aportaciones:** Recopila información en modo de guía para la predicción del éxito académico.<br>**Área de oportunidad:** El estudio está enfocado en proporcionar los pasos para aplicar las técnicas de minería de datos para predecir el éxito estudiantil, no especifica el periodo estudiado, no identifica los factores que influyen en la deserción escolar, ni tampoco las técnicas utilizadas para predecirla, no especifica si utiliza el enfoque PRISMA. |
| *A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques.* (Albreiki et al., 2021) | El objetivo de este estudio es conocer las técnicas empleadas para predecir el desempeño académico de los estudiantes para determinar a los estudiantes que están en riesgo de desertar, con la utilización de datos estáticos y dinámicos, también el estudio de los planes remediales implementados.<br>**Aportaciones:** Este artículo presenta una descripción general de la técnica de ML utilizada en la minería de datos educativos, centrándose en dos aspectos críticos: 1) la predicción precisa de estudiantes en riesgo y 2) la predicción precisa de la deserción escolar.<br>**Área de oportunidad:** El estudio está enfocado en la predicción del desempeño académico, el periodo de estudio es del 2009 al 2021, no especifica si utiliza el enfoque PRISMA. |

| | |
|---|---|
| *Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature.* (Fahd et al., 2022) | El objetivo de este estudio es una revisión sistemática y meta–análisis de estudios de investigación que han informado sobre la aplicación de ML en la educación superior.<br>**Aportaciones:** El estudio contribuyó a la literatura sobre tecnología educativa al proporcionar hallazgos valiosos. Siguiendo un protocolo de selección restringido, además de hacer un meta-análisis y una síntesis estadística de los estudios.<br>**Área de oportunidad:** El estudio está enfocado en el rendimiento académico, el riesgo y la deserción de los estudiantes, el periodo de estudio es del 2010 al 2020. |
| *Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review.* (Balaji et al., 2021) | El objetivo de este estudio extraer los algoritmos y las características que se han utilizado en los estudios de predicción del desempeño académico.<br>**Aportaciones:** Este estudio identificó los algoritmos más utilizados para el desempeño académico.<br>**Área de oportunidad:** El estudio está enfocado a las técnicas utilizadas para la predicción del desempeño académico, no especifica sobre las técnicas para la predicción de la deserción escolar y tampoco sobre los factores que la influyen, el periodo de investigación es extenso de 1959 a 2020, no especifica si utiliza el enfoque PRISMA. |

| | |
|---|---|
| *Current stance on predictive analytics in higher education: opportunities, challenges and future directions.* (Umer et al., 2021) | El objetivo del estudio es identificar los datos más utilizados en la predicción de desempeño académico, identifica las técnicas de ML utilizada para predecirla; así como las métricas de evaluación de las técnicas.<br>**Aportaciones:** Realizó una revisión de los trabajos en el ámbito de la minería de datos educativos y destaca dimensiones como los datos, las técnicas de clasificación y las métricas de evaluación en la predicción del rendimiento estudiantil.<br>**Área de oportunidad:** El estudio está enfocado en el desempeño académico, el periodo de estudio es del 2008 al 2018, no habla sobre los factores que influyen a la deserción escolar; así como de las técnicas de ML utilizadas para predecirla, no especifica si utiliza el enfoque PRISMA. |
| *Predicting Dropout in Higher Education: a Systematic Review.* (da Silva & Roman, 2021) | El objetivo del estudio es resaltar las investigaciones sobre el uso de técnicas de ML y minería de datos para identificar a estudiantes en riesgo de deserción.<br>**Aportaciones:** Este estudio se realizó para contribuir en el sistema de información de Brasil, identificando las técnicas de ML de estudiantes en riesgo.<br>**Área de oportunidad:** Es un artículo de congreso, no explica claramente los criterios de inclusión y exclusión, el periodo de estudio es desde 2016 al 2020, no especifica si utiliza el enfoque PRISMA. |

| | |
|---|---|
| *Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision.* (Munir et al., 2022) | El estudio propone dos contribuciones principales. Primero, el estudio sigue un proceso repetible y objetivo de exploración de la literatura. En segundo lugar, el estudio describe y explica los temas de la literatura relacionados con el uso de algoritmos basados en IA en la educación digital. Es fundamental resaltar que el ámbito de estudio se circunscribe a la educación superior.<br>**Aportaciones:** Este estudio contribuyó en identificar qué temas y conceptos que giran en torno a la IA y los modelos basados en ML, además de recomendaciones útiles para los responsables políticos y educadores en educación digital.<br>**Área de oportunidad:** El estudio analiza temáticas como: using intelligent tutors, dropout predictions, performance predictions, adaptive and predictive learning and learning styles, analytics and group-based learning, and automation; el periodo de estudio es del 2000 al 2020; solamente se identificaron 9 estudios relacionados a la deserción escolar, cursos en línea. |
| *Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms.* (Oqaidi et al., 2022) | El objetivo es brindar un diagnóstico de las técnicas de aprendizaje automático utilizadas para detectar la deserción de los estudiantes en programas de educación superior.<br>**Aportaciones:** Este articulo plantea la necesidad de un modelo global que considere las particularidades del programa de estudios y los datos disponibles en la IES en cuestión, con la finalidad de facilitar la elección del mejor Algoritmo de ML para las mejores variables disponibles, y la mejor métrica de evaluación para predecir la deserción escolar con la mayor eficacia posible.<br>**Área de oportunidad:** El periodo de estudio es del 2018 al 2020, el estudio no se apega a la guía PRISMA, no es especifica claramente los criterios de inclusión y exclusión. |

*Fuente:* Elaboración propia.

Se observa que algunos estudios no indican de forma explícita la metodología utilizada para llevar a cabo la revisión sistemática. Por ejemplo, los trabajos de Cui et al. (2019) y Munir et al. (2022) solo describen los pasos que siguieron, sin hacer referencia a una metodología formal. Por otro lado, el estudio de (Albreiki et al., 2021) empleó la guía propuesta por Okoli, mientras que Fahd et al. (2022) aplicó la metodología PRISMA, la cual está específicamente diseñada para revisiones sistemáticas e incluye procedimientos de síntesis estadística y meta-análisis de los estudios seleccionados.

Respecto a la temática abordada, varios estudios se enfocan en el desempeño escolar y el éxito académico, junto con factores relacionados con la deserción estudiantil. El estudio más completo es el de Fahd et al. (2022), ya que, además de aplicar una metodología rigurosa, realiza una síntesis estadística y un meta-análisis. Sin embargo, los estudios incluidos en su revisión están centrados en el rendimiento académico, el riesgo de deserción y la deserción misma.

El principal valor diferencial de este estudio radica en que se enfoca exclusivamente a estudios sobre la predicción de la deserción universitaria utilizando algoritmos de ML, apoyado de la metodología PRISMA, para la generación de análisis estadísticos y de meta-análisis.

Cardona et al. (2020) en su estudio de revisión de la literatura producida antes del 31 de agosto de 2018, recomienda realizar estudios de actualización en fechas posteriores a su trabajo, y como podemos ver que se han incrementado las investigaciones y el citado de los estudios sobre el tema, en el presente trabajo se realizará una revisión sistemática de la literatura reciente, para conocer los factores que ocasionan la deserción escolar; así como las técnicas de ML más utilizadas para la predicción de la misma.

# 3. Metodología

Las revisiones sistémicas son de gran utilidad, debido a que puede proporcionar una síntesis del estado del arte en un tema determinado y ayuda a identificar las necesidades de investigación en el futuro (Page et al., 2021). El presente trabajo se apegó en la metodología PRISMA, esta metodología permite a los autores documentar paso a paso para poder lograr el objetivo de la generación de una revisión sistemática (Page et al., 2021)).

## *3.1.* Estrategia de búsqueda en las bases de datos digitales

Para realizar esta revisión sistemática se seleccionaron las siguientes bases de datos digitales: Web of Science (WOS) y SCOPUS, se eligieron estas bases de datos referenciales por la amplia cobertura en literatura científica revisada por pares, la facilidad en la recuperación de la información, porque provee de filtros avanzados con operadores booleanos, además de que ofrecen herramientas de análisis de fuentes y autores, el estudio de Boukhlif et al. (2024) menciona que por sus características, estas bases de datos se convierten en una opción ideal para realizar investigaciones, de acuerdo al estudio de Azevedo et al. (2024) en su búsqueda de literatura en WOS, SCOPUS e IEEE, las mayorías de los estudios encontrados en IEEE se encontraban en las

otras bases de datos, por lo que podemos decir que genera una disminución en el sesgo al no incluir otras bases de datos.

Se consideraron publicaciones de acceso abierto ((*open Access*) y publicaciones de acceso restringido (por suscripción).

*En cada una de las bases de datos se filtró la literatura que cumplía con los términos de búsqueda; dentro de las búsquedas se utilizaron operadores booleanos OR y AND, además de la utilización de comodines como $ y * entre los términos de búsqueda; los términos de búsqueda que se utilizaron en ambas bases de datos digitales fueron los siguientes:"attrition", "dropout", "drop-out", "at-risk", "dropping out", student$, school, "Machine learning", "artificial intelligence", "deep learning", "learning system*", "neural network", "support vector machine*", "decision tree*", "artificial neural network*", "supervised learning", "unsupervised learning", "random forest", "natural language processing", "reinforcement learning", "text processing", "fuzzy logic", "Bayesian network*", "fuzzy test", "stochastic systems", "logistic regression", LR, AI, SVR, RFR, LSTM, hft y ML; para la base de datos WOS se filtraron a través de los topics con un resultado de 2,781 publicaciones y en la base de datos de SCOPUS fue a través de title, abstract, keywords, en esta base de datos se obtuvo un resultado de 3,691 publicaciones, a continuación se muestran las consultas utilizadas para cada base de datos digital.*

**Consulta en WOS**

TS=((attrition or dropout$ or drop-out$ or at-risk or "dropping out") AND (student$ or school)) AND TS=((("artificial intelligence" OR "machine learning" OR "machine$learning" OR "learning system*" OR "*neural network*" OR "deep learning" OR "support vector machine*" OR "decision tree*" OR "artificial neural network*" OR "supervised learning" OR "unsupervised learning" OR "random forest" OR "natural language processing" OR "reinforcement learning" OR "text processing" OR "fuzzy logic" OR "Bayesian network*" OR "fuzzy test" OR "stochastic systems" OR "logistic regression" OR "LR" OR "AI" OR "SVR" OR "RFR" OR "LSTM" OR "hft")))

**Consulta en SCOPUS**

TITLE-ABS-KEY=((attrition or dropout$ or drop-out$ or at-risk or "dropping out") AND (student$ or school)) AND TITLE-ABS-KEY=((("artificial intelligence" OR "machine learning" OR "machine$learning" OR "learning system*" OR "*neural network*" OR "deep learning" OR "support vector machine*" OR "decision tree*" OR "artificial neural network*" OR "supervised learning" OR "unsupervised learning" OR "random forest" OR "natural language processing" OR "reinforcement learning" OR "text processing" OR "fuzzy logic" OR "Bayesian network*" OR "fuzzy test" OR "stochastic systems" OR "logistic regression" OR "LR" OR "AI" OR "SVR" OR "RFR" OR "LSTM" OR "hft")))

## *3.2.* Criterios de selección de literatura (inclusión y exclusión)

Los resultados de los términos de búsqueda en la base de datos, se les aplicaron criterios de inclusión y exclusión, ver Tabla 2; en la base de datos solamente se seleccionó la base de datos digital WOS core colletions con 2,603 publicaciones; para ambas bases de datos se filtró con el siguiente período de búsqueda, del 1ero de enero de 2018 al 21 de noviembre de 2022. Es importante señalar que el estudio se limitó a este periodo de búsqueda porque abarca años clave en la evolución del uso de ML y coincide con el impacto de la pandemia

por COVID-19, un evento que transformó significativamente la educación superior y pudo haber influido en los patrones de deserción escolar.

Como resultado de la búsqueda se obtuvieron 1,435 publicaciones en WOS y 2,002 publicaciones en SCOPUS, posteriormente las bases de datos se filtraron por tipo de publicación, incluyendo solamente los artículos y excluyendo el resto de los tipos de publicaciones (ejemplo: review, conferencias, libro, capítulos de libros, entre otros), este criterio dio los siguientes resultados: WOS con 1,186 publicaciones y SCOPUS con 1,306 publicaciones; el siguiente criterio de inclusión fue por el área de investigación, para WOS se filtró por matemáticas, ciencias computacionales, ingeniería e investigación en educación con resultado de 780 publicaciones y en SCOPUS el filtro que se aplicó fue matemáticas, ciencias computacionales e ingeniería con un resultado de 358 publicaciones y para ambas bases de datos digitales se incluyeron solamente publicaciones en idioma inglés con el resultado de 750 publicaciones para WOS y 344 publicaciones para SCOPUS.

**Tabla 2**. Criterios de inclusión y exclusión.

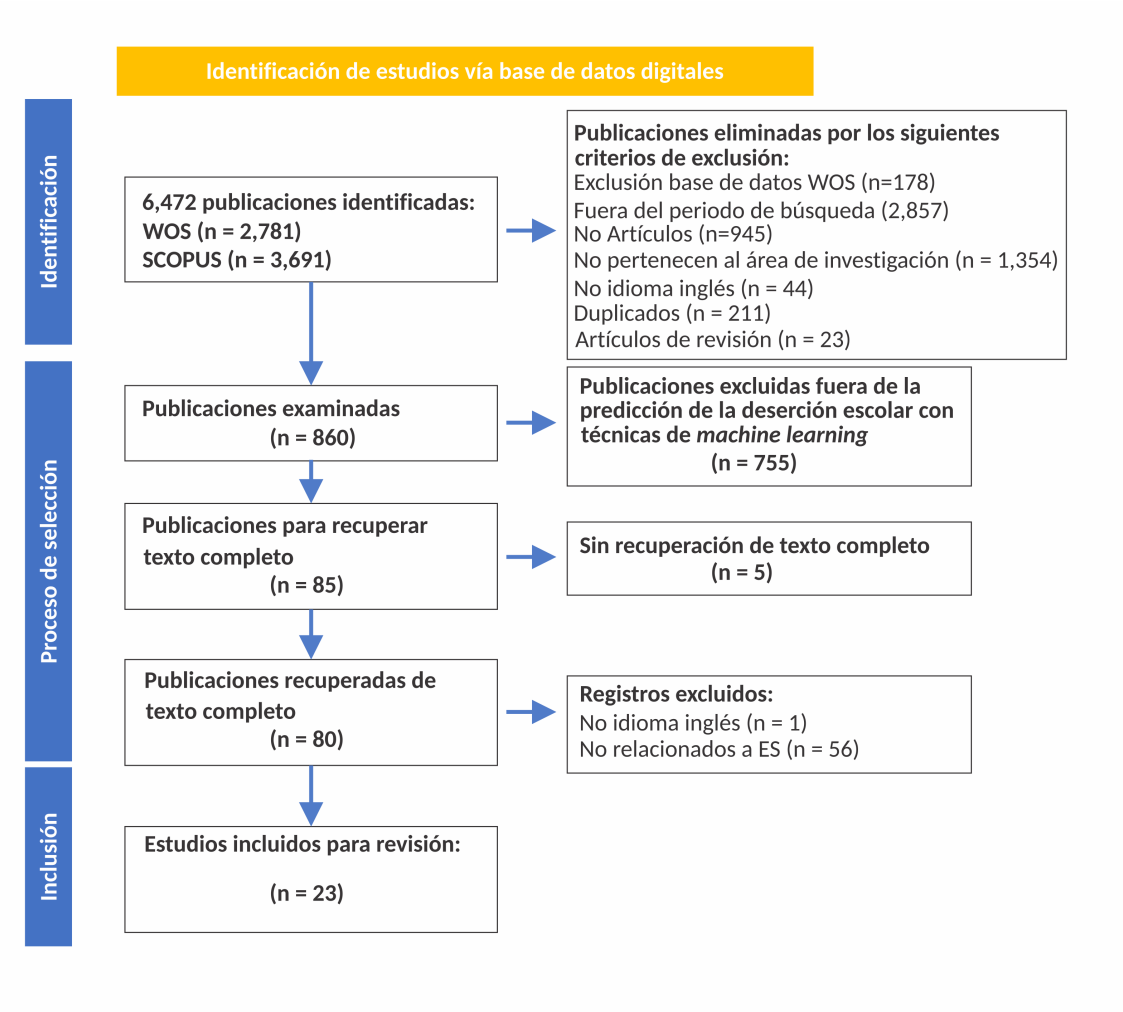
| **Inclusión** | **Exclusión** |
|---|---|
| WOS, solamente la base de datos *WOS core colletions* SCOPUS todas las bases de datos | WOS resto de bases de datos |
| Período del 1ero de enero de 2018 al 21 de noviembre de 2022 | Publicaciones fuera de período del 1ero de enero de 2018 al 21 de noviembre de 2022 |
| Publicaciones de las siguientes áreas de investigación: WOS, Matemáticas, ciencias computacionales, ingeniería e investigación en educación. SCOPUS, matemáticas, ciencias computacionales e ingeniería | Publicaciones ubicadas en otras áreas de investigación. |
| Publicaciones de acceso abierto (open Access) y publicaciones de acceso restringido (por suscripción) | |
| Publicaciones en idioma inglés | Publicaciones diferentes al idioma inglés |

*Fuente:* Elaboración propia.

Los registros obtenidos al aplicar los criterios de inclusión y exclusión en las bases de datos digitales se exportaron en formato RIS (por sus siglas en inglés *Research Information Systems*), posteriormente los archivos se importaron al software para administración de

referencias EndNote versión 20, se unieron las dos bases de datos, obteniendo 1,094 publicaciones; mediante este *software* EndNote se identificaron 211 publicaciones duplicadas, las cuales se eliminaron, por lo que quedaron 883 publicaciones, como siguiente paso se identificaron 23 estudios de revisión, los cuales fueron eliminados, el siguiente paso fue la lectura del título y del resumen de 860 publicaciones para elegir los trabajos que den respuesta a las preguntas de investigación, como siguiente paso se recuperaron los textos completos de las 85 publicaciones seleccionadas, 5 de ellas no se tuvo acceso al texto completo; las 80 publicaciones restantes se realizó lectura del trabajo completo, de las cuales se eligieron 23 publicaciones que utilizaron técnicas de ML para la predicción de la deserción escolar en la educación superior (Figura 1).

Figura 1: Proceso de selección utilizando el protocolo PRISMA.




*Fuente:* Elaboración propia.

## *3.3.* Extracción de información y análisis de publicaciones

La Tabla 3 presenta la información extraída de los estudios seleccionados para el análisis y la síntesis estadística, así como para el meta-análisis relacionados con la predicción de la deserción escolar en el nivel universitario.

**Tabla 3**. Información obtenida de estudios seleccionados.

| Característica | Descripción |
|---|---|
| Base de datos | Base de datos donde se encontró la publicación (WOS o SCOPUS). |
| Año de publicación | Año de la publicación (2018 – 2022). |
| Título de la publicación | Título único del estudio. |
| Autores | Lista de autores que participan en la publicación. |
| País de origen autor | País de origen de la institución del primer autor del estudio. |
| Citas | Cantidad de citas de la publicación. |
| Revista | Nombre de la revista donde se publicó el estudio. |
| Ranking | Ranking de reconocimiento de la revista cuando fue publicado el estudio (Q1 – Q4). |
| País | País de la revista. |
| H-Index | H-Index de la revista. |
| Objetivo de la publicación | Objetivo principal de la publicación. |
| Metodología del estudio | Metodología utilizada en el estudio (experimento/encuestas/casos de estudios). |
| *Keywords* | Palabras claves utilizadas dentro del estudio. |
| Tipo de ML | Tipo de modelo de ML utilizada en la publicación (Supervisada/No supervisada). |
| Tamaño de la base de datos | Cantidad de registros contenidos en las bases de datos utilizadas en el estudio. |
| División de los datos | Técnica de ML utilizada para la división de datos para entrenamiento y prueba de modelos. |
| Modelos | Modelos de ML utilizadas para predecir la deserción dentro de la publicación. |
| Atributos | Tipos de atributos utilizados dentro del estudio para la predicción de la deserción escolar. |
| Clasificación de los atributos | Clasificación de los atributos mencionada por los autores del estudio. |

| | |
|---|---|
| Selección de atributos | Técnicas de ML utilizadas para la selección de atributos. |
| Balanceo | Técnicas de ML para balancear los datos. |
| Métricas | Métricas de ML utilizadas para evaluar el desempeño de los modelos. |
| Herramientas | Herramientas de *software* utilizadas dentro del estudio para aplicar técnicas de ML. |

*Fuente:* Elaboración propia.

### *3.4.* Limitaciones del estudio

El presente estudio de revisión sistemática se llevó a cabo de manera minuciosa, la estrategia de búsqueda se limitó para cumplir el objetivo del estudio; la búsqueda se realizó solamente en las bases de datos digitales de WOS y SCOPUS, las estrategias de búsquedas se limitaron a los criterios de exclusión e inclusión, dentro de cada base de datos se limitó a 5 años del año 2018 al 2022 en ciertas áreas del conocimiento. Se seleccionaron estudios solamente en idioma inglés, excluyendo aquellos que están escritos en otros idiomas, la búsqueda se limitó a artículos, excluyendo capítulos de libro, entre otros.

## 4. Resultados y discusiones

### *4.1.* Calidad de las revistas y distribución de las publicaciones

La selección de trabajos de estudio fue de 23 publicaciones, solamente 4 de los trabajos se publicaron en revistas con clasificación Q1, representando el 17 % de los estudios seleccionados, el 78 % de los trabajos se publicaron en revistas con clasificación de Q1 al Q3, sin embargo 5 de los trabajos se publicaron en revistas que no están clasificadas (SR) dentro de *Scimago rank* (ver Figura 2). Las revistas utilizadas para las publicaciones de los trabajos son muy diversas, pero dos revistas fueron las que tuvieron la mayor cantidad de publicaciones, en la revista *International Journal of Advanced Computer Science and Applications* y la revista *Procedia Computer Science*, con dos publicaciones cada una de ellas, en la Tabla 4 se enumeran la cantidad de trabajos publicados en cada revista.

Se realizó un análisis de cuanto se han mencionado los estudios seleccionados dentro de otros trabajos; debido a que se utilizaron dos bases de datos, fue necesario realizar la unión de las citas registradas en ambas y se observó que la publicación de Tsai et al. (2020) es el trabajo más citado con 40, seguido del trabajo de Delen et al. (2020) con 35 citas, ambas publicaciones se publicaron en revistas con una clasificación de Q1, en tercer lugar tenemos el trabajo de Silva et al. (2020) con 20 citas publicado en una revista con una clasificación de Q3; en la Tabla 5 se enlistan las citas para cada publicación.

**Tabla 4**. Información obtenida de estudios seleccionados.

| Ranking | Revista | País | Continente | Calidad de estudios |
|---|---|---|---|---|
| Q1 | *Education and Information Technologies* | Estados Unidos | América | 1 |
| Q1 | *European Journal of Operational Research* | Países Bajos | Europa | 1 |
| Q1 | *International Journal of Educational Technology in Higher Education* | Países Bajos | Europa | 1 |
| Q1 | *Journal of College Student Retention-Research Theory & Practice* | Estados Unidos | América | 1 |
| Q2 | *Applied Sciences* | Suiza | Europa | 1 |
| Q2 | *Electronics* | Suiza | Europa | 1 |
| Q2 | *Future Internet* | Suiza | Europa | 1 |
| Q2 | *Journal of Supercomputing* | Países Bajos | Europa | 1 |
| Q2 | *Journal of Theoretical and Applied Information Technology* | Pakistán | Asia | 1 |
| Q2 | *Mathematics* | Suiza | Europa | 1 |
| Q3 | *Advances in Science, Technology and Engineering Systems* | Estados Unidos | América | 1 |
| Q3 | *International Journal of Emerging Technologies in Learning* | Austria | Europa | 1 |
| Q3 | *International Journal of Machine Learning and Computing* | Singapur | Asia | 1 |
| Q3 | *Iraqi Journal of Science* | Iraq | Asia | 1 |

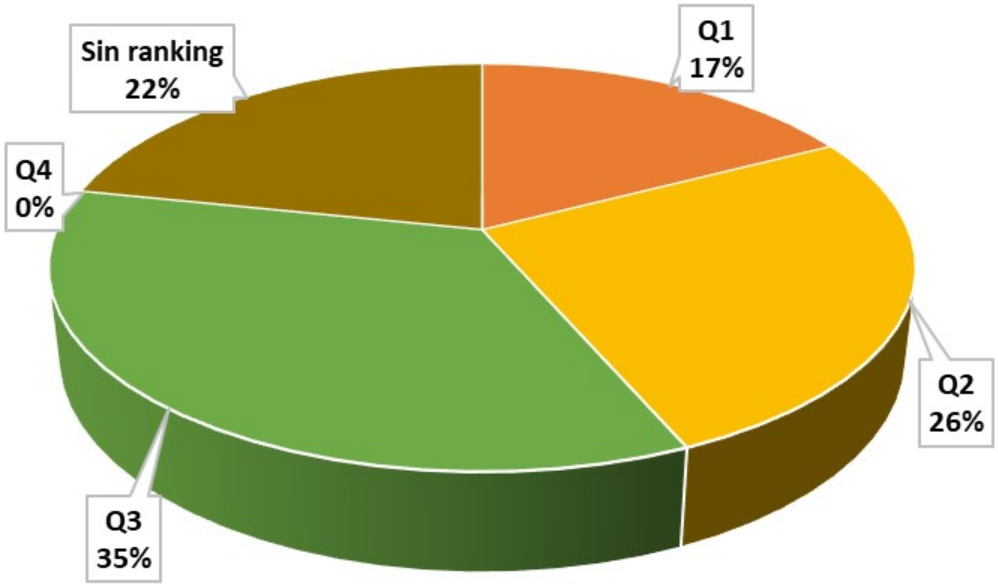
*Continúa en la siguiente página*

| **Ranking** | **Revista** | **País** | **Continente** | **Calidad de estudios** |
|---|---|---|---|---|
| Q3 | *Journal of Applied Research in Higher Education* | Reino Unido | Europa | 1 |
| Q3 | *Journal of E-Learning and Knowledge Society* | Italia | Europa | 1 |
| SR | *Computers and Education: Artificial Intelligence* | Reino Unido | Europa | 1 |
| SR | *International Journal on Information Technologies and Security* | Bulgaria | Europa | 1 |
| SR | *Procedia Computer Science* | Países Bajos | Europa | 2 |
| SR | *Relieve-Revista Electrónica de Investigación y Evaluación Educativa* | España | Europa | 1 |

*Fuente:* Elaboración propia.

De las 23 publicaciones, 16 (70 %) de ellas las podemos encontrar en ambas bases de datos digitales (WOS y SCOPUS), seis (26 %) publicaciones solamente en SCOPUS y una (4 %) solamente en WOS, como se muestra en la Figura 3.

Figura 2: Distribución de las revistas de acuerdo a su clasificación.




*Fuente:* Elaboración propia.

**Tabla 5**. Citas de estudios seleccionados.

| Año | Título del estudio | Autores | Revista | Base de datos | Citas |
|---|---|---|---|---|---|
| 2020 | *Precision education with statistical learning and deep learning: a case study in Taiwan* | Tsai, S. C.; Chen, C. H.; Shiao, Y. T.; Ciou, J. S.; Wu, T. N. | *International Journal of Educational Technology in Higher Education* | WOS/SCOPUS | 40 |
| 2020 | *Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition* | Delen, D.; Topuz, K.; Eryarsoy, E. | *European Journal of Operational Research* | WOS/SCOPUS | 35 |

*Continúa en la siguiente página*

| Año | Título del estudio | Autores | Revista | Base de datos | Citas |
|---|---|---|---|---|---|
| 2020 | *Drop-Out Prediction in Higher Education Among B40 Students* | Sani, N. S.; Nafuri, A. F. M.; Othman, Z. A.; Nazri, M. Z. A.; Nadiyah Mohamad, K. | *International Journal of Advanced Computer Science and Applications* | WOS/SCOPUS | 20 |
| 2021 | *Towards predicting student's dropout in university courses using different machine learning techniques* | Kabathova, J.; Drlik, M. | *Applied Sciences (MDPI)* | WOS/SCOPUS | 19 |
| 2019 | *Dropout situation of business computer students, University of Phayao* | Nuankaew, P. | *International Journal of Emerging Technologies in Learning* | WOS/SCOPUS | 18 |
| 2019 | *Predictive modelling of student dropout using ensemble classifier method in higher education* | Hutagaol, N.; Suharjito, | *Advances in Science, Technology and Engineering Systems* | SCOPUS | 18 |
| 2019 | *Integration of Data Technology for Analyzing University Dropout* | Viloria, Amelec; Garcia Padilla, Jholman; Vargas-Mercado, Carlos; Hernandez-Palma, Hugo; Orellano Llinas, Nataly; Arrozola David, Monica | *Procedia Computer Science* | WOS/SCOPUS | 15 |

*Continúa en la siguiente página*

| Año | Título del estudio | Autores | Revista | Base de datos | Citas |
|---|---|---|---|---|---|
| 2019 | *Neural networks to predict dropout at the universities* | Alban, M.; Mauricio, D. | *International Journal of Machine Learning and Computing* | SCOPUS | 13 |
| 2020 | *IoT system for school dropout prediction using machine learning techniques based on socioeconomic data* | Freitas, F. A. D. S.; Vasconcelos, F. F. X.; Peixoto, S. A.; Hassan, M. M.; Ali Akber Dewan, M.; de Albuquerque, V. H. C.; Rebouças Filho, P. P. | Electronics | WOS/SCOPUS | 13 |
| 2020 | *Classification models for determining types of academic risk and predicting dropout in university students* | Bedregal-Alpaca, N.; Cornejo-Aparicio, V.; Zarate-Valderrama, J.; Yanque-Churo, P. | *International Journal of Advanced Computer Science and Applications* | WOS/SCOPUS | 12 |
| 2020 | *Deep learning approach for predicting university dropout: A case study at roma tre university* | Agrusti, F.; Mezzini, M.; Bonavolontà, G. | *Journal of E-Learning and Knowledge Society* | WOS/SCOPUS | 12 |

*Continúa en la siguiente página*

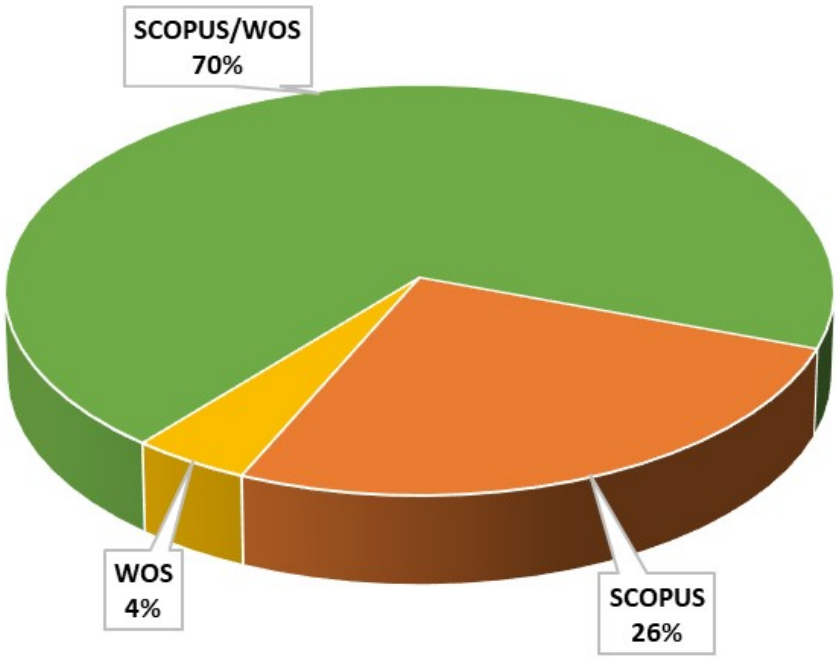
| Año | Título del estudio | Autores | Revista | Base de datos | Citas |
|---|---|---|---|---|---|
| 2020 | *University dropout: Prevention patterns through the application of educational data mining* | Urbina-Najera, A. B.; Camino-Hampshire, J. C.; Barbosa, Cruz R. | *Relieve-Revista Electrónica de Investigación y Evaluación Educativa* | WOS/SCOPUS | 10 |
| 2022 | *Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization* | Niyogisubizo, J.; Liao, L.; Nziyumva, E.; Murwanashyaka, E.; Nshimyumukiza, P. C. | *Computers and Education: Artificial Intelligence* | SCOPUS | 10 |
| 2020 | *A parallel intelligent algorithm applied to predict students dropping out of university* | Lee, Z. J.; Lee, C. Y. | *Journal of Supercomputing (Springer)* | WOS/SCOPUS | 6 |
| 2021 | *Analysis of first-year university student dropout through machine learning models: A comparison between universities* | Opazo, D.; Moreno, S.; Álvarez-Miranda, E.; Pereira, J. | Mathematics | WOS/SCOPUS | 5 |
| 2019 | *Bayesian Classifier Applied to Higher Education Dropout* | Viloria, Amelec; Pineda Lezama, Omar Bonerge; Varela, Noel | *Procedia Computer Science* | WOS/SCOPUS | 4 |
| 2022 | *Prediction of student attrition risk using machine learning* | Barramuno, Mauricio; Meza-Narvaez, Claudia; Galvez-Garcia, German | *Journal of Applied Research in Higher Education* | WOS/SCOPUS | 3 |

*Continúa en la siguiente página*

| Año | Título del estudio | Autores | Revista | Base de datos | Citas |
|---|---|---|---|---|---|
| 2022 | *Forecasting Students Dropout: A UTAD University Study* | Moreira da Silva, D. E.; Solteiro Pires, E. J.; Reis, A.; de Moura Oliveira, P. B.; Barroso, J. | *Future Internet* | SCOPUS | 2 |
| 2022 | *A stacking based hybrid technique to predict student dropout at universities* | Daza, A | *Journal of Theoretical and Applied Information Technology* | SCOPUS | 1 |
| 2022 | *Predicting Freshmen Attrition in Computing Science using Data Mining* | Naseem, Mohammed; Chaudhary, Kaylash; Sharma, Bibhya | *Education and Information Technologies* | WOS/SCOPUS | 1 |
| 2020 | *A Method for Estimating Students' Desertion in Educational Institutions Using the Analytic Hierarchy Process* | Silva, Hernan A.; Quezada, Luis E.; Oddershede, A. M.; Palominos, Pedro I.; O'Brien, Christopher | *Journal of College Student Retention-Research Theory & Practice* | WOS/SCOPUS | 0 |
| 2020 | *Using machine learning to analyze university students' dropout rate - a case study* | Nedeva, Veselina; Pehlivanova, Tanya | *International Journal on Information Technologies and Security* | WOS | 0 |
| 2021 | *The use of predictive analyzes for university dropout cases* | Alaoui, H. H.; Hachem, E.; Ziti, C.; Bassiri, M. | *Iraqi Journal of Science* | SCOPUS | 0 |

*Fuente:* Elaboración propia.

Figura 3: Publicaciones encontradas en las bases de datos digitales.

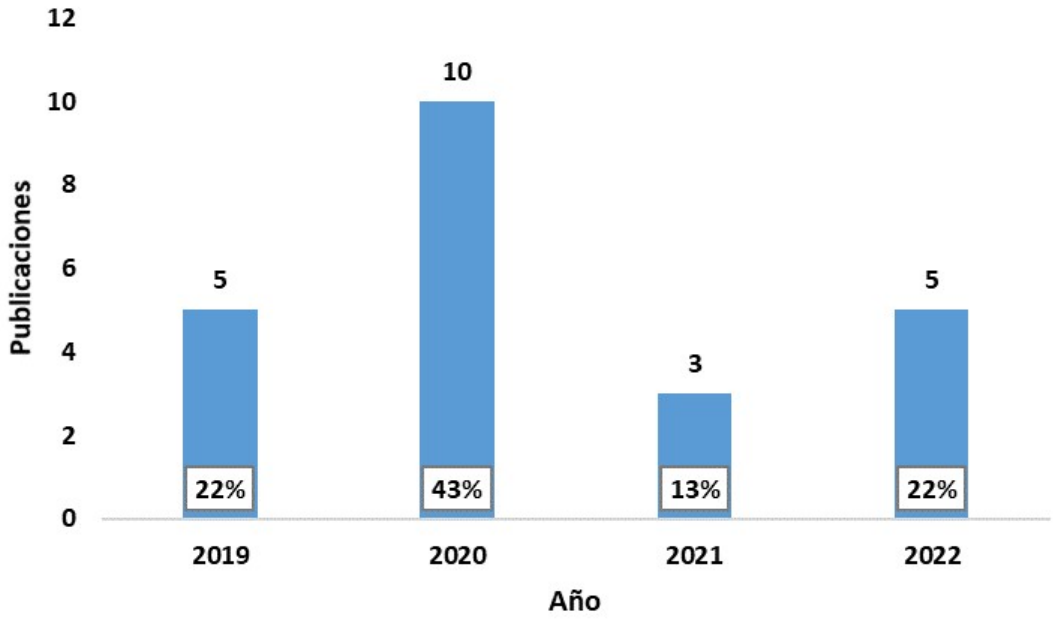



*Fuente:* Elaboración propia.

## *4.2.* Síntesis demográfica de los estudios seleccionados.

Se realizó un análisis de los estudios seleccionados que utilizan modelos de ML para la predicción de la deserción escolar en las IES, la búsqueda estuvo limitada a enero de 2018 a noviembre de 2022, se puede observar que no hubo ningún estudio publicado en el año 2018, en el año 2019 se publicaron cinco, representando un 22 % de los estudios seleccionados, pero en el 2020 incrementó a diez publicaciones, representando un 43 %, en el año 2021 y 2022 hubo un decremento de publicaciones, con tres y cinco respectivamente, presentando un 13 y 22 % (Figura 4), al ver este decremento nos podríamos hacer la siguiente pregunta ¿La pandemia ocasionada por el COVID-19 podría haber causado este decremento en la publicaciones en el año 2021 y 2022?

Figura 4: Distribución estudios publicados por año.

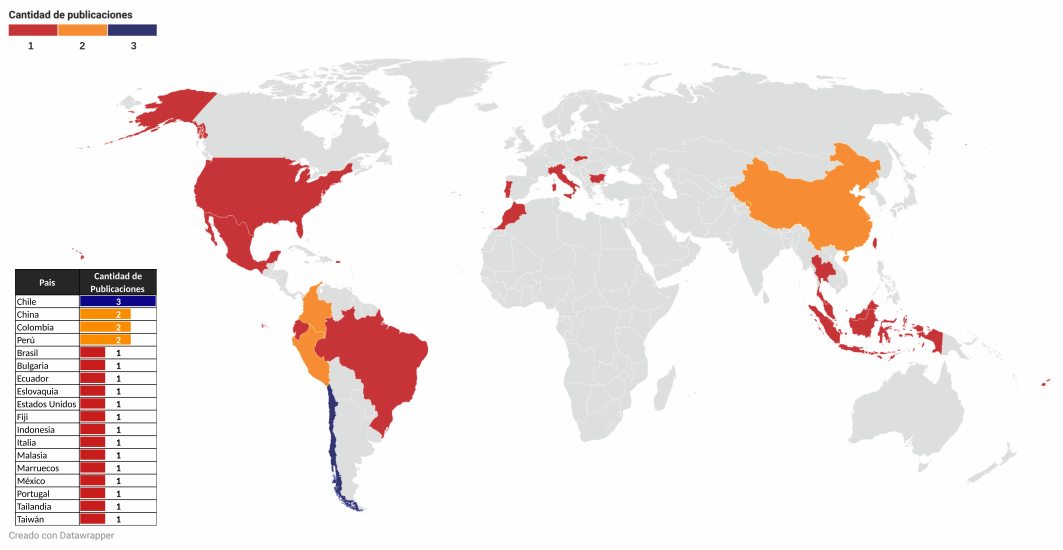



*Fuente:* Elaboración propia.

*Nota.* Aunque el año 2018 fue incluido en los criterios de selección, tras el cribado no se identificaron estudios correspondientes a dicho año.

Haciendo un análisis por el país de origen de la revista de publicación, se observó que cinco publicaciones fueron en los países bajos, siendo el país con mayor publicación en distribuidas en cuatro revistas diferentes, Reino Unido y Suiza son los siguientes países con mayor cantidad publicaciones con cuatro cada uno de ellos; también podemos observar que 74 % de los países pertenecen al continente europeo, la revistas de Asia y América representan el 26 % de las publicaciones con 13 % para cada continente, como se muestra en la Tabla 4. Pero si el análisis lo realizamos de acuerdo a país de origen de la institución del primer autor, se observa que Chile liderea en las publicaciones con tres estudios, con dos publicaciones observamos a China, Colombia y Perú con dos estudios cada uno de los países, seguido de Brasil, Bulgaria, Ecuador, Eslovaquia, Estados Unidos de América, Fiji, Indonesia, Italia, Malasia, Marruecos, México, Portugal, Tailandia y Taiwán con una publicación cada país; el 47.8 % de estas instituciones están ubicados en el continente americano, en Asia el 26.1 %, en Europa el 17.4 % y con 4.3 % para África y Oceanía, como se muestra en la Figura 5.

Figura 5: Distribución de publicaciones por país de origen de la institución del primer autor.

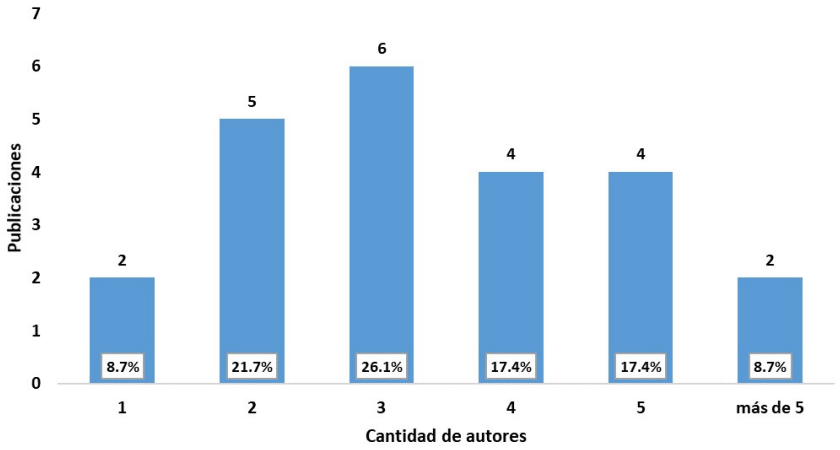



*Fuente:* Elaboración propia.

*Nota.* La figura fue generada con el software Datawrapper (2025).

Al analizar los estudios seleccionados de acuerdo la cantidad de autores dentro del estudio, se observa que seis (26.1 %) estudios tienen tres autores, cinco de los estudios contienen a dos autores, cuatro (17.4 %) estudios contenía cuatro autores y de igual forma cuatro (17.4 %) publicaciones con cinco autores, y con un (8.7 %) solo autor fueron dos estudios, y de igual manera dos estudios con más de cinco autores (ver Figura 6).

Figura 6: Distribución de publicaciones por cantidad de autores.

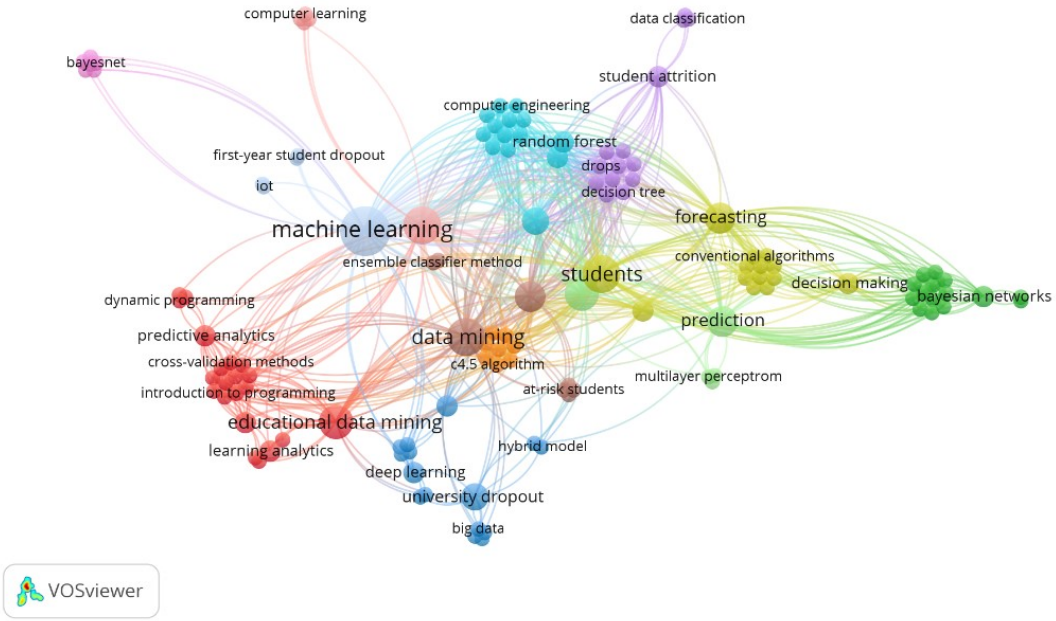



*Fuente:* Elaboración propia.

## 4.3. Análisis de palabras claves de los estudios seleccionados.

De los 23 estudios seleccionados, se realizó un análisis de las palabras claves utilizadas dentro de cada uno de los estudios, ML fue la palabra clave más utilizada y fue en ocho estudios, seguido de Educational Data Mining en cinco de los estudios y Prediction en cuatro de los estudios seleccionados; se observó que la mayor cantidad de palabras claves fue dentro de los estudios que se publicaron en el año 2020, año en donde se publicó casi la mitad (43 %) de los estudios. En la Figura 7 se puede observar los patrones, la distancia y la conexión existente de todas las palabras claves de los estudios seleccionados. Dentro de las palabras reservadas se puede observar algunos algoritmos de ML como: Convolutional Neural Network, Decision tree, ID3 algorithm, k-nearest neighbors, logistic regression, multilayer perceptrom, NaiveBayes, OneR, Random Forest y XGBoost; así como técnicas de selección de atributos como Elastic Net y también el software empleado en los estudios como WEKA.

Figura 7: Mapa generado con palabras claves de estudios seleccionados.




*Fuente:* Elaboración propia.

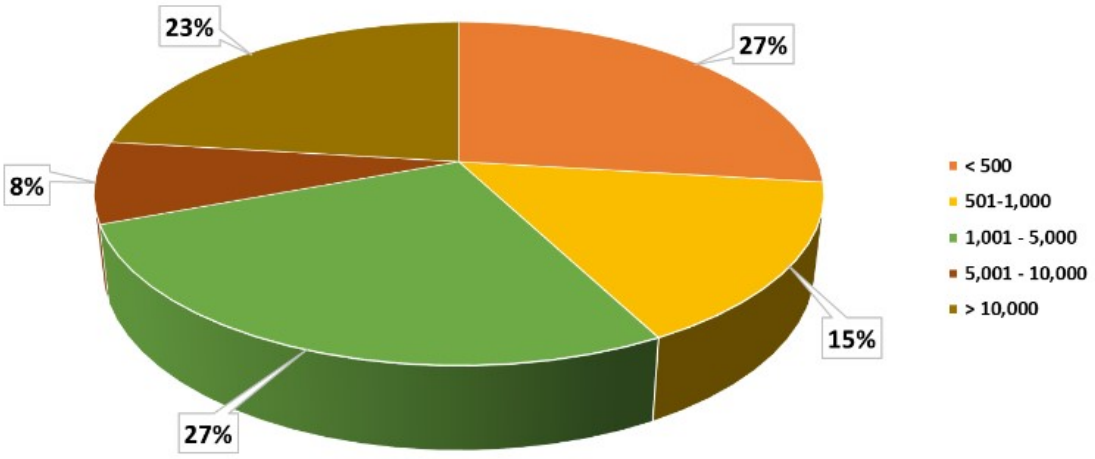
*Nota.* La Figura fue generada con el software VOSviewer versión 1.6.19, (Eck & Waltman, 2023).

## 4.4. Análisis de bases de datos y procesamiento de datos.

Se analizaron las bases de datos utilizadas dentro los estudios seleccionados, todos ellos utilizaron información de escenarios del mundo real, el estudio de Naseem et al. (2022) utilizó tres bases de datos diferentes dentro de su estudio, el trabajo de Opazo et al.

(2021) utilizó dos bases de datos, en el trabajo de Lee y Lee (2020), adicionalmente se utilizó una base de datos del repositorio UCI (por su siglas en inglés, Universidad de California em Irvine) para verificar la corrección del algoritmo propuesto. Adicionalmente, estás se analizaron de acuerdo a la cantidad inicial de registros utilizados en cada uno de los estudios, los 23 estudios utilizaron 26 bases de datos, en la Figura 8 se puede observar que se utilizaron siete (27 %) base de datos con menos de 500 registros, de igual manera los estudios utilizaron siete bases de datos con 1,001 a 5,000 registros y también se utilizaron 6 (23 %) bases de datos con más de 10,000, por nombrar los rangos con mayores frecuencias.

Figura 8: Distribución de estudios de acuerdo al tamaño inicial de la base de datos.




*Fuente:* Elaboración propia.

Al analizar los atributos utilizados dentro de los estudios, se observó que 22 estudios utilizaron entre 5 y 33 atributos, un estudio utilizó 56 y otro estudio 89 variables, uno de los trabajos no especificó los atributos utilizados. En relación con la clasificación de los atributos, en 16 estudios los clasificaron como académicos, en seis estudios como demográficos y seis estudios como financieros, por mencionar las tres características más utilizadas; analizando los atributos, género fue utilizado en 15 estudios, siendo el más utilizado, edad en segundo lugar con 11 estudios y los ingresos familiares en siete de los estudios, en la Tabla 6 se puede observar los cinco atributos más utilizados.

Para la selección de atributos solo algunos de los estudios mencionan cuales fueron los métodos utilizados, a continuación se en listan los métodos mencionados: *Ranking method* (Alban & Mauricio, 2019), *Elastic net* (Delen et al., 2020), *Correlation matrix* (Kabathova & Drlik, 2021),*The Permutation Feature Importance (PFI) technique* (Moreira da Silva et al., 2022), *Boruta* (Naseem et al., 2022), *InfoGainAttributeEval technique* (Nedeva & Pehlivanova, 2020), *Correlation heat map* (Niyogisubizo et al., 2022), *Forward selection approach* (Opazo et al., 2021), el estudio de Urbina-Najera et al. (2020) utilizó dos métodos, *GainRatioAttributeEval* y *InfoGainAttributeEval*, así como el estudio de Hutagaol y Suharjito (2019), *Ensemble Bagging Tree* y *Learning Vector Quantization*, solo

**Tabla 6**. Atributos utilizados dentro de los estudios.

| Atributo | Estudios |
|---|---|
| Género | 15 |
| Edad | 11 |
| Ingresos familiares | 7 |
| Estado civil | Tab6 |
| Estado laboral | 5 |

*Fuente:* Elaboración propia.

el estudio de *Chi-Square test* para la selección, pero adicionalmente utilizó para cada atributo seleccionado, los métodos *Phi* y *Cramer's V* para medir la fuerza de asociación.

Para la división de la base de datos para el entrenamiento y prueba, el 39 % de los estudios seleccionados no mencionan cuál es la técnica utilizada, el 52 % utilizaron *k -fold cross-validation* y el 9 % *stratified k -fold cross-validation*; en la Tabla 7 podemos ver más detalladamente la utilización de los métodos, podemos destacar que el más utilizado es el *10 -fold cross-validation*, el estudio de Nuankaew (2019) utilizó los siguientes cuatro métodos dentro del estudio: *15 -fold cross-validation, 10 -fold cross-validation, 5 -fold cross-validation* y *Leave-one-out cross-validation*.

**Tabla 7**. Métodos empleados para la división de la base de datos (entrenamiento y prueba).

| Método | Estudios |
|---|---|
| *15 -fold cross-validation* | 1 |
| *10 -fold cross-validation* | 11 |
| *5 -fold cross-validation* | 2 |
| *Leave-one-out cross-validation* | 1 |
| *Stratified 10 -fold cross-validation* | 2 |
| *No especifica* | 9 |

*Fuente:* Elaboración propia.

En el análisis de los métodos utilizados para el manejo del desbalanceo de las bases de datos, solamente siete de los estudios seleccionados menciona cual utilizaron, se puede observar en la Tabla 8 que el método más utilizado fue *Synthetic Minority Oversampling Technique (SMOTE)*, mencionado en cuatro estudios, el trabajo de Lee y Lee (2020) utilizó 3 métodos, SMOTE, *Tomek link* y *Clustering Based Oversampling* (CBO).

**Tabla 8**. Métodos utilizados para el manejo del desbalanceo en las bases de datos.

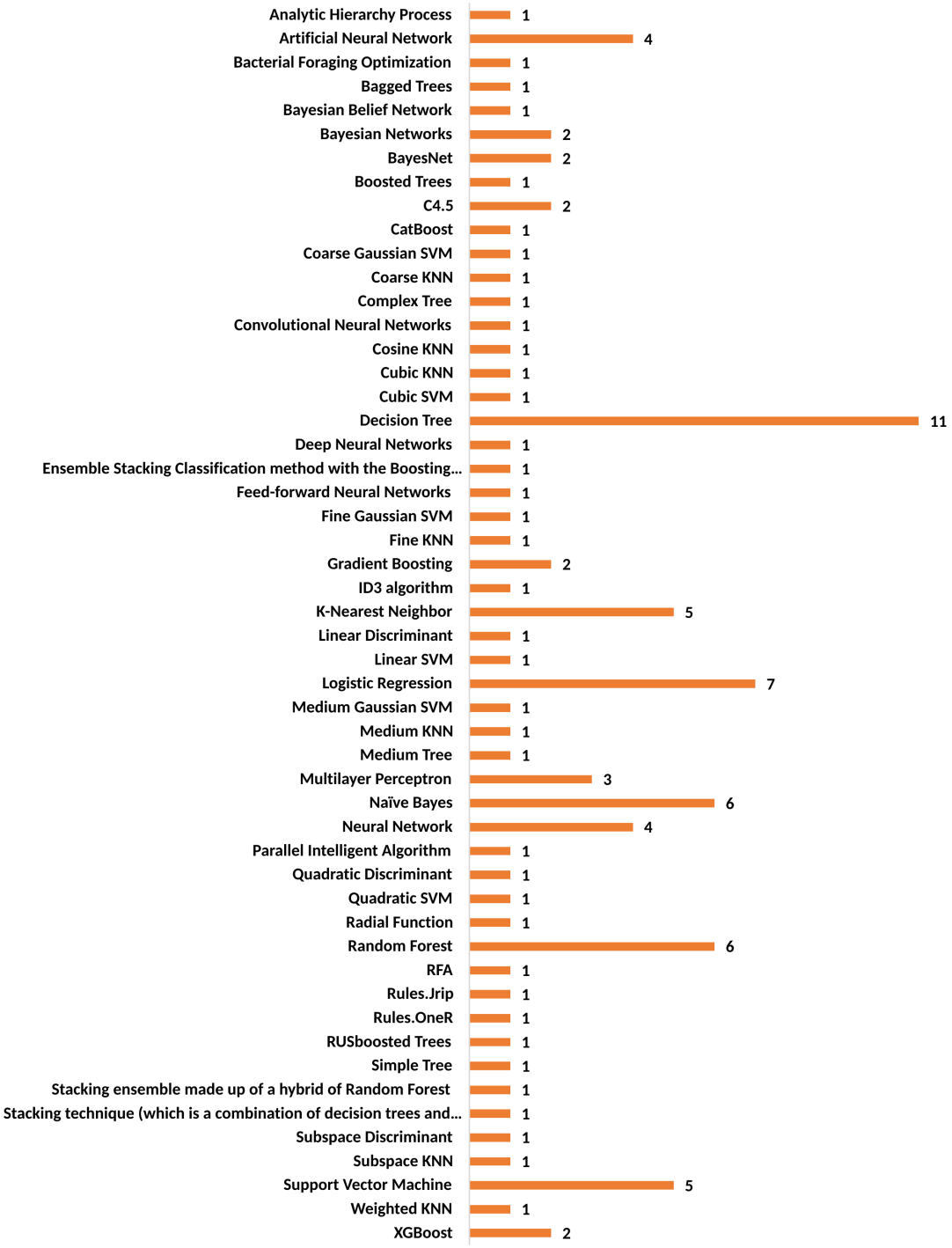
| Método | Estudios |
|---|---|
| *Synthetic Minority Oversampling Technique* (SMOTE) | 4 |
| *Clustering Based Oversampling*(CBO) | 1 |
| *Random Over-Sampling* (ROS) | 1 |
| *RUSboost* | 1 |
| *Tomek link* | 1 |
| *Undersampling technique* | 1 |
| *No especifica* | 16 |

*Fuente:* Elaboración propia.

## *4.5.* Análisis de los algoritmos de ML y métricas de evaluación

El 100 % de los estudios utilizaron algoritmos de ML de aprendizaje supervisado con el objetivo de predecir la deserción escolar en las IES, el 84 % de los algoritmos empleados son para clasificación, el 7 % de regresión y el resto, el 3 % de *clustering*; los algoritmos de clasificación más utilizados fueron,*DT, RF, NB, SVM, K-Nearest Neighbor* (KNN) y *Neural Network* (NN), respecto a los algoritmos de regresión el más utilizado fue LR y de los algoritmos de *clustering* se utilizaron los siguientes: *Subspace Discriminant, Radial Function y Linear Discriminant*. Pero podemos resaltar que el algoritmo más empleado dentro de los 23 seleccionados fue *DT* en once (11.1 %) estudios como lo podemos observar en la Figura 9. Para la implementación de las técnicas de ML, solo 17 estudios indicaron las herramientas utilizadas, entre las que se encuentran: Phyton, WEKA, R, Matlab, RapidMiner Studio, Apache Spark y SPSS *Statistics*.

Figura 9: Cantidad de algoritmos de ML empleados en los estudios seleccionados.



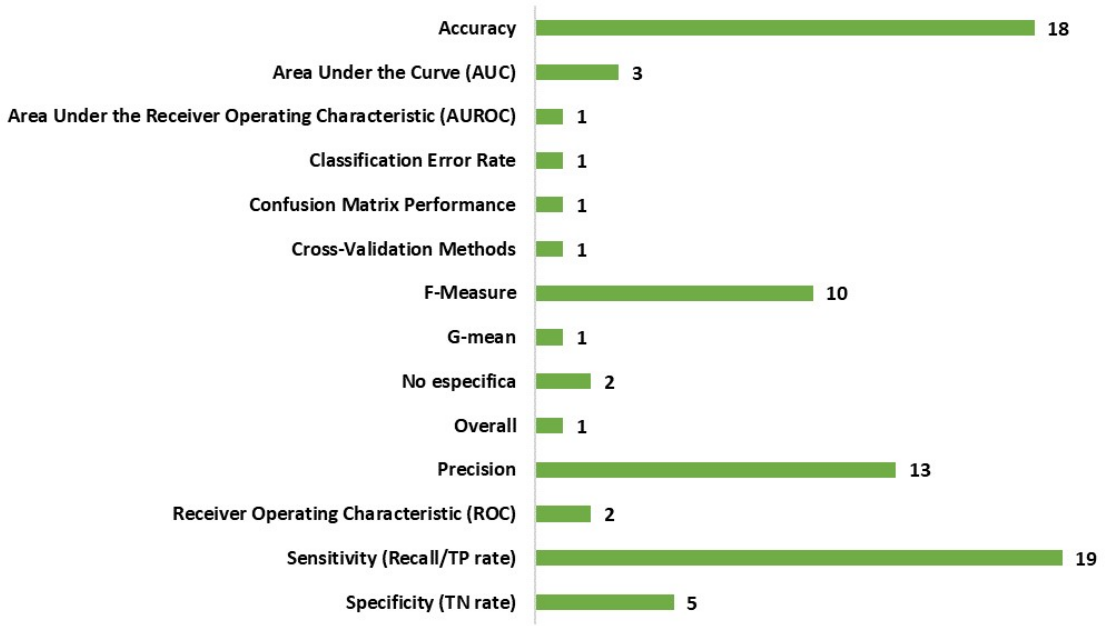

*Fuente:* Elaboración propia.

En la Anexo 1 se organizan las métricas (columnas) y algoritmos predictivos (renglones) empleados en los estudios seleccionados. En la intersección de cada columna

y renglón, se encuentran los estudios que emplearon un algoritmo predictivo junto a la métrica de evaluación específica para medir el desempeño del modelo. La estructura facilita la comparación entre técnicas y métricas, lo que permite identificar patrones de investigación. Así como también ayuda a visualizar cuales combinaciones han sido más utilizadas.

En la Figura 10 se puede observar el análisis de las métricas de evaluación, en esta se observa que la métrica más utilizada es *Sensitivity* (también conocida como *Recall* o *TP rate*) representando el 24.4 % de los estudios, el 23.1 % de los estudios utilizó la métrica de *Accuracy*, el 16.7 % empleo *Precision* y 12.8 % utilizó la métrica *F-Measure*, por nombrar las cuatro métricas más empleadas en los estudios seleccionados.

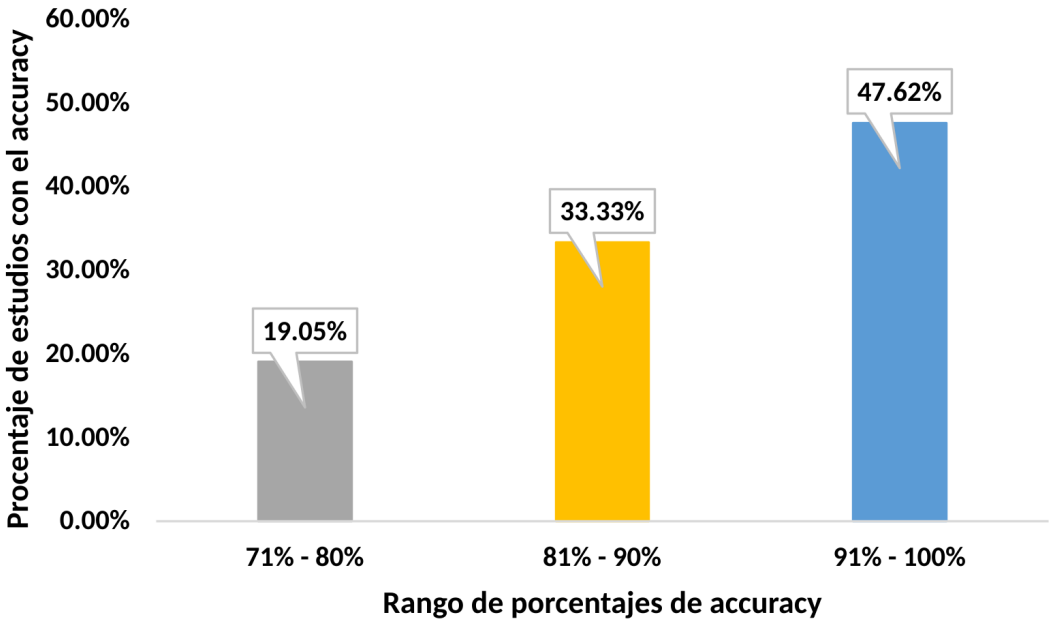
Figura 10: Métricas de evaluación empleadas en los estudios seleccionados.




*Fuente:* Elaboración propia.

La mayoría de los estudios seleccionados utilizaron la métrica de exactitud *(accuracy* para medir el desempeño de los algoritmos utilizados para la predicción de la deserción escolar, como se puede observar en la Figura 11, la mayoría de los estudios tiene un alto porcentaje de rendimiento, el 80 % de los estudios tuvo una exactitud de 81 % o más.

Figura 11: Distribución del porcentaje más alto del *accuracy* en los estudios seleccionados.




*Fuente:* Elaboración propia.

# 5. Conclusiones e investigaciones futuras

El objetivo principal de esta investigación fue realizar una revisión sistemática de la literatura del periodo 2018 a 2022, esta investigación nos permite conocer los patrones de las investigaciones existentes en WOS y SCOPUS sobre la deserción escolar universitaria; en el estudio se hace un análisis de las técnicas de *machine learning* utilizadas de inicio a fin en el proceso de predicción de la deserción escolar, iniciando en la selección de atributos, división de la base de datos, manejo de *overfitting*, balanceo de la información, técnicas predictivas hasta la evaluación de los modelos.

Para esto, se aplicó la metodología PRISMA, utilizando un protocolo de búsqueda con criterios de inclusión y exclusión. Como resultado, se seleccionaron 23 estudios pertinentes para responder las preguntas de investigación planteadas. Inicialmente, se presenta información sobre la calidad de los estudios, datos demográficos, años de publicación, técnicas de selección de atributos, tamaño de las bases de datos y otros criterios metodológicos relevantes.

El estudio aborda el análisis y la síntesis desde la perspectiva del ML predictivo. En relación con la primera pregunta de investigación (RQ1), se identificaron tres atributos recurrentes en los estudios: género, edad e ingresos familiares. Para responder a la segunda pregunta (RQ2), se realizó un análisis detallado del uso de técnicas de ML predictivo a lo largo del proceso predictivo, entre las técnicas más empleadas se encuentran: validación cruzada (*10-fold cross-validation*) para la partición de datos,

SMOTE para el balanceo de clases, y algoritmos como DT, LR, NB, RF y KNN como los más comunes en la predicción. Las métricas más frecuentemente reportadas fueron *sensibility, accuracy, precision* y *F-measure*.

Este tipo de estudios permite evaluar, comparar y sintetizar evidencia significativa en torno a preguntas de investigación específicas, ofreciendo una visión general y estructurada del estado actual del tema. Para futuras investigaciones, se sugiere ampliar el rango temporal de análisis, explorar nuevas tendencias emergentes, incluir estudios provenientes de conferencias, capítulos de libros y otras fuentes académicas, considerar publicaciones en diferentes idiomas y expandir la búsqueda en otras bases de datos digitales y áreas del conocimiento.

## Referencias bibliográficas

Agrusti, F., Mezzini, M., & Bonavolontà, G. (2020). Deep learning approach for predicting university dropout: A case study at roma tre university. *Journal of e-learning and knowledge society*, *16*(1), 44-54. https://doi.org/10.20368/1971-8829/1135192

Ahmad, F., Ismail, N. H., & Aziz, A. A. (2015). The Prediction of Students' Academic Performance Using Classification Data Mining Techniques. *Applied Mathematical Sciences*, *9*(129), 6415-6426. https://doi.org/10.12988/ams.2015.53289

Alaoui, H. H., Hachem, E., Ziti, C., & Bassiri, M. (2021). The use of predictive analyzes for university dropout cases. *Iraqi Journal of Science*, 44-51. https://doi.org/10.24996/ijs.2021.SI.1.7

Alban, M., & Mauricio, D. (2019). Neural networks to predict dropout at the universities. *International Journal of Machine Learning and Computing*, *9*(2), 149-153. https://doi.org/10.18178/ijmlc.2019.9.2.779

Albreiki, B., Zaki, N., & Alashwal, H. (2021). A Systematic Literature Review of Student' Performance Prediction Using Machine Learning Techniques. *Education Sciences*, *11*(9), 552. https://doi.org/10.3390/educsci11090552

Arizmendi, C. J., Bernacki, M. L., Raković, M., Plumley, R. D., Urban, C. J., Panter, A. T., Greene, J. A., & Gates, K. M. (2022). Predicting student outcomes using digital logs of learning behaviors: Review, current standards, and suggestions for future work. *Behavior Research Methods*. https://doi.org/10.3758/s13428-022-01939-9

Azevedo, B. F., Rocha, A. M. A. C., & Pereira, A. I. (2024). Hybrid approaches to optimization and machine learning methods: a systematic literature review. *Machine Learning*, *113*(7), 4055-4097. https://doi.org/10.1007/s10994-023-06467-x

Balaji, P., Alelyani, S., Qahmash, A., & Mohana, M. (2021). Contributions of Machine Learning Models towards Student Academic Performance Prediction: A Systematic Review. *Applied Sciences*, *11*(21), 10007. https://doi.org/10.3390/app112110007

Barramuno, M., Meza-Narvaez, C., & Galvez-Garcia, G. (2022). Prediction of student attrition risk using machine learning. *Journal of Applied Research in Higher Education*, *14*(3), 974-986. https://doi.org/10.1108/jarhe-02-2021-0073

Bedregal-Alpaca, N., Cornejo-Aparicio, V., Zarate-Valderrama, J., & Yanque-Churo, P. (2020). Classification models for determining types of academic risk and predicting dropout in university students. *International Journal of Advanced Computer Science and Applications*, *11*(1), 266-272. https://doi.org/10.14569/ijacsa.2020.0110133

Boukhlif, M., Hanine, M., Kharmoum, N., Noriega, A. R., Obeso, D. G., & Ashraf, I. (2024). Natural Language Processing-Based Software Testing: A Systematic Literature Review. *IEEE Access*, *12*, 79383-79400. https://doi.org/10.1109/ACCESS.2024.3407753

Brannon, P. M., Taylor, C. L., & Coates, P. M. (2014). Use and Applications of Systematic Reviews in Public Health Nutrition. *Annual Review of Nutrition*, *34*(1), 401-419. https://doi.org/10.1146/annurev-nutr-080508-141240

Cardona, T., Cudney, E. A., Hoerl, R., & Snyder, J. (2020). Data Mining and Machine Learning Retention Models in Higher Education. *Journal of College Student Retention: Research, Theory & Practice*, 1521025120964920. https://doi.org/10.1177/1521025120964920

Chen, L., Chen, P., & Lin, Z. (2020). Artificial Intelligence in Education: A Review. *IEEE Access*, *8*, 75264-75278. https://doi.org/10.1109/ACCESS.2020.2988510

Chu, H.-C., Hwang, G.-H., Tu, Y.-F., & Yang, K.-H. (2022). Roles and research trends of artificial intelligence in higher education: A systematic review of the top 50 most-cited articles. *Australasian Journal of Educational Technology*, *38*(3), 22-42. https://doi.org/10.14742/ajet.7526

Cui, Y., Chen, F., Shiri, A., & Fan, Y. (2019). Predictive analytic models of student success in higher education. *Information and Learning Sciences*, *120*(3/4), 208-227. https://doi.org/10.1108/ILS-10-2018-0104

Dalipi, F., Imran, A. S., & Kastrati, Z. (2018). MOOC dropout prediction using machine learning techniques: Review and research challenges. *2018 IEEE Global Engineering Education Conference (EDUCON)*, 1057-1064. https://doi.org/10.1109/EDUCON.2018.8363340

Dancot, J., Petre, B., Dardenne, N., Donneau, A.-F., Detroz, P., & Guillaume, M. (2021). Exploring the relationship between first-year nursing student self-esteem and dropout: A cohort study. *Journal of Advanced Nursing*, *77*(6), 2748-2760. https://doi.org/10.1111/jan.14806

da Silva, J. J., & Roman, N. T. (2021). Predicting Dropout in Higher Education: a Systematic Review. *Anais do XXXII Simpósio Brasileiro de Informática na Educação*.

Datawrapper. (2025). *Datawrapper Map [Aplicación web]* [Aplicación web]. https://www.datawrapper.de/maps

Daza, A. (2022). A stacking based hybrid technique to predict student dropout at universities. *Journal of Theoretical and Applied Information Technology*, *100*(13), 4790-4801. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85134385654

Delen, D. (2010). A comparative analysis of machine learning techniques for student retention management. *Decision Support Systems*, *49*, 498-506. https://doi.org/10.1016/j.dss.2010.06.003

Delen, D., Topuz, K., & Eryarsoy, E. (2020). Development of a Bayesian Belief Network-based DSS for predicting and understanding freshmen student attrition. *European Journal of Operational Research*, *281*(3), 575-587. https://doi.org/10.1016/j.ejor.2019.03.037

Eck, N., & Waltman, L. (2023). VOSviewer [Version 1.6.19 [Aplicación de escritorio]. Centro de Estudios de Ciencia y Tecnología (CWTS), Universidad de Leiden].

Fahd, K., Venkatraman, S., Miah, S. J., & Ahmed, K. (2022). Application of machine learning in higher education to assess student academic performance, at-risk, and attrition: A meta-analysis of literature. *Education and Information Technologies*, *27*(3), 3743-3775. https://doi.org/10.1007/s10639-021-10741-7

Freitas, F. A. D. S., Vasconcelos, F. F. X., Peixoto, S. A., Hassan, M. M., Ali Akber Dewan, M., de Albuquerque, V. H. C., & Rebouças Filho, P. P. (2020). IoT system for school dropout prediction using machine learning techniques based on socioeconomic data. *Electronics (Switzerland)*, *9*(10), 1613. https://doi.org/10.3390/electronics9101613

Guarín, C. E. L., Guzmán, E. L., & González, F. A. (2015). A Model to Predict Low Academic Performance at a Specific Enrollment Using Data Mining. *IEEE Revista Iberoamericana de Tecnologias del Aprendizaje*, *10*(3), 119-125. https://doi.org/10.1109/RITA.2015.2452632

Hernández González, A. G., Melendez Armenta, R. A., Morales Rosales, L. A., Garcia Barrientos, A., Tecpanecatl Xihuitl, J. L., & Algredo, I. (2016). Comparative Study of Algorithms to Predict the Desertion in the Students at the ITSM-Mexico. *IEEE Latin America Transactions*, *14*(11), 4573-4578. https://doi.org/10.1109/TLA.2016.7795831

Huang, J., Saleh, S., & Liu, Y. (2021). A Review on Artificial Intelligence in Education. *Academic Journal of Interdisciplinary Studies*, *10*(3), 206. https://doi.org/10.36941/ajis-2021-0077

Hutagaol, N., & Suharjito. (2019). Predictive modelling of student dropout using ensemble classifier method in higher education. *Advances in Science, Technology and Engineering Systems*, *4*(4), 206-211. https://doi.org/10.25046/aj040425

Jin, Q., Imbrie, P. K., Lin, J. J. J., & Chen, X. (2011). A Multi-Outcome Hybrid Model for Predicting Student Success in Engineering. *2011 ASEE Annual Conference & Exposition*.

Kabathova, J., & Drlik, M. (2021). Towards predicting student's dropout in university courses using different machine learning techniques. *Applied Sciences (Switzerland)*, *11*(7), 3130. https://doi.org/10.3390/app11073130

Kaddoura, S., Popescu, D. E., & Hemanth, J. D. (2022). A systematic review on machine learning models for online learning and examination systems. *PeerJ Comput Sci*, *8*, e986. https://doi.org/10.7717/peerj-cs.986

Lackey, L. W., Lackey, W. J., Grady, H. M., & Davis, M. T. (2003). Efficacy of Using a Single, Non-Technical Variable to Predict the Academic Success of Freshmen Engineering Students. *Journal of Engineering Education*, *92*(1), 41-48. https://doi.org/10.1002/j.2168-9830.2003.tb00736.x

Lee, Z. J., & Lee, C. Y. (2020). A parallel intelligent algorithm applied to predict students dropping out of university. *Journal of Supercomputing*, *76*(2), 1049-1062. https://doi.org/10.1007/s11227-019-03093-0

Moreira da Silva, D. E., Solteiro Pires, E. J., Reis, A., de Moura Oliveira, P. B., & Barroso, J. (2022). Forecasting Students Dropout: A UTAD University Study. *Future Internet*, *14*(3), 76. https://doi.org/10.3390/fi14030076

Munir, H., Vogel, B., & Jacobsson, A. (2022). Artificial Intelligence and Machine Learning Approaches in Digital Education: A Systematic Revision. *Information*, *13*(4), 203. https://doi.org/10.3390/info13040203

Naseem, M., Chaudhary, K., & Sharma, B. (2022). Predicting Freshmen Attrition in Computing Science using Data Mining. *Education and Information Technologies*, *27*(7), 9587-9617. https://doi.org/10.1007/s10639-022-11018-3

Nedeva, V., & Pehlivanova, T. (2020). Using machine learning to analyze university students' dropout rate - a case study. *International Journal on Information Technologies and Security*, *12*(3), 37-50.

Nguyen, M., Chaudhry, S. I., Desai, M. M., Chen, C., Mason, H. R. C., McDade, W. A., Fancher, T. L., & Boatright, D. (2022). Association of Sociodemographic Characteristics With US Medical Student Attrition. *JAMA Internal Medicine*, *182*(9), 917-924. https://doi.org/10.1001/jamainternmed.2022.2194

Nikolaidis, P., Ismail, M., Shuib, L., Khan, S., & Dhiman, G. (2022). Predicting Student Attrition in Higher Education through the Determinants of Learning Progress: A Structural Equation Modelling Approach. *Sustainability (Switzerland)*, *14*(20), 13584. https://doi.org/10.3390/su142013584

Niyogisubizo, J., Liao, L., Nziyumva, E., Murwanashyaka, E., & Nshimyumukiza, P. C. (2022). Predicting student's dropout in university classes using two-layer ensemble machine learning approach: A novel stacked generalization. *Computers and Education: Artificial Intelligence*, *3*, 100066. https://doi.org/10.1016/j.caeai.2022.100066

Nuankaew, P. (2019). Dropout situation of business computer students, University of Phayao. *International Journal of Emerging Technologies in Learning*, *14*(19), 117-131. https://doi.org/10.3991/ijet.v14i19.11177

Onder, E., & Batar, A. S. (2021). Investigation of Factors Affecting the Intention of Drop out of School in Academic and Vocational High School Students via Logistic Regression Analysis. *Cukurova University Faculty of Education Journal*, *50*(1), 150-179. https://doi.org/10.14812/cufej.733087

Opazo, D., Moreno, S., Álvarez-Miranda, E., & Pereira, J. (2021). Analysis of first-year university student dropout through machine learning models: A comparison between universities. *Mathematics*, *9*(20), 2599. https://doi.org/10.3390/math9202599

Oqaidi, K., Aouhassi, S., & Mansouri, K. (2022). Towards a Students' Dropout Prediction Model in Higher Education Institutions Using Machine Learning Algorithms. *International Journal of Emerging Technologies in Learning*, *17*(18), 103-117. https://doi.org/10.3991/ijet.v17i18.25567

Otero Escobar, A. D. (2021). Deserción escolar en estudiantes universitarios: estudio de caso del área económico-administrativa. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(23), 19. https://doi.org/10.23913/ride.v12i23.1084

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., & Alonso-Fernández, S. (2021). Declaración PRISMA 2020: una guía actualizada para la publicación de revisiones sistemáticas. *Revista Española de Cardiología*, *74*(9), 790-799. https://doi.org/10.1016/j.recesp.2021.06.016

Raidal, S. L., Lord, J., Hayes, L., Hyams, J., & Lievaart, J. (2019). Student selection to a rural veterinary school. 2: predictors of student performance and attrition. *Australian Veterinary Journal*, *97*(7), 211-219. https://doi.org/10.1111/avj.12816

Rochin Berumen, F. L. (2021). Deserción escolar en la educación superior en México: revisión de literatura. *Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, *12*(22), 11. https://doi.org/10.23913/ride.v11i22.821

Ryan, J. M., Potier, T., Sherwin, A., & Cassidy, E. (2021). Identifying factors that predict attrition among first year physiotherapy students: a retrospective analysis. *Physiotherapy*, *110*, 26-33. https://doi.org/10.1016/j.physio.2017.04.001

Sacală, M. D., Pătărlăgeanu, S. R., Popescu, M. F., & Constantin, M. (2021). Econometric research of the mix of factors influencing first-year students' dropout decision at the faculty of agri-food and environmental economics. *Economic Computation and Economic Cybernetics Studies and Research*, *55*(3), 203-220. https://doi.org/10.24818/18423264/55.3.21.13

Salas-Pilco, S. Z., & Yang, Y. (2022). Artificial intelligence applications in Latin American higher education: a systematic review. *International Journal of Educational Technology in Higher Education*, *19*(1), 21. https://doi.org/10.1186/s41239-022-00326-w

Sani, N. S., Nafuri, A. F. M., Othman, Z. A., Nazri, M. Z. A., & Nadiyah Mohamad, K. (2020). Drop-Out Prediction in Higher Education Among B40 Students. *International Journal of Advanced Computer Science and Applications*, *11*(11), 550-559. https://doi.org/10.14569/IJACSA.2020.0111169

Schlosser, R. W., Raghavendra, P., & Sigafoos, J. (2013). Chapter 3 Appraising Systematic Reviews: From Navigating Synopses of Reviews to Conducting One's Own Appraisal. En B. G. Cook, M. Tankersley & T. J. Landrum (Eds.), *Evidence-Based Practices* (pp. 45-64, Vol. 26). Emerald Group Publishing Limited. https://doi.org/10.1108/S0735-004X(2013)0000026005

Silva, H. A., Quezada, L. E., Oddershede, A. M., Palominos, P. I., & O'Brien, C. (2020). A Method for Estimating Students' Desertion in Educational Institutions Using the Analytic Hierarchy Process. *Journal of College Student Retention-Research Theory & Practice*, 1521025120971227. https://doi.org/10.1177/1521025120971227

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

Tinto, V. (1975). Dropout from Higher Education: A Theoretical Synthesis of Recent Research. *Review of Educational Research*, *45*(1), 89-125. https://doi.org/10.2307/1170024

Tinto, V. (1989). Definir la Deserción: Una Cuestión de Perspectiva. *Revista de la Educación Superior*, *18*(71). http://publicaciones.anuies.mx/pdfs/revista/Revista71_S1A3ES.pdf

Tsai, S. C., Chen, C. H., Shiao, Y. T., Ciou, J. S., & Wu, T. N. (2020). Precision education with statistical learning and deep learning: a case study in Taiwan. *International Journal of Educational Technology in Higher Education*, *17*(1), 12. https://doi.org/10.1186/s41239-020-00186-2

Umer, R., Susnjak, T., Mathrani, A., & Suriadi, L. (2021). Current stance on predictive analytics in higher education: opportunities, challenges and future directions. *Interactive Learning Environments*, 1-26. https://doi.org/10.1080/10494820.2021.1933542

Urbina-Najera, A. B., Camino-Hampshire, J. C., & Barbosa, C. R. (2020). University dropout: Prevention patterns through the application of educational data mining. *Relieve-Revista Electronica De Investigacion Y Evaluacion Educativa*, *26*(1), 4. https://doi.org/10.7203/relieve.26.1.16061

Utami, S., Winarni, I., Handayani, S. K., & Zuhairi, F. R. (2020). When and who dropouts from distance education? *Turkish Online Journal of Distance Education*, *21*(2), 141-152.

Viloria, A., Pineda Lezama, O. B., & Varela, N. (2019). Bayesian Classifier Applied to Higher Education Dropout. *10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN) / 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH)*.

**Anexo 1.** Métricas empleadas para evaluar el desempeño de los algoritmos de ML.

| | **Métricas** | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Algoritmo de ML** | **Accuracy** | **Precision** | **F-measure** | **Sensibility** | **Sensitivity (Recall/TP Rate)** | **Specifity (TN Rate)** | **AUC** | **ROC** | **G-mean** | **AUROC** | **Classification Error Rate** | **Confusion matrix** | **Cross-validation methods** |
| **Artificial Neural Networks** | (Bedregal-Alpaca et al., 2020; Moreira da Silva et al., 2022; Sani et al., 2020) | (Bedregal-Alpaca et al., 2020; Moreira da Silva et al., 2022; Sani et al., 2020) | (Moreira da Silva et al., 2022; Sani et al., 2020) | | (Bedregal-Alpaca et al., 2020; Moreira da Silva et al., 2022; Sani et al., 2020) | | | | | (Moreira da Silva et al., 2022) | | | |

*Continúa en la siguiente página*

| Naïve Bayes | Bayesian Networks |
|---|---|
| (Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Naseem et al., 2022; Nedeva & Pehlivanova, 2020; Opazo et al., 2021) | (Agrusti et al., 2020) |
| (Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Nedeva & Pehlivanova, 2020; Opazo et al., 2021) | (Agrusti et al., 2020) |
| (Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Agrusti et al., 2020) |
| | (Agrusti et al., 2020) |
| (Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Naseem et al., 2022; Nedeva & Pehlivanova, 2020; Opazo et al., 2021) | |
| (Naseem et al., 2022) | |
| (Naseem et al., 2022) | (Viloria et al., 2019) |
| (Naseem et al., 2022) | |
| | |
| | |
| (Kabathova & Drlik, 2021) | |
| | |
| | |

*Continúa en la siguiente página*

| **RulesJrip** | **BayesNet** | **Analytic Hierarchy Process** |
|---|---|---|
| (Nedeva & Pehlivanova, 2020) | (Nedeva & Pehlivanova, 2020) | (Silva et al., 2020) |
| (Nedeva & Pehlivanova, 2020) | (Nedeva & Pehlivanova, 2020) | |
| | | |
| | | |
| (Nedeva & Pehlivanova, 2020) | (Nedeva & Pehlivanova, 2020) | (Silva et al., 2020) |
| | | (Silva et al., 2020) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| Random Forest | Rules.OneR |
|---|---|
| (Kabathova & Drlik, 2021; Moreira da Silva et al., 2022; Naseem et al., 2022; Opazo et al., 2021; Sani et al., 2020) | (Nedeva & Pehlivanova, 2020) |
| (Kabathova & Drlik, 2021; Moreira da Silva et al., 2022; Opazo et al., 2021; Sani et al., 2020) | (Nedeva & Pehlivanova, 2020) |
| (Kabathova & Drlik, 2021; Moreira da Silva et al., 2022; Opazo et al., 2021; Sani et al., 2020) | |
| | |
| (Kabathova & Drlik, 2021; Moreira da Silva et al., 2022; Naseem et al., 2022; Opazo et al., 2021; Sani et al., 2020) | (Nedeva & Pehlivanova, 2020) |
| (Naseem et al., 2022) | |
| (Naseem et al., 2022) | |
| (Naseem et al., 2022) | |
| | |
| (Moreira da Silva et al., 2022) | |
| (Kabathova & Drlik, 2021) | |
| | |
| | |

*Continúa en la siguiente página*

| ID3 algorithm | C4.5 algorithm | Convolutional Neural Network |
|---|---|---|
| (Bedregal-Alpaca et al., 2020) | (Bedregal-Alpaca et al., 2020; Urbina-Najera et al., 2020) | (Agrusti et al., 2020) |
| (Bedregal-Alpaca et al., 2020) | (Bedregal-Alpaca et al., 2020; Urbina-Najera et al., 2020) | (Agrusti et al., 2020) |
| | (Urbina-Najera et al., 2020) | (Agrusti et al., 2020) |
| | | |
| (Bedregal-Alpaca et al., 2020) | (Bedregal-Alpaca et al., 2020; Urbina-Najera et al., 2020) | (Agrusti et al., 2020) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Bayesian Belief Network** | (Delen et al., 2020) | | | | (Delen et al., 2020) | (Delen et al., 2020) | (Delen et al., 2020) | | (Delen et al., 2020) | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **CatBoost** | (Moreira da Silva et al., 2022) | (Moreira da Silva et al., 2022) | (Moreira da Silva et al., 2022) | | (Moreira da Silva et al., 2022) | | | | | (Moreira da Silva et al., 2022) | | | |

*Continúa en la siguiente página*

| **Decision Tree** |
| --- |
| (Daza, 2022; Freitas et al., 2020; Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Naseem et al., 2022; Nuankaew, 2019; Opazo et al., 2021; Sani et al., 2020; Urbina-Najera |
| (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Nuankaew, 2019; Opazo et al., 2021; Sani et al., 2020; Urbina-Najera et al., 2020) |
| (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Opazo et al., 2021; Sani et al., 2020) |
| |
| (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Kabathova & Drlik, 2021; Naseem et al., 2022; Nuankaew, 2019; Opazo et al., 2021; Sani et al., 2020; Urbina-Najera et al., 2020) |
| (Naseem et al., 2022) |
| (Naseem et al., 2022) |
| (Naseem et al., 2022; Viloria et al., 2019) |
| |
| |
| (Kabathova & Drlik, 2021) |
| (Naseem et al., 2022) |
| (Naseem et al., 2022) |

*Continúa en la siguiente página*

| **Deep Neural Networks** | **K-Nearest Neighbors** |
|---|---|
| (Freitas et al., 2020) | (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Naseem et al., 2022; Opazo et al., 2021) |
| (Freitas et al., 2020) | (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Opazo et al., 2021) |
| (Freitas et al., 2020) | (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Opazo et al., 2021) |
| | (Freitas et al., 2020; Hutagaol & Suharjito, 2019; Naseem et al., 2022; Opazo et al., 2021) |
| (Freitas et al., 2020) | (Naseem et al., 2022) |
| | (Naseem et al., 2022) |
| | (Naseem et al., 2022) |
| | |
| | |
| | |
| | |
| | |
| | |

*Continúa en la siguiente página*

| **Support Vector Machine** | **Neural Networks** |
|---|---|
| (Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Daza, 2022; Kabathova & Drlik, 2021; Opazo et al., 2021) |
| (Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Kabathova & Drlik, 2021; Opazo et al., 2021) |
| (Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Kabathova & Drlik, 2021; Opazo et al., 2021) |
| (Alaoui et al., 2021) | |
| (Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Kabathova & Drlik, 2021; Opazo et al., 2021) |
| | |
| | |
| | (Viloria et al., 2019) |
| | |
| | |
| (Kabathova & Drlik, 2021) | (Kabathova & Drlik, 2021) |
| | |
| | |

*Continúa en la siguiente página*

| Complex tree | Multilayer Perceptron | Radial Function |
|---|---|---|
| (Barramuno et al., 2022) | (Freitas et al., 2020; Tsai et al., 2020) | |
| | (Alban & Mauricio, 2019; Freitas et al., 2020) | (Alban & Mauricio, 2019) |
| (Barramuno et al., 2022) | (Freitas et al., 2020) | |
| | (Alban & Mauricio, 2019) | (Alban & Mauricio, 2019) |
| (Barramuno et al., 2022) | (Freitas et al., 2020; Tsai et al., 2020) | |
| (Barramuno et al., 2022) | (Tsai et al., 2020) | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Linear discriminant** | **Simple tree** | **Medium tree** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| Logistic Regression | Quadratic discriminant |
|---|---|
| (Barramuno et al., 2022; Freitas et al., 2020; Kabathova & Drlik, 2021; Naseem et al., 2022; Silva et al., 2020; Tsai et al., 2020) | (Barramuno et al., 2022) |
| (Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | |
| (Barramuno et al., 2022; Freitas et al., 2020; Kabathova & Drlik, 2021; Opazo et al., 2021) | (Barramuno et al., 2022) |
| | |
| (Barramuno et al., 2022; Freitas et al., 2020; Kabathova & Drlik, 2021; Naseem et al., 2022; Opazo et al., 2021; Silva et al., 2020; Tsai et al., 2020) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022; Naseem et al., 2022; Silva et al., 2020; Tsai et al., 2020) | (Barramuno et al., 2022) |
| (Naseem et al., 2022) | |
| (Naseem et al., 2022) | |
| | |
| | |
| (Kabathova & Drlik, 2021) | |
| | |
| | |

*Continúa en la siguiente página*

| **Cubic SVM** | **Quadratic SVM** | **Linear SVM** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Coarse Gaussian SVM** | **Medium Gaussian SVM** | **Fine Gaussian SVM** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Coarse KNN** | **Medium KNN** | **Fine KNN** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Weighted KNN** | **Cubic KNN** | **Cosine KNN** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| **Subspace discriminant** | **Bagged trees** | **Boosted trees** |
|---|---|---|
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| (Barramuno et al., 2022) | (Barramuno et al., 2022) | (Barramuno et al., 2022) |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |
| | | |

*Continúa en la siguiente página*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Subspace KNN** | (Barramuno et al., 2022) | | (Barramuno et al., 2022) | | (Barramuno et al., 2022) | (Barramuno et al., 2022) | | | | | | |
| **RUSboosted trees** | (Barramuno et al., 2022) | | (Barramuno et al., 2022) | | (Barramuno et al., 2022) | (Barramuno et al., 2022) | | | | | | |

*Continúa en la siguiente página*

| Gradient Boosting | XGBoost |
|---|---|
| (Niyogisubizo et al., 2022; Opazo et al., 2021) | (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022) |
| (Niyogisubizo et al., 2022; Opazo et al., 2021) | (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022) |
| (Niyogisubizo et al., 2022; Opazo et al., 2021) | (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022) |
| | |
| (Niyogisubizo et al., 2022; Opazo et al., 2021) | (Moreira da Silva et al., 2022; Niyogisubizo et al., 2022) |
| | (Niyogisubizo et al., 2022) |
| (Niyogisubizo et al., 2022; Opazo et al., 2021) | |
| | |
| | (Moreira da Silva et al., 2022) |
| | |
| | |
| | |

*Continúa en la siguiente página*

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Feed-forward Neural Networks** | (Niyogisubizo et al., 2022) | (Niyogisubizo et al., 2022) | (Niyogisubizo et al., 2022) | | (Niyogisubizo et al., 2022) | | (Niyogisubizo et al., 2022) | | | | | |
| **Stacking Technique (which is a combination of decision trees and neural networks)** | (Daza, 2022) | | | | | | | | | | | |

*Continúa en la siguiente página*

| **Stacking Ensemble made up of a hybrid of RF** | **Ensemble Stacking Classification method with the Boosting Gradient algorithm** |
|---|---|
| (Niyogisubizo et al., 2022) | (Hutagaol & Suharjito, 2019) |
| (Niyogisubizo et al., 2022) | (Hutagaol & Suharjito, 2019) |
| (Niyogisubizo et al., 2022) | (Hutagaol & Suharjito, 2019) |
| | |
| (Niyogisubizo et al., 2022) | (Hutagaol & Suharjito, 2019) |
| | |
| (Niyogisubizo et al., 2022) | |
| | |
| | |
| | |
| | |
| | |
| | |

*Fuente:* Elaboración propia.