Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), septiembre-octubre 2025,
Volumen 9, Número 5.

https://doi.org/10.37811/cl_rcm.v9i5

MINERÍA DE DATOS EDUCATIVA CON WEKA: APLICACIONES Y MODELADO PREDICTIVO

EDUCATIONAL DATA MINING WITH WEKA: APPLICATIONS AND PREDICTIVE MODELING

Yolanda Moyao Martínez
Benemérita Universidad Autónoma de Puebla

DOI: https://doi.org/10.37811/cl_rcm.v9i5.21172

Minería de datos educativa con Weka: aplicaciones y modelado predictivo.

Yolanda Moyao Martínez[1]
yolanda.moyao@correo.buap.mx
https://orcid.org/0000-0002-7259-3525
Benemérita Universidad Autónoma de Puebla
México

RESUMEN

El presente trabajo tiene como objetivo explorar la aplicación de la minería de datos educativa para mejorar las técnicas de enseñanza y aprendizaje con base en el análisis de datos obtenidos por medio de un cuestionario. Para ello, se utiliza el software libre Weka 3.8.6, que permite el preprocesamiento, regresión y agrupamiento de datos educativos, facilitando la identificación de patrones y elementos clave en el aprovechamiento académico. La metodología incluye la preparación del conjunto de datos en formato compatible para su análisis, la aplicación de filtros para limpieza y transformación de los datos, y el uso del modelo predictivo de regresión M5P de Weka basado en árboles de decisión, para analizar variables relevantes como ausencias escolares, edad, y participación en actividades extracurriculares. Entre los principales hallazgos, se destaca la capacidad del modelo para predecir el rendimiento académico con una correlación moderada, lo que ayuda a identificar estudiantes que se encuentran en riesgo y apoyar la toma de decisiones didácticas. Estos resultados evidencian que la minería de datos puede influir significativamente a la personalización del aprendizaje y al progreso continuo de los métodos de enseñanza. La integración de estas técnicas representa un recurso clave en el contexto educativo de México, para reforzar la toma de decisiones didácticas y también para impulsar la equidad de oportunidades académicas.

Palabras clave: minería de datos educativa, modelado predictivo, weka, rendimiento académico

[1] Autor principal
Correspondencia: yolanda.moyao@correo.buap.mx

Educational Data Mining with Weka: Applications and Predictive Modeling

ABSTRACT

The present work aims to explore the application of educational data mining to improve teaching and learning techniques based on the analysis of data obtained through a questionnaire. For this purpose, the free software Weka 3.8.6 is used, which allows preprocessing, regression, and clustering of educational data, facilitating the identification of patterns and key elements in academic achievement. The methodology includes preparing the dataset in a compatible format for analysis, applying filters for data cleaning and transformation, and using Weka's M5P regression predictive model based on decision trees to analyze relevant variables such as school absences, age, and participation in extracurricular activities. Among the main findings, the model's ability to predict academic performance with moderate correlation stands out, helping to identify students at risk and support educational decision-making. These results show that data mining can significantly influence the personalization of learning and the continuous improvement of teaching methods. The integration of these techniques represents a key resource in Mexico's educational context to strengthen educational decision-making and also to promote equity in academic opportunities.

***Keywords*:** educational data mining, predictive modeling, weka, academic performance

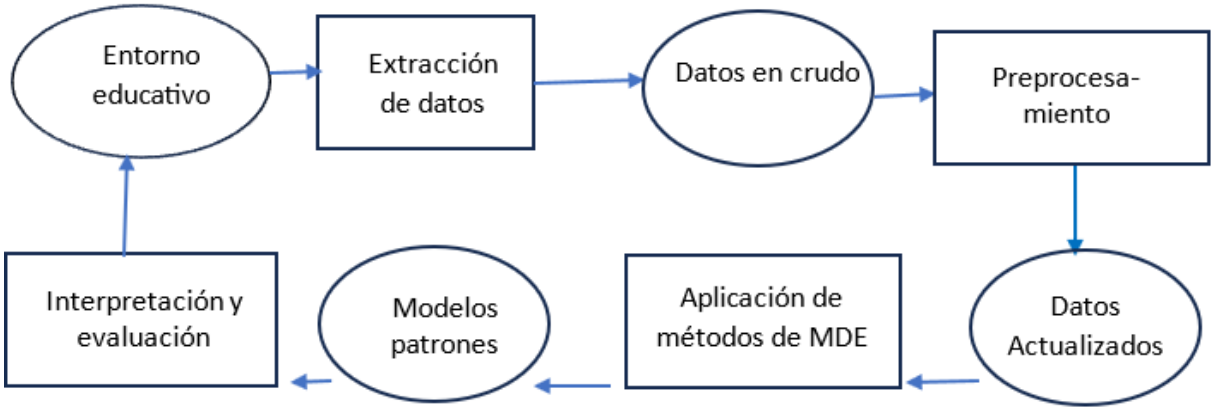
Artículo recibido 02 setiembre 2025
Aceptado para publicación: 29 setiembre 2025

INTRODUCCIÓN

La Minería de Datos (MD) es un campo de la estadística y las ciencias de la computación que ayuda a identificar patrones, tendencias y relaciones significativas, las cuales no son inmediatamente visibles en grandes volúmenes de conjuntos de datos, la minería de datos ayuda a desarrollar modelos capaces de descubrir conexiones entre millones o miles de millones de registros, con el objetivo de transformar estos datos en "bruto" en conocimiento práctico (Romero y Ventura, 2024).

Esta disciplina se utiliza en diversas áreas como, investigación científica, detección de fraude, educación, salud, ciencia, ingeniería, entre otras. En particular, en el contexto educativo, la minería de datos tiene por objetivo el desarrollo de métodos que utilicen datos de plataformas educacionales para comprender mejor a los estudiantes y el entorno en el que aprenden (Romero y Ventura, 2024).

Figura1. Proceso de la Minería de Datos Educativa.



Fuente: Adaptada de https://www.researchgate.net/figure/Figura-1-Proceso-de-la-mineria-de-datos-educativa_fig1_351055258.

La Minería de Datos Educativa es una disciplina emergente que se enfoca en técnicas, herramientas y la investigación diseñadas para extraer automáticamente conocimiento a partir de grandes volúmenes de datos educativos en el ámbito de la enseñanza, con el objetivo de descubrir información nueva y útil para mejorar los procesos educativos (Bhatt, Sajja, y Liyanage, 2020). Este proceso se muestra en la figura 1. En la última década esta disciplina ha evolucionado a gran velocidad y gracias a ello han surgido diferentes términos, como Analítica Académica, Analítica Institucional, Analítica de la Enseñanza, Educación Basada en Datos, Toma de Decisiones Basada en Datos en Educación, Big Data

en Educación y Ciencia de Datos Educativos (Romero y Ventura, 2024).

Esta disciplina tiene diferentes aplicaciones como el análisis de datos de estudiantes, de docentes, de las evaluaciones, de plataformas virtuales, entre otras. Las dependencias educativas en basa a los resultados obtenidos pueden llegar a conclusiones y tomar decisiones bien informadas, como la actualización o el diseño de nuevos contenidos temáticos o atender a estudiantes con problemas de rendimiento y así prevenir la deserción escolar.

Los docentes pueden mejorar sus prácticas profesionales, diseñarlas o adaptarlas a las necesidades en base a los contenidos temáticos. También pueden identificar conceptos de aprendizaje complejo para los estudiantes y así planear una mejor práctica de enseñanza o incluso personalizar la enseñanza según las necesidades de cada alumno (Bhatt, Sajja, y Liyanage, 2020).

Este campo de estudio tiene por objetivo contar con información suficiente para la toma de decisiones en los procesos educativos, mediante la identificación de factores clave que influyen en el rendimiento académico de los estudiantes, en la deserción escolar, en los estilos de aprendizaje, o en el uso de recursos, contribuyendo así a mejorar las prácticas y resultados educativos en las escuelas de cualquier nivel educativo (Lampropoulos, 2023; Cedillo Arce et al., 2024).

La aplicación de minería de datos en la educación permite obtener múltiples beneficios, entre los que destacan la disminución de la deserción escolar a través del análisis de los resultados y de la identificación de patrones de riesgo, la mejora de la práctica docente a partir del análisis de resultados obtenidos y de las necesidades educativas particulares del alumnado, y la personalización del aprendizaje adaptado a las necesidades y ritmo personal de aprendizaje de cada estudiante (Ordoñez-Avila et al., 2023).

Actualmente las sociedades se encuentran inmersas en la era digital que crece a gran velocidad, en esta las fuentes de datos se presentan en diversas formas y tamaños, por lo tanto, su gestión requiere de herramientas tecnológicas que sean capaces de procesarlos. La minería de datos educativa no se queda atrás, pues esta permite la gestión de una gran cantidad de datos educativos que requieren del uso de herramientas digitales para ser procesados y posteriormente analizados y con ello descubrir patrones que no son obvios, con el objetivo de proponer estrategias en aras de la mejora continua en el contexto educativo (Cedillo Arce et al., 2024; Andrade-Girón et al., 2023).

A pesar de que existen múltiples investigaciones sobre minería de datos educativa, la mayoría abordan desafíos en contextos internacionales y principlamente con un enfoque teórico. Esto deja un vacío significativo en el desarrollo de estudios que integren estos métodos con las demandas partículares de los sistemas educativos latinoamericanos, en partícular, del contexto mexicano. El presente estudio responde a esa demanda al aplicar métodos de modelado predictivo como el M5P, con el objetivo de proporcionar información clave que pueda influir en las políticas educativas y prácticas educativas.

METODOLOGÍA

El enfoque de la investigación es cuantitativo, dado que se utiliza minería de datos para el análisis y modelado predictivo sobre datos educativos numéricos. El tipo de estudio es aplicativo y predictivo, ya que se aplican técnicas de análisis de datos para prever el desempeño académico de estudiantes.

El diseño es observacional y transversal, pues se trabaja con un conjunto de datos recolectados en un momento específico, sin ninguna intervención experimental. La población de estudio está conformada por 383 estudiantes del complejo regional de Tehuacán de la Benemérita Universidad Autónoma de Puebla, para el entrenamiento del modelo se tomaron 307 instancias.

La recolección de datos se realizó a través de un cuestionario estructurado de 31 preguntas, el cual se diseñó de tal manera que se pudieraa recopilar información clave para el estudio, tal como, factores académicos, personales y demográficos, familiares y sociales y contextuales con el objetivo de obtener una visión integral de los factores que afectan el aprovechamiento académico de los estudiantes de nivel medio superior en Puebla. Se emplearon técnicas computacionales de preprocesamiento de datos, como limpieza y transformación mediante filtros no supervisados y supervisados, para preparar la información para el modelado con algoritmos de regresión en la herramienta Weka.

Entre los materiales de apoyo se utilizaron los algoritmos específicos del software, como M5P para regresión, y métricas estadísticas para la validación del modelo: coeficiente de correlación, error absoluto medio, error cuadrático medio y error relativo.

En cuanto a consideraciones éticas, se garantizó la confidencialidad y anonimato de los datos utilizados, respetando las normativas institucionales para el manejo de información estudiantil. Los criterios de inclusión fueron registros completos con las variables requeridas, mientras que se excluyeron datos con valores faltantes o inconsistentes.

Esta metodología permitió desarrollar un análisis riguroso y replicable para entender y predecir variables educativas a través de minería de datos.

El proceso utilizado para el análisis de los datos comienza con definir el origen de los datos, que pueden ser diversos formatos y fuentes como archivos planos, bases de datos SQL, URLs, etc. Luego, los datos se preprocesan, lo cual incluye la limpieza y transformación de los datos mediante filtros específicos para preparar los datos para los modelos de aprendizaje automático.

Se aplica el modelo de regresión en la herramienta Weka, utilizando algoritmos como M5P para la regresión. El análisis se hace sobre un conjunto de datos con 383 registros y 31 variables, donde la variable dependiente es la calificación del primer semestre, una variable numérica continua que se pretende predecir. El modelo fue entrenado con 307 instancias. Se diseñaron modelos lineales basados en reglas para predecir la calificación según variables como edad, ausencias escolares y materias reprobadas, entre otras.

Los datos fueron procesados con filtros no supervisados y supervisados según el objetivo del análisis. Se utilizaron técnicas de selección y transformación de variables para optimizar el proceso de modelado. La evaluación del modelo incluye métricas como coeficiente de correlación, error absoluto medio, error cuadrático medio y error relativo, con una interpretación sobre la capacidad predictiva del modelo de estudio.

RESULTADOS Y DISCUSIÓN

Los resultados obtenidos después de aplicar el modelo de regresión M5P se muestran en la figura 2. El modelo se basó en las siguientes características:

- Se entrenó con los datos de 307 estudiantes de una población total de 383.
- El modelo M5P construyó 3 modelos lineales (LM1, LM2 y LM3) basados en reglas.
- Para el proceso de predicción, cada nueva instancia o estudiante puede ser evaluado con alguna de esas reglas, y su calificación de primer semestre se puede estimar en base a sus datos, como edad, ausencias, materias reprobadas, etc.

Estructura del árbol:

El árbol que construyó el modelo M5P se compone de 3 reglas o ramas terminales, y cada una de estas tiene su propio modelo lineal:

- LM1: para alumnos con materias_reprobadas = dos, una o ninguna ≤ 0.5, esto se interpreta como que el estudiante "sí reprobó" alguna materia. De tal forma, que esta regla se aplica solamente a aquellos alumnos que sí reprobaron.
- LM2: para alumnos que reprobaron bajo otra condición adicional (por ejemplo, reprobar solo “dos” o “una” materia) que define un subgrupo adicional dentro del grupo LM1. Esta regla se aplica a un subgrupo de materias_reprobadas = dos, una, ninguna > 0.5. Para ello, WEKA realiza una subdivisión en el grupo de alumnos que si reprobaron.
- LM3: para alumnos con materias_reprobadas = ninguna > 0.5. Esta regla se aplica a aquellos alumnos que “no reprobaron” ninguna materia.

Se identificaron variables con impacto positivo en la calificación, como edad, nivel educativo del padre y profesión de la madre, actividades extracurriculares, asistir a guarderías y mantener buenas relaciones familiares. Del mismo modo se identificaron variables con impacto negativo como número de ausencias escolares, en algunos modelos también la edad tuvo impacto negativo.

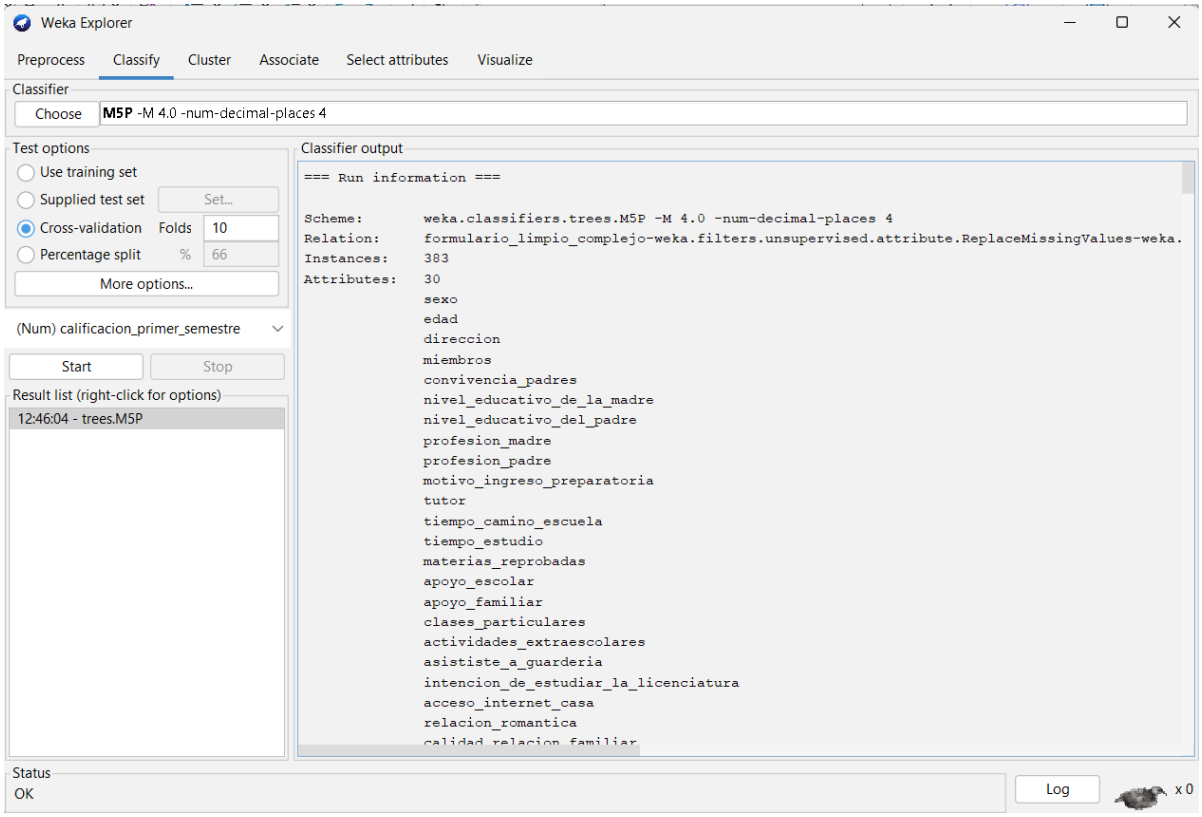
Los resultados indican que el modelo tiene una capacidad predictiva moderada, no excelente, con un coeficiente de correlación 0.4212 y error promedio absoluto de 0.6449.

La minería de datos aplicada permitió identificar variables relevantes que influyen en el rendimiento académico, tales como ausencias escolares, edad, materias reprobadas, nivel educativo y profesión de los padres, así como aspectos sociales y familiares.

Estos hallazgos sugieren que la integración de datos sociodemográficos y académicos es indispensable para comprender el rendimiento y apoyar decisiones educativas adaptadas a las necesidades de cada alumno.

Además, se destaca la importancia de seguir explorando modelos más robustos para mejorar la capacidad predictiva y facilitar intervenciones tempranas para reducir la deserción escolar y mejorar el éxito académico.

Figura 2. Resultados de la aplicación del modelo M5P.



Fuente: Tomada de Hall et al., 2019.

En base a los resultados arrojados por Weka, se puede observar que cada uno de los modelos lineales LM1, LM2 y LM3 calculan una predicción para la calificación del primer semestre, a través de la combinación ponderada de diferentes variables:

Para el caso de LM1 con n = 15 instancias o estudiantes del conjunto de entrenamiento fueron clasificadas con una precisión interna del 97.02 % en esta hoja del árbol, es decir, que cumplieron las condiciones para ser predichas por la regla lineal LM1.

Se pudo observar que la variable edad impacta de forma positiva en la calificación, es decir que a mayor edad se tiene una mayor calificación, mientras que el número de ausencias escolares tiene un impacto negativo, es decir a mayor número de faltas, se obtiene una menor calificación. Finalmente, se observó que hubo otras variables con un impacto ligeramente positivo, como, por ejemplo, tiempo libre promedio y relaciones familiares positivas.

Para cada estudiante, se identifica primero a qué regla o rama del árbol corresponde según su valor en

"materias_reprobadas" (si reprobaron o no, y si reprobaron cuántas). Posteriormente, se calcula la variable a predecir “calificación”, de atal forma que para cada regla, se usa una fórmula lineal que asigna coeficientes a las variables relevantes para ese grupo

calificación = +1.11* edad - 4.70 * ausencias_escolares + otros coeficientes pequeños

Para el caso de LM2 con n = 58 instancias o estudiantes del conjunto de entrenamiento fueron clasificadas con una precisión interna del 68.31 % en esta hoja del árbol, es decir, que cumplieron las condiciones para ser predichas por la regla lineal LM2.

Se pudo observar que las variables como nivel educativo del padre y profesión de la madre impactan de forma positiva en la calificación, mientras que la edad tiene un impacto positivo, pero menos fuerte que el que tuvo en LM1. Mientras que el número de ausencias escolares sigue teniendo un impacto negativo, pero mucho menos fuerte que el que tuvo en LM1. Finalmente, se observó que hubo otras variables con un impacto ligeramente negativo, por ejemplo, clases particulares.

calificación = + 0.4572 * edad - 0.3675 * ausencias_escolares + otros coeficientes pequeños

Para el caso de LM3 con n = 234 instancias o estudiantes del conjunto de entrenamiento fueron clasificadas con una precisión interna del 62.25 % en grupo del árbol, es decir, que cumplieron las condiciones para ser predichas por la regla lineal LM3.

Se pudo observar que las variables como actividades extracurriculares, asistir a guarderías y mantener buenas relaciones familiares impactan de forma positiva en la calificación, mientras que la edad tiene un impacto negativo. Mientras que el número de ausencias escolares sigue teniendo un impacto negativo, pero mucho menos fuerte que el que tuvo en LM1. También se detectó que este modelo representa adecuadamente el impacto combinado de factores sociales y académicos además del rendimiento previo obtenido. Finalmente, se identificaron otras variables con un impacto ligeramente positivo, por ejemplo, sexo = femenino, direccion = urbana, convivencia de los padres = juntos, nivel educativo de la madre, profesiones del padre y la madre, tutor = madre u otro, participación en actividades extraescolares, asistencia o no a guardería, calidad de la relación familiar y frecuencia de interacción social con amigos.

calificación = -0.6852 * edad - 0.0637 * ausencias_escolares + otros coeficientes pequeños.

Esto se puede interpretar como que a mayor edad del estudiante, se espera una disminución moderada en la calificación. Este efecto negativo puede reflejar posibles rezagos escolares. Por otro lado, el

impacto de las ausencias es negativo, pero más leve.

En la tabla 1, se muestra el resumen de la aplicación del modelo M5P.

Tabla 1. Resumen de las métricas obtenidas.

Métrica	**Valor**	**Interpretación**
Coeficiente de Correlación	0.4212	El resultado muestra que hay una correlación moderada, lo cual indica que el modelo tiene algo de poder predictivo.
Error absoluto medio	0.6449	Representa el error promedio absoluto en las calificaciones.
Raíz del error cuadrático medio	0.9003	Penaliza los errores grandes.
Error absoluto relativo	88.1%	Es mejor que predecir con el promedio (menos de 100%)
Error cuadrático relativo	92.68%	Es mejor esto que predecir con la media.

Fuente: Elaboración propia.

CONCLUSIONES

Se puede concluir que las variables más significativas por su aporte en el proceso de predicción en las diferentes ramas del modelo M5P fueron: edad, materias reprobadas, ausencias escolares, tiempo libre semanal, nivel educativo del padre o madre, profesión del padre o madre, relación familiar y actividades extraescolares.

El modelo M5P mostró un rendimiento moderado para la predicción del desempeño académico, identificando patrones relevantes, aunque con limitaciones en precisión. Por ejemplo, la edad y el número de ausencias influyen significativamente en las calificaciones estimadas. Además, la integración de variables sociodemográficas junto con datos académicos resultó crucial para mejorar la capacidad predictiva del modelo M5P, resaltando la multifactorialidad del rendimiento estudiantil y la necesidad

de abordajes integrales en las políticas educativas.

Por otro lado, los resultados reflejan que el modelo M5P, aunque tiene un desempeño moderado, ofrece una base útil para identificar estudiantes en riesgo, lo que abre oportunidades para desarrollar estrategias de intervención temprana y personalización del aprendizaje que puedan reducir la deserción y mejorar el éxito académico.

Con este estudio se abre la posibilidad de continuar con la investigación en diferentes direcciones. En primer lugar, se pueden explorar otros modelos predictivos más robustos, como Random Forest y redes neuronales, los cuales podrían mejorar la habilidad para predecir en comparación con el modelo M5P. También, es conveniente llevar a cabo estudios longitudinales, pues esto permitiría analizar el comportamiento del aprovechamiento académico a lo largo del tiempo.

Otro trabajo a futuro es considerar datos cualitativos, por ejemplo, cuestionarios abiertos con el objetivo de enriquecer el enfoque estadístico para tener una comprensión más a fondo de variables socioemocionales. Finalmente se propone, el desarrollo de sistemas didácticas virtuales que utilicen los modelos predictivos en tiempo real, para proporcionar información anticipada y significativa a docentes y alumnos, con el propósito de influir en la prevensión de la decersión escolar y al progreso en el desarrollo de estrategias educativas.

De acuerdo con los aportes de Cerón Garnica et al. (2025) acerca del impacto de Recursos Educativos Abiertos (REA) en el bachillerato en la etapa posterior a la pandemia, los resultados obtenidos en este trabajo de investigación sugieren que combinar técnicas de minería de datos con el uso de REA podría influir significativamente en la detección anticipada de rezagos y la planeación de estrategias individualizadas en el contexto educativo mexicano.

REFERENCIAS BIBLIOGRÁFICAS

Andrade-Girón, C. A., Yepes-Gómez, J. D., Toloza-Ruíz, C. J., & Durán-Moreno, A. A. (2023). Machine and deep learning algorithms for early dropout prediction in higher education: A systematic review. *EAI Endorsed Transactions on Scalable Information Systems, 10*(2), e6. https://doi.org/10.4108/eetsis.3586.

Bhatt, C., Sajja, P. S., & Liyanage, S. (Eds.). (2020). Utilizing educational data mining techniques for improved learning: Emerging research and opportunities. *IGI Global*. https://doi.org/10.4018/978-1-7998-0010-1.

Bouckaert, R. R., Frank, E., Hall, M., Kirkby, R., Reutemann, P., Seewald, A., & Scuse, D. (2018, September 4). WEKA manual for version 3-8-3. University of Waikato. https://user.eng.umd.edu/~austin/ence688p.d/handouts/WekaManual2018.pdf.

Cedillo Arce, J. M., Beltrán Abreo, H. M., Saltos Arce, M. I., & Soriano Barzola, F. R. (2024). Explorando la minería de datos en la gestión educativa superior: desafíos y oportunidades en la era digital. *Reincisol, 3*(5), 1368–1385. https://doi.org/10.59282/reincisol.V3(5)1367-1385.

Cerón Garnica, C., Moyao Martínez, Y., Martínez Guzmán, G., y Mila Avendaño, V. M. (2025). Análisis de los Recursos Educativos Abiertos en el rezago de aprendizajes en el bachillerato postpandemia. *EDUCATECIENCIA, 33*(2), 1-33. https://educateconciencia.com/index.php/revistaeducate/article/view/380

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2019). WEKA 3.8.6 [Software]. Universidad de Waikato. https://www.cs.waikato.ac.nz/ml/weka/.

Han, J., Kamber, M., & Pei, J. (2012). Data mining: Concepts and techniques (3rd ed.). Morgan Kaufmann. https://doi.org/10.1016/C2009-0-61819-5.

Information Resources Management Association. (2018). Information Retrieval and Management: Concepts, Methodologies, Tools, and Applications (4 vols.). *IGI Global*. https://doi.org/10.4018/978-1-5225-5191-1.

Lampropoulos, G. (2023). Educational data mining and learning analytics in the 21st century. En Encyclopedia of data science and machine learning (pp. 1642–1651). IGI Global. https://doi.org/10.4018/978-1-7998-9220-5.ch098.

Ordoñez-Avila, R., Salgado Reyes, N., Meza, J., & Ventura, S. (2023). Data mining techniques for predicting teacher evaluation in higher education: A systematic literature review. *Heliyon, 9*(3), e13939. https://doi.org/10.1016/j.heliyon.2023.e13939.

Romero, C., & Ventura, S. (2024). Educational data mining and learning analytics: An updated survey. *Journal of Educational Data Mining*. https://doi.org/10.1002/widm.1355.

University of Waikato. (s. f.). https://www.cs.waikato.ac.nz/ml/weka/.

Witten, I. H., Frank, E., & Hall, M. A. (2011). Data Mining: Practical Machine Learning Tools and Techniques (3rd ed.). Morgan Kaufmann. https://dl.acm.org/doi/book/10.5555/1972514.