



DOI: <https://doi.org/10.23913/cemys.v12i24.423>

Artículos científicos

Modelos de predicción de aprobación de estudiantes de media superior mediante examen diagnóstico y datos demográficos

Prediction models for high school students' passing grades using diagnostic tests and demographic data

Nora Diana Gaytán Ramírez

Instituto Politécnico Nacional

nora_diana@hotmail.com

<https://orcid.org/0000-0002-5159-9194>

Andrés Rico Páez

Instituto Politécnico Nacional

aricop.ipn@gmail.com

<https://orcid.org/0000-0002-6450-318X>

Resumen

El propósito de la presente investigación es elaborar modelos predictivos de la aprobación final de una asignatura de nivel medio superior de estudiantes a partir de datos demográficos y la calificación de un examen diagnóstico al inicio del periodo. La población objetivo fue de 80 estudiantes de bachillerato. De este grupo se recopilieron 5 variables demográficas de los estudiantes y la calificación de un examen diagnóstico. Para realizar los modelos de predicción se utilizaron los algoritmos de aprendizaje automático *Naïve Bayes* y árbol de decisión, con este último se obtuvo el modelo con mayor exactitud de 80%. Además, para observar el potencial uso de este modelo, se llevó a cabo la predicción de 10 estudiantes de un periodo académico posterior y se obtuvo una exactitud del 70%. Este tipo de metodología permite utilizar información que es recabada comúnmente por docentes para ser empleada en la detección de estudiantes con problemas académicos desde poco tiempo de inicio de un curso.

Palabras clave: aprendizaje automático, examen diagnóstico, modelos de predicción, árbol de decisión.

Abstract

The purpose of this research is to develop predictive models for students' final passing grades in a high school course based on demographic data and the grades on a diagnostic exam at the beginning of the academic year. The target population was 80 high school students. From this group, five demographic variables and a diagnostic exam grade were collected. Naïve Bayes and decision tree machine learning algorithms were used to develop the predictive models, with the latter yielding the highest accuracy of 80%. Furthermore, to explore the potential use of this model, 10 students from a subsequent academic

year were predicted for a subsequent academic year, yielding an accuracy of 70%. This type of methodology allows for the use of information commonly collected by teachers to identify students with academic problems early in a school year.

Keywords: machine learning, diagnostic testing, prediction models, decision tree.

Fecha Recepción: Enero 2025

Fecha Aceptación: Mayo 2025

Introducción

Un problema de interés en las instituciones educativas es la implementación de estrategias de aprendizaje más eficaces y en nuevas modalidades de enseñanza como la virtualidad asincrónica y sincrónica (Parra *et al.*, 2024). Asimismo, se ha avanzado en la creación de bases de datos educativas que permiten su análisis para apoyar la toma de decisiones ante problemáticas como la reprobación y la deserción escolar. Estos análisis también posibilitan la identificación de hábitos de estudio que favorecen el rendimiento académico y contribuyen a la personalización del aprendizaje, adaptando los contenidos a las necesidades específicas de cada estudiante. Los avances tecnológicos también han incentivado la participación estudiantil a través del uso de diversas herramientas tecnológicas. Una ventaja adicional es la creación de modelos predictivos mediante técnicas de aprendizaje automático, construidos a partir de bases de datos estudiantiles, que permiten anticipar los porcentajes de aprobación y reprobación en una asignatura antes del inicio del curso. Asimismo, este tipo de modelo se ha empleado para predecir el rendimiento académico en distintos niveles educativos (Del Carpio, 2024). Es importante destacar que estas predicciones ofrecen una ventaja significativa al docente, al permitirle identificar tempranamente a los estudiantes con altas probabilidades de reprobación y proporcionarles el apoyo y los recursos necesarios según sus dificultades académicas.

En los últimos años se han desarrollado diversos estudios que abordan la predicción del rendimiento académico, así como la identificación de probabilidades de aprobación y reprobación en cursos específicos mediante algoritmos de aprendizaje automático, a partir de modelos predictivos construidos con datos académicos y personales de los estudiantes (Castrillón, Sarache y Ruiz, 2020; Contreras, Fuentes y Rodríguez, 2020; Yağcı, 2022).

También se han desarrollado modelos predictivos basados en técnicas de aprendizaje automático que estiman la reprobación o aprobación a partir de resultados obtenidos en actividades realizadas por estudiantes de diversos niveles educativos, evaluando su exactitud mediante validación cruzada. Morilla *et al.* (2020) utilizaron calificaciones de actividades académicas para predecir el desempeño en matemáticas mediante distintos algoritmos de aprendizaje automático. Por su parte, Contreras *et al.* (2024) emplearon calificaciones de cursos previos para predecir el rendimiento en exámenes de ingreso al bachillerato.

La persistencia de elevados índices de reprobación en asignaturas de nivel medio superior ha motivado nuevos estudios enfocados en identificar a los estudiantes con mayor probabilidad de reprobación (Castillo, Gamboa e Hidalgo, 2020; Fuentes y Rivera, 2021; Rojas, Figueroa y Gallegos, 2025), con el objetivo de que el profesorado pueda implementar estrategias de intervención antes de que ocurra la reprobación. Es por ello, que este trabajo se enfoca en el desarrollo de una metodología que permita la construcción de modelos predictivos con diferentes algoritmos de aprendizaje automático a través de datos

demográficos y una evaluación diagnóstica de los estudiantes con la finalidad de determinar la probabilidad de aprobación y reprobación en la calificación final de una asignatura. Este estudio, de carácter investigativo, plantea las siguientes preguntas de investigación: ¿Cómo se construyen modelos predictivos con algoritmos de aprendizaje automático a partir de datos demográficos y un examen diagnóstico? ¿Cómo se puede evaluar la exactitud de dichos modelos al predecir la aprobación y reprobación en una asignatura? Por ello, el objetivo de este estudio es desarrollar modelos predictivos, a partir de los resultados de un examen diagnóstico y ciertos datos demográficos de los estudiantes, que permitan estimar la probabilidad de aprobación o reprobación en una asignatura de nivel medio superior. El artículo describe la metodología de recolección y análisis de datos, los resultados obtenidos, la discusión de hallazgos, las conclusiones y futuras líneas de investigación para perfeccionar el modelo propuesto.

Metodología

La investigación se llevó a cabo utilizando un enfoque cuantitativo con el propósito de estimar el desempeño final en una asignatura de nivel medio superior. Esta metodología permite detectar, desde el inicio del curso, a los estudiantes con altas probabilidades de aprobar o reprobar mediante algoritmos de inteligencia artificial.

La población objetivo estuvo conformada por 80 estudiantes de un bachillerato tecnológico ubicado en la Ciudad de México, quienes cursaron la asignatura de Mediciones Eléctricas. La Tabla 1 presenta las variables recabadas y sus respectivas categorías. Cabe señalar que la información fue recabada con el consentimiento informado de los estudiantes, a quienes se les explicó que los datos serían utilizados exclusivamente con fines de investigación. Para preservar la confidencialidad, los nombres fueron reemplazados por códigos alfanuméricos, dado que los algoritmos empleados no requieren identificación personal. A través de una encuesta, se recolectaron datos personales de cada participante mostrados en los cinco primeros renglones de la Tabla 1. Asimismo, se aplicó una evaluación diagnóstica compuesta por preguntas de opción múltiple relacionadas con los contenidos del curso. De esta manera, se utiliza como variable predictora la calificación de una evaluación diagnóstica, la cual se registró si el estudiante aprobó, reprobó o no asistió mostrada en el renglón seis de la Tabla 1. La calificación final de la asignatura se clasifica en dos valores: aprobatoria y reprobatoria

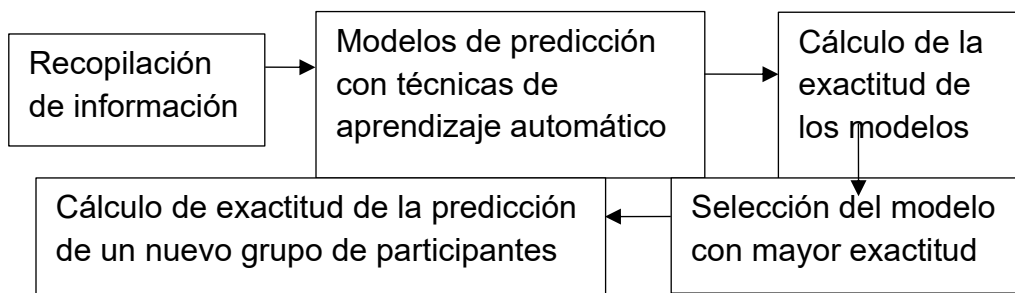
Tabla 1. Lista de datos recabados con valoraciones

Nivel académico del padre	Secundaria o menor, medio superior, universidad
Nivel académico de la madre	Secundaria o menor, medio superior, universidad
Promedio obtenido en el período anterior	Menor a 6, de 6 a 8, 8 u más
Numero de asignaturas reprobadas	0, 1, 2 ó mas
Periodicidad de estudio	Diario, semanal, mensual
Evaluación diagnostica	A (aprobatoria), R (reprobatoria), NA (no asistió)
Final	A (aprobatoria), R (reprobatoria)

Fuente: Elaboración propia

Una vez recopilada la información de los estudiantes se realizan modelos de predicción con distintas técnicas de aprendizaje automático y se calcula la exactitud de dichos modelos. Posteriormente, se selecciona el modelo con mayor exactitud para ser aplicado para predecir el rendimiento académico de otro grupo de estudiantes para mostrar la utilidad de dicho modelo. De manera resumida, la metodología empleada se muestra en la Figura 1.

Figura 1. Metodología empleada en la investigación.



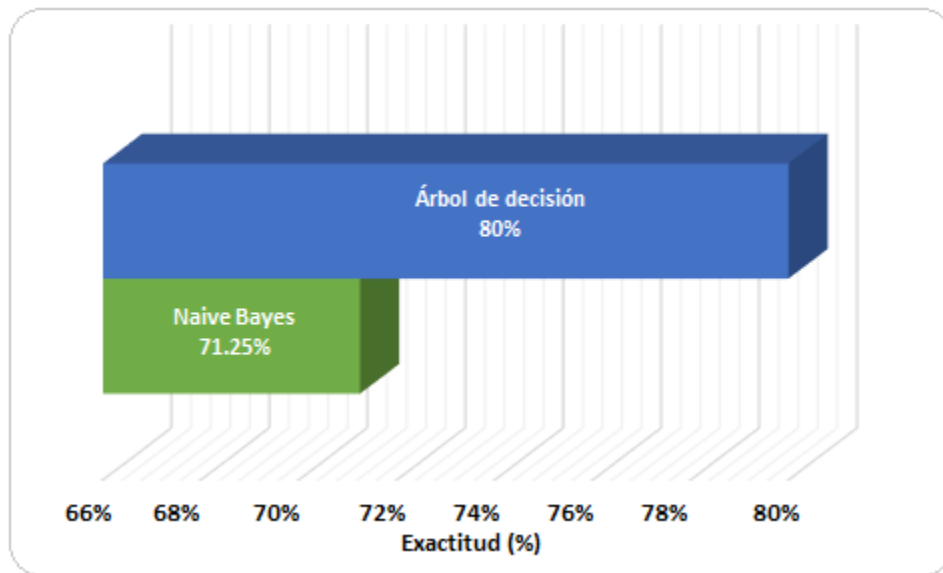
Fuente: Elaboración propia

En las siguientes secciones se describe más a detalle cada una de las etapas de la Figura 1.

Resultados

En esta etapa se construyeron modelos predictivos utilizando los datos recopilados previamente. Para ello, se aplicaron los algoritmos de aprendizaje automático Naïve Bayes (Chen et al., 2021; Angdresey, Sitanayah y Tangka, 2025; Islam et al., 2025) y árbol de decisión C4.5 (Caballero, López y Bautista, 2015; Lintang, Pandiangan y Hyronimus, 2022; Maulani et al., 2025), con el objetivo de predecir la probabilidad de aprobación o reprobación en la calificación final. Una vez contruidos los modelos de predicción, se evaluaron utilizando la métrica conocida como exactitud predictiva (Zapeta et al., 2022; Opitz, 2024). Esta se calcula dividiendo el número de predicciones correctas entre el total de predicciones realizadas. La exactitud se determinó mediante validación cruzada de diez segmentos (Ghasemzadeh, Hillman y Mehta, 2023). Este método consiste en dividir el conjunto de datos en diez partes: una se utiliza para validar las predicciones, mientras que las nueve restantes se emplean para construir el modelo. El proceso se repite alternando la parte utilizada para validación. Así, se repite el proceso alternando la parte destinada a validación, y se calcula la exactitud del modelo como el promedio de las tasas de acierto obtenidas en las diez iteraciones. La Figura 2 presenta una comparación entre las exactitudes obtenidas con los modelos predictivos desarrollados mediante los algoritmos *Naïve Bayes* y árbol de decisión.

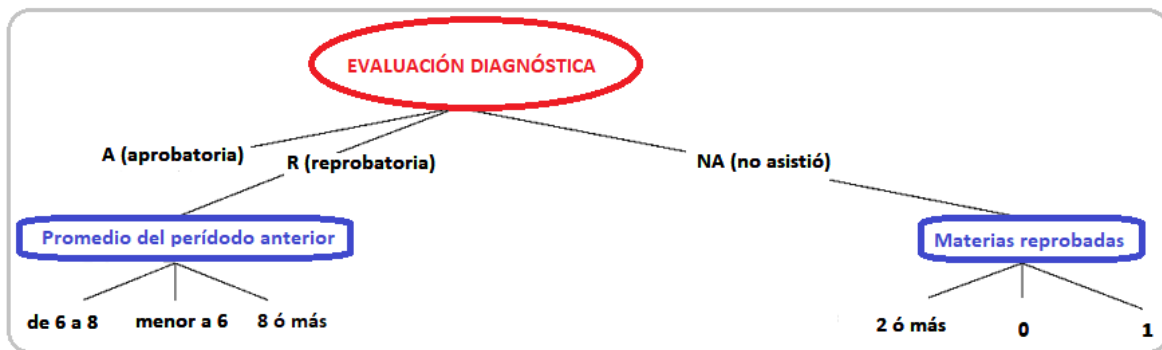
Figura 2. Exactitud de las técnicas de aprendizaje automático.



Fuente: Elaboración propia

La Figura 2 evidencia que el modelo basado en el algoritmo de árbol de decisión alcanzó una exactitud superior a la obtenida con el modelo construido mediante *Naïve Bayes*. El modelo de árbol de decisión empleado para obtener la exactitud de la figura anterior se presenta en la Figura 3.

Figura 3. Estructura del modelo predictivo con el algoritmo árbol de decisión.



Fuente: Elaboración propia

Del modelo de predicción del árbol de decisión se puede observar que los factores que contribuyen más en la predicción del rendimiento académico son el promedio del periodo anterior, las materias reprobadas y la evaluación diagnóstica, siendo esta última la variable más significativa del modelo ya que se encuentra en el nodo origen de la estructura del árbol de decisión.

Con el fin de demostrar la eficacia de los modelos, se realizaron predicciones sobre la calificación final de 10 estudiantes pertenecientes a un curso posterior de la misma asignatura. Estas predicciones se efectuaron utilizando el algoritmo de árbol de decisión, dado que obtuvo una mayor exactitud en la validación cruzada. Posteriormente, los resultados obtenidos se compararon con las calificaciones reales para evaluar la precisión del modelo.

En la Tabla 2 se observan las comparaciones de las predicciones, así como los resultados finales obtenidos de los 10 estudiantes.

Tabla 2. Predicciones de 10 estudiantes.

Estudiante	Predicción final de aprobación	Aprobación final con datos reales	Predicciones correctas
1	si	si	😊
2	si	si	😊
3	no	si	x
4	si	no	x
5	si	si	😊
6	si	si	😊
7	si	si	😊
8	si	si	😊
9	si	si	😊
10	no	si	x

Fuente: Elaboración propia.

En la Tabla 2 se observa que los estudiantes 3, 4 y 10 recibieron una predicción incorrecta. Esto representa una exactitud del 70 % en las predicciones realizadas por el modelo basado en el algoritmo de árbol de decisión.

Discusión

En la sección de resultados se describieron los modelos de predicción desarrollados mediante las técnicas de inteligencia artificial *Naïve Bayes* y árbol de decisión. Asimismo, se comparó la exactitud obtenida por ambos modelos, destacando que el árbol de decisión presentó una superioridad de aproximadamente 8 % respecto al algoritmo *Naïve Bayes*, resultado que coincide con hallazgos de otros estudios (Normah *et al.*, 2020; Alam, Alana y Julianne, 2023). En este modelo, la variable más significativa fue la evaluación diagnóstica, ubicada en el nodo raíz del árbol de decisión. Empleando este modelo se realizó la predicción de aprobación de 10 estudiantes de un curso posterior obteniendo una exactitud del 70 %, una de las razones de la disminución de la exactitud puede ser a los pocos datos sobre los cuales se aplicó el modelo, no obstante, se puede observar su utilidad en grupos pequeños, los cuales son comunes en grupos presenciales en instituciones educativas. En comparación con métodos tradicionales de identificación de estudiantes en riesgo como el promedio de materias previas o criterios de profesores, el método propuesto puede emplear éstos como variables predictoras para construir o mejorar el modelo de predicción. Cabe señalar que la información empleada en la presente investigación puede recabarse desde el primer día de clases mediante encuestas y la aplicación del examen diagnóstico, lo que permite a los docentes intervenir oportunamente con estudiantes que presenten riesgo académico. Cabe mencionar, que, en este tipo de modelos, es necesario proteger la privacidad de los participantes en el estudio para evitar divulgar información personal que puede causar perjuicios, por lo que, en esta investigación todos los nombres fueron sustituidos por códigos alfanuméricos desde el inicio del estudio.

Diversos estudios en la literatura han empleado técnicas de inteligencia artificial para predecir el desempeño académico de los estudiantes. Por ejemplo, Morilla et al. (2020) realizaron un estudio con 144 participantes y cinco variables predictoras, logrando exactitudes de 73.61 % y 72.22 % mediante validación cruzada con los algoritmos *Naïve Bayes* y árbol de decisión, respectivamente. En comparación, el presente estudio obtuvo una exactitud del 80 % con el modelo de árbol de decisión, a pesar de contar con una muestra menor y un número similar de variables.

Por otro lado, Contreras et al. (2024) emplearon un conjunto de datos mucho más amplio, con 7118 registros de estudiantes y cuatro variables predictoras, alcanzando una exactitud del 79.26 % con *Naïve Bayes* y 81.1 % con árbol de decisión. Aunque la exactitud obtenida en ese estudio fue ligeramente superior (1 %) a la del presente trabajo, la diferencia puede atribuirse al volumen considerablemente mayor de registros utilizados. Cabe destacar que, en esta investigación, también se realizaron predicciones con estudiantes de un curso posterior al periodo de recolección de datos, lo que aporta un valor añadido a la validación del modelo.

Conclusiones

Este estudio desarrolló una metodología orientada a determinar el desempeño académico de los estudiantes en una asignatura de nivel medio superior y a identificar a aquellos con alta probabilidad de aprobar o reprobar. Los modelos predictivos fueron elaborados mediante los algoritmos de aprendizaje automático *Naïve Bayes* y árbol de decisión. Para su construcción se utilizaron 80 registros de estudiantes de bachillerato, integrando variables demográficas y los resultados obtenidos en una evaluación diagnóstica aplicada al inicio del curso.

La predicción temprana de estudiantes con alta probabilidad de reprobación, a través de algoritmos de inteligencia artificial, ofrece al profesorado la posibilidad de intervenir oportunamente. Esto facilita la implementación de estrategias de enseñanza que pueden contribuir a reducir el índice de reprobación en la asignatura.

El modelo predictivo que alcanzó la mayor exactitud fue el basado en el algoritmo de árbol de decisión, con un 80 %, frente al 71.25 % obtenido por el modelo *Naïve Bayes*. Debido a este rendimiento superior, se eligió el modelo de árbol de decisión para realizar predicciones sobre 10 estudiantes de un curso posterior, alcanzando una exactitud del 70 %. Además, la evaluación diagnóstica se identificó como la variable predictora más influyente en dicho modelo. La metodología desarrollada en este estudio constituye una herramienta útil para que los docentes identifiquen estudiantes con alta probabilidad de reprobación. Esto permite diseñar e implementar estrategias de enseñanza-aprendizaje, así como fomentar hábitos de estudio, que contribuyan a reducir el número de estudiantes reprobados en distintas asignaturas y niveles educativos.

Futuras líneas de investigación

Aunque este estudio abordó un tema relevante como la predicción del rendimiento académico mediante modelos construidos con algoritmos de aprendizaje automático, aún existen áreas susceptibles de

exploración y mejora. Entre ellas se encuentran la posibilidad de emplear otros algoritmos de clasificación y aumentar el tamaño de la muestra para optimizar la precisión de los modelos.

Referencias

- Alam, A., Alana, D. A. F., & Juliane, C. (2023). Comparison of the C.45 and Naive Bayes algorithms to predict diabetes. *Sinkron: Jurnal Dan Penelitian Teknik Informatika*, 7(4), 2641–2650. <https://doi.org/10.33395/sinkron.v8i4.12998>
- Angdresey, A., Sitanayah, L., & Tangka, I. L. H. (2025). Sentiment analysis for political debates on YouTube comments using BERT labeling, random oversampling, and multinomial Naïve Bayes. *Journal of Computing Theories and Applications*, 2(3), 342–354. <https://doi.org/10.62411/jcta.11668>
- Caballero, L. A., López, A., & Bautista, J. (2015). Árbol de decisión C4.5 basado en entropía minoritaria para clasificación de conjuntos de datos no balanceados. *Research in Computing Science*, 92, 23–34. https://www.rcs.cic.ipn.mx/2015_92/Arbol%20de%20decision%20C4_5%20basado%20en%20entropia%20minoritaria%20para%20clasificacion%20de%20conjuntos%20de%20datos.pdf
- Castillo, M., Gamboa, R., & Hidalgo, R. (2020). Factors that influence student dropout and failing grades in a university mathematics course. *Uniciencia*, 34(1), 219–245. <https://doi.org/10.15359/ru.34-1.13>
- Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). Improved naive Bayes classification algorithm for traffic risk management. *EURASIP Journal on Advances in Signal Processing*, 2021(30). <https://doi.org/10.1186/s13634-021-00742-6>
- Contreras, L. E., Fuentes, H. J., & Rodríguez, J. I. (2020). Predicción del rendimiento académico como indicador de éxito/fracaso de los estudiantes de ingeniería, mediante aprendizaje automático. *Formación Universitaria*, 13(5), 233–246. <https://doi.org/10.4067/S0718-50062020000500233>
- Contreras, R., Cerón, C., Archundia, E., & Rivera, K. J. (2024). Técnicas de clasificación para predecir el desempeño de los estudiantes en pruebas estandarizadas. *EDUCATECONCIENCIA*, 32(3). <https://doi.org/10.58299/edutec.v32i2.793>
- Del Carpio, R. (2024). Predicción del rendimiento académico utilizando modelos de aprendizaje automático: Una revisión sistemática de la literatura. *593 Digital Publisher CEIT*, 9(6), 1038–1054. <https://doi.org/10.33386/593dp.2024.6.2797>
- Fuentes, J., & Rivera, M. E. (2021). Factores personales, contextuales y académicos asociados a la reprobación en educación superior: Una muestra nacional mexicana. *Revista Mexicana de Investigación Educativa*, 27(95), 1039–1062. <https://www.redalyc.org/journal/140/14074127002/html/>

- Ghasemzadeh, H., Hillman, R. E., & Mehta, D. D. (2023). Toward generalizable machine learning models in speech, language, and hearing sciences: Estimating sample size and reducing overfitting. *Journal of Speech, Language, and Hearing Research*, 67(3), 753–781. https://doi.org/10.1044/2023_JSLHR-23-00273
- Islam, A. R. M. T., Mia, M. U., Nova, N. A., Chakraborty, R., Khan, M. S. I., Ghose, B., Pal, S. C., Bari, A. B. M. M., Alam, E., Islam, M. K., Alshehri, M. A., Abdo, H. G., & Costache, R. (2025). Enhancing flood susceptibility mapping in Meghna River basin by introducing ensemble Naive Bayes with stacking algorithms. *Geomatics, Natural Hazards and Risk*, 16(1). <https://doi.org/10.1080/19475705.2025.2464049>
- Lintang, M., Pandiangan, N., & Hyronimus, D. (2022). Use of the C4.5 algorithm to analyze student interest in continuing to college. *SHS Web of Conferences*, 149. <https://doi.org/10.1051/shsconf/202214901048>
- Maulani, N., Farell, G., Hadi, A., Irfan, D., Alzyoud, M., & Nikolaevna, S. R. (2025). Automated academic supervisor allocation using the C4.5 decision tree algorithm: A scalable web-based solution. *Journal of Hypermedia & Technology-Enhanced Learning*, 3(1), 37–63. <https://doi.org/10.58536/j-hytel.165>
- Morilla, R. C., Omabe, R. D., Tolibas, C. J. S., Cornillez, E. E. C., & Treceñe, J. K. D. (2020). Application of machine learning algorithms in predicting the performance of students in mathematics in the modern world. *Journal of Educational Research and Technology Management*, 1(1), 49–57. <https://journals.indexcopernicus.com/api/file/viewByFileId/1449649>
- Normah, N., Yulianti, I., Novianti, D., Winnarto, M. N., Zumarniansyah, A., & Linawati, S. (2020). Comparison of classification C4.5 algorithms and Naïve Bayes classifier in determining merchant acceptance on sponsorship program. *Journal of Physics: Conference Series*, 1641(1), 012006. <https://doi.org/10.1088/1742-6596/1641/1/012006>
- Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *Transactions of the Association for Computational Linguistics*, 12, 820–836. https://doi.org/10.1162/tacl_a_00675
- Parra, M. E., Trujillo, J. C., Álvarez, D. R., Arias, A. S., & Santillán, E. (2024). El impacto de la inteligencia artificial en la educación. *Revista Científica Retos de la Ciencia*, 1(4), 169–181. <https://doi.org/10.53877/rc.8.19e.202409.14>
- Rojas, F., Figueroa, G. S., & Gallegos, N. (2025). Identificación temprana de estudiantes en riesgo de reprobación mediante la tutoría y su impacto académico. *Espacio I+D, Innovación más Desarrollo*, 14(40), 30–43. <https://www.espacioimasd.unach.mx/index.php/Inicio/article/view/440>
- Yağcı, M. (2022). Educational data mining: Prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, 9(11). <https://doi.org/10.1186/s40561-022-00192-z>

Zapeta, A., Galindo, G. A., Juan, H. J., & Martínez, M. (2022). Métricas de rendimiento para evaluar el aprendizaje automático en la clasificación de imágenes petroleras utilizando redes neuronales convolucionales. *Ciencia Latina Revista Científica Multidisciplinar*, 6(5), 4624–4637. https://doi.org/10.37811/cl_rcm.v6i5.3420