Predicción de la aprobación a través de datos personales de estudiantes de medio superior

Prediction of approval through personal data of high school students

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 objetivo de este trabajo es utilizar datos personales de los estudiantes para predecir la aprobación y reprobación de un curso determinado mediante modelos predictivos con diferentes técnicas de aprendizaje automático. En este estudio participaron 96 estudiantes de nivel medio superior. De cada uno de ellos se obtuvieron 7 datos personales con el propósito de elaborar modelos predictivos mediante el uso de técnicas de aprendizaje automático *Naïve Bayes* y *k* vecinos más cercanos. El porcentaje de exactitud de las predicciones más alto obtenido fue 73.95% con la técnica *k* vecinos más cercanos. Se mostró la potencial utilidad de este modelo al realizar la predicción de la evaluación final de 10 estudiantes de un curso posterior de la misma unidad de aprendizaje y se consiguió un porcentaje de exactitud de 70%. La metodología empleada puede ser un apoyo para los docentes con el propósito de intervenir con cierto tiempo de antelación en la recuperación de estudiantes con probabilidades de reprobación e incluso antes de que el curso inicie.

Palabras clave: exactitud, rendimiento académico, modelos predictivos, técnicas de aprendizaje automático.

Abstract

The objective of this work is to use students' personal data to predict the approval and failure of a given course using predictive models with different machine learning techniques. From each of them, 7 personal data were obtained with the purpose of developing predictive models through the use of Naïve Bayes and k nearest neighbors machine learning techniques. The highest prediction accuracy percentage obtained was 73.95% with the k nearest neighbors technique. The potential usefulness of this model was shown when predicting the final evaluation of 10 students from a subsequent course of the same learning unit and an accuracy percentage of 70% was achieved. The methodology used can be a support

for teachers with the purpose of intervening with a certain amount of time in advance in the recovery of students who are likely to fail and even before the course begins.

Keywords: accuracy, academic performance, predictive models, machine learning techniques.

Introducción

El avance tecnológico en años recientes ha permitido que el área educativa haya sufrido una evolución informática en cuanto almacenamiento y análisis de datos (Morales, Jiménez y Casas, 2023, Vega, Rueda y Niño, 2024, Bermúdez, 2024). En consecuencia, ha demostrado mejoras en los procesos de enseñanza aprendizaje y ser un gran apoyo para los docentes en la implementación de estrategias de aprendizaje, y principalmente, en la toma de decisiones a diferentes problemas académicos en instituciones educativas (Donoso y Calvopiña, 2024, Cruz, Arafet y Herrera, 2024). Una de las herramientas utilizadas para el análisis de datos educativos son los modelos predictivos los cuales nos permiten encontrar patrones asimismo comprender los procesos de aprendizaje para mejorar el desempeño académico de los estudiantes (Proaño, Ulloa, Hernández y Gunsha, 2023; Quijije y Maldonado, 2023; Álvarez, Ponce, Alban y Zambrano, 2024; Grasso, 2024, Barahona, Vega, Moyota y Porras, 2024).

Actualmente, existen diversos trabajos dedicados a la predicción del rendimiento académico implementando distintas técnicas de aprendizaje automático (Castillo y Martínez, 2023; Forero y Negre, 2024; Pérez y Quimbayo, 2024). En el estudio realizado por Contreras, Fuentes y Rodríguez, (2020) predijeron el desempeño académico de estudiantes universitarios a través de la selección de diversas variables implementando diferentes técnicas de aprendizaje automático. Por otra parte, Daza, Castro y Ávila (2024) propone una metodología para construir modelos predictivos de rendimiento académico en estudiantes de ingeniería basada en el análisis de datos académicos empleando tres técnicas de aprendizaje automático.

Guanin, Guaña y Casillas (2024) utilizan los conocimientos previos de estudiantes universitarios con distintas técnicas de aprendizaje automático para predecir el éxito académico. Asimismo, otros trabajos estiman la exactitud de la predicción con técnicas de aprendizaje automático utilizando métodos de validación cruzada (Fuentes y Rivas, 2021; Jahuey *et al.*, 2024; Villarreal, Ángeles, Marín y Cano, 2024).

Se puede notar que, a pesar de distintos trabajos para la mejora del rendimiento académico, sigue existiendo un problema de reprobación en estudiantes de nivel medio superior. Por lo que es necesario seguir desarrollando metodologías que ayuden a minimizar este tipo de problemas mediante el análisis de datos escolares con el objetivo de mejorar la enseñanza en las escuelas. En esta investigación se formulan los siguientes cuestionamientos: ¿Cómo se puede predecir la aprobación y reprobación de un curso determinado? y ¿Como se construyen los modelos con diferentes técnicas de aprendizaje automático para predecir la aprobación y reprobación de estudiantes? De tal forma, el objetivo de este estudio es realizar modelos con diferentes técnicas de aprendizaje automático que permitan predecir la aprobación y reprobación de estudiantes de nivel medio superior de un curso determinado a través de sus datos personales. En este trabajo se presenta la metodología de recopilación y análisis de datos, las

gráficas y resultados conseguidos, así como las discusiones y conclusiones. Al final se presenta el trabajo a futuro que se puede desarrollar a partir de esta investigación.

Metodología

La metodología usada en este trabajo es utilizar los datos personales de los estudiantes con el objetivo de predecir la aprobación y reprobación de un curso determinado. Esta metodología permite identificar a los estudiantes con altas probabilidades de reprobación en un curso con la finalidad de que el profesor pueda intervenir y tomar medidas pertinentes para reducir la reprobación.

En este estudio participaron 96 estudiantes de nivel medio superior que estuvieron en el curso de la asignatura Diseño Digital, donde se obtuvieron 7 datos personales de cada estudiante y su correspondiente evaluación final del curso. Estos valores son descritos en la tabla 1.

Tabla 1. Datos personales con sus respectivos valores.

Datos personales	Valores
Escolaridad del padre	Secundaria o menor, bachillerato, superior
Escolaridad de la madre	Secundaria o menor, bachillerato, superior
Ingreso familiar	Bajo (Menos de $7000), medio (entre $7000 y $15000), alto (más de $15000)
Promedio del semestre anterior	Menos de 6, entre 6 y 8, más de 8
Cantidad de materias reprobadas	0, 1, más de 2
Promedio del semestre anterior	Menos de 6, entre 6 y 8, más de 8
Frecuencia de estudio	Diario, semanalmente, mensualmente
Evaluación final	A (aprobación), R (reprobación)

Fuente: Elaboración propia

La evaluación final del curso se identifica con los valores de “A” aprobado y “R” reprobado. La información proporcionada por los estudiantes se recabo con su consentimiento para fines estadísticos de investigación.

Con estos datos se realiza una tabla con 96 registros (estudiantes) y 8 columnas (datos personales de los estudiantes), ilustrada en la tabla 2, la cual servirá para elaborar modelos predictivos con diferentes técnicas de aprendizaje automático que permitan identificar el modelo con mayor exactitud en la predicción del atributo “Aprobación del curso.

Tabla 2. Parte de la tabla de los datos personales de los estudiantes.

Registros	Escolaridad del padre	Escolaridad de la madre	Ingreso familiar	Promedio del semestre anterior
1	bachillerato	superior	medio (entre $7000 y $15000)	entre 6 y 8
2	secundaria o menor	bachillerato	medio (entre $7000 y $15000)	Menos de 6
3	bachillerato	superior	bajo (Menos de $7000)	entre 6 y 8
4	universidad	bachillerato	medio (entre $7000 y $15000)	Menos de 6
5	bachillerato	bachillerato	medio (entre $7000 y $15000)	más de 8
6	bachillerato	bachillerato	bajo (Menos de $7000)	Menos de 6
7	secundaria o menor	secundaria o menor	medio (entre $7000 y $15000)	más de 8

Fuente: Elaboración propia

Resultados

En esta parte se realizan modelos que permitan predecir a los alumnos con más probabilidad de aprobar mediante los datos proporcionados en el bloque anterior y las técnicas de aprendizaje automático *Naïve Bayes* (Daza *et al.*, 2024), *k* vecinos más cercanos (Quimbayo *et al.*, 2024). Una vez elaborados los modelos predictivos, estos se evalúan utilizando una métrica conocida como exactitud predictiva. Esta métrica se determina con el número de predicciones realizadas correctamente entre el total de predicciones. (Guanin *et al.*, 2024). La exactitud de las predicciones se determinada por medio de la validación cruzada con 10 particiones (Jahuey *et al.*, 2024), el cual se calcula al dividir el total de los datos en 10 particiones, donde se elige uno para elaborar las predicciones y los nueve restantes para elaborar el modelo, con ello, se determina la exactitud, posteriormente se repite el procedimiento cambiando de partición. La exactitud se calcula con el promedio de las exactitudes obtenidas de las 10 particiones. En la tabla 3 se muestran los resultados de las probabilidades de aprobación y reprobación en la evaluación final con los datos personales del modelo predictivo elaborado con la técnica *Naïve Bayes*.

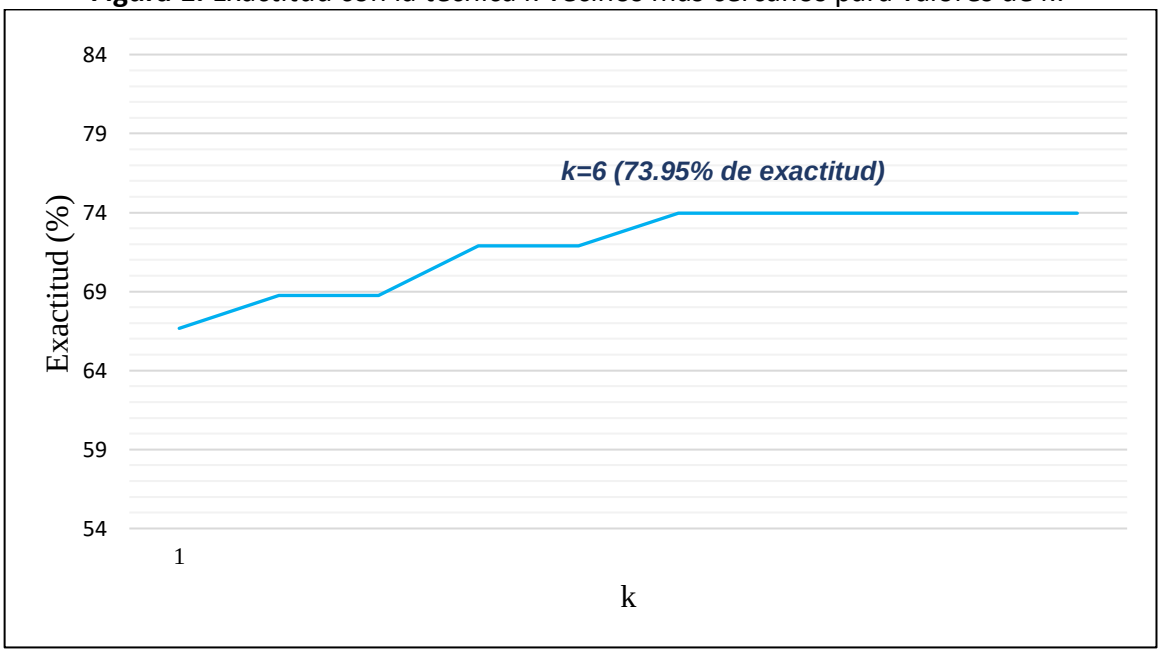
Tabla 3. Probabilidades de aprobación y reprobación con la técnica *Naïve Bayes*

Probabilidades	valor=A	valor=R
P(Aprobación en la evaluación final del curso=valor)	0.73	0.27
P(Escolaridad padre=bachillerato /evaluación final del curso=valor)	0.45	0.46
P(Escolaridad padre=secundaria o menor/evaluación final del curso=valor)	0.18	0.10
P(Escolaridad padre=universidad /evaluación final del curso=valor)	0.35	0.42
P(Escolaridad madre=bachillerato /evaluación final del curso=valor)	0.43	0.42
P(Escolaridad madre=secundaria o menor/evaluación final del curso=valor)	0.20	0.17
P(Escolaridad madre=universidad /evaluación final del curso =valor)	0.36	0.39
P(Ingreso familiar=bajo (Menos de $7000)/ evaluación final del curso=valor)	0.22	0.21
P(Ingreso familiar=medio (entre $7000 y $15000)/evaluación final del curso=valor)	0.48	0.46
P(Ingreso familiar=alto (más de $15000)/evaluación final del curso=valor)	0.28	0.32
P(Promedio del semestre anterior=menos de 6/evaluación final del curso=valor)	0.41	0.46
P(Promedio del semestre anterior=entre 6 y 8/evaluación final del curso=valor)	0.51	0.42
P(Promedio del semestre anterior=más de 8/evaluación final del curso=valor)	0.06	0.10
P(Cantidad de materias reprobadas=0/evaluación final del curso=valor)	0.31	0.5
P(Cantidad de materias reprobadas=1/evaluación final del curso=valor)	0.29	0.17
P(Cantidad de materias reprobadas=más de 2/evaluación final del curso=valor)	0.39	0.32
P(Promedio del semestre anterior=menos de 6/evaluación final del curso=valor)	0.48	0.53
P(Promedio del semestre anterior=entre 6 y 8/evaluación final del curso=valor)	0.47	0.39
P(Promedio del semestre anterior=más de 8/evaluación final del curso=valor)	0.04	0.07
P(Frecuencia de estudio=diario/evaluación final del curso=valor)	0.39	0.25
P(Frecuencia de estudio=semanalmente/evaluación final del curso=valor)	0.37	0.46
P(Frecuencia de estudio=mensualmente/evaluación final del curso=valor)	0.22	0.28

Fuente: Elaboración propia

Para poder usar la técnica de *k* vecinos más cercanos se debe elegir un valor para el parámetro *k*. Una forma de obtenerlo es eligiendo el valor que logre la máxima exactitud con este algoritmo. Así que, utilizado esta forma, se calcula la exactitud asignando distintos valores de *k* que varían desde uno hasta 96, para elegir el valor con mayor exactitud. En la figura 1 se muestra la gráfica de exactitud para valores de *k*.

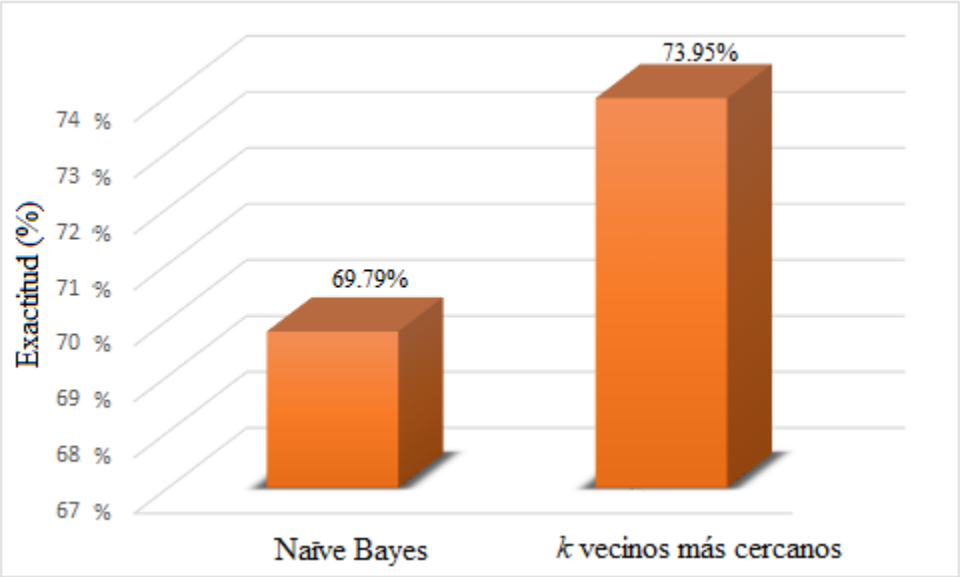
Figura 1. Exactitud con la técnica *k* vecinos más cercanos para valores de *k*.



Fuente: Elaboración propia

Como se indica en la gráfica anterior, a partir de *k*=6 se obtienen valores máximos de exactitud, el cual es de 73.95 %, por lo que se utilizará este valor. Al determinar el valor de *k* se realizan los modelos predictivos que permitan predecir a los alumnos con más probabilidad de aprobar en el curso. En la figura 2 se presenta la exactitud de las técnicas de aprendizaje automático *Naïve Bayes* y *k* vecinos más cercanos.

Figura 2. Exactitud de las técnicas de aprendizaje automático.

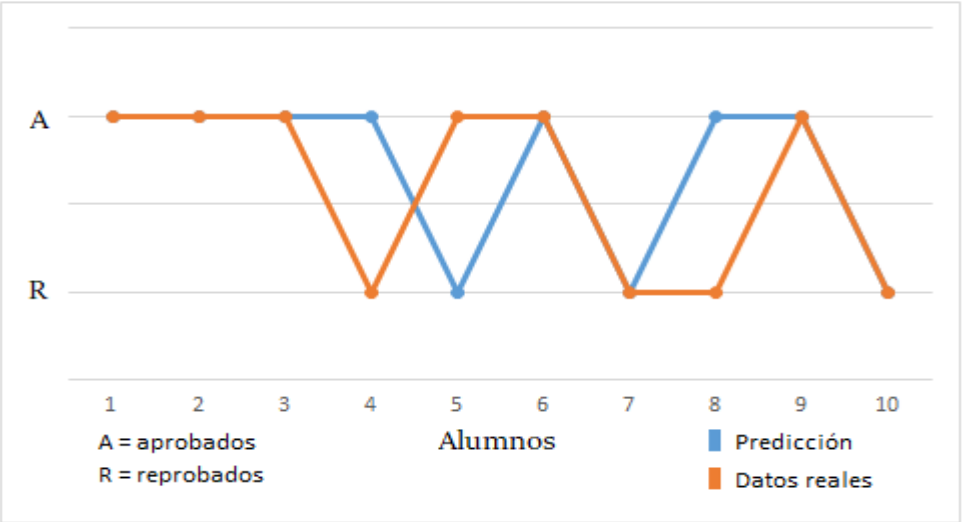


Fuente: Elaboración propia

Como se observa en la Figura 2 la exactitud con la técnica *k* vecinos más cercanos es la más alta con respecto a la técnica de aprendizaje automático *Naïve Bayes*.

Para mostrar la utilidad de este tipo de modelos, se realiza la predicción de la evaluación final de 10 estudiantes de un curso posterior de la misma unidad de aprendizaje. Estas predicciones se realizan mediante el modelo *k* vecinos más cercanos, debido a que obtuvo la mayor exactitud en la validación cruzada. Al realizarse todas las predicciones con cada dato de prueba se comparan las predicciones con los resultados obtenidos en la evaluación final del curso. En la figura 3 se determinan las predicciones, así como los resultados finales.

Figura 3. Predicciones de 10 estudiantes.



Fuente: Elaboración propia.

En la figura anterior se puede observar que los alumnos 4, 5 y 7 tuvieron una predicción incorrecta del total de los 10 alumnos, es decir, se obtuvo una exactitud de la predicción de 70 % con la técnica *k* vecinos más cercanos.

Discusión

En el bloque anterior se mostraron las exactitudes de los modelos predictivos con las técnicas de aprendizaje automático *Naïve Bayes* y *k* vecinos más cercanos de la aprobación y reprobación de un curso. La técnica *k* vecinos más cercanos presenta un 4.16% mayor exactitud con respecto a la técnica *Naïve Bayes.* Los resultados con estas técnicas de aprendizaje automático son parecidos a los observados en (Daza *et al.,* 2024). Cabe resaltar que los resultados se obtuvieron con datos de estudiantes que pueden ser recabados al comienzo del curso de manera sencilla de forma presencial o con algún formulario digital.

En la actualidad existen trabajos que predicen el rendimiento académico de los estudiantes mediante diferentes técnicas de aprendizaje automático. Daza *et al.* (2024) emplea 492 datos de muestra con los que se obtuvieron exactitudes del 59.2% con la técnica árbol de decisión J48, 64.5% con *k* vecinos más cercanos y 62.7% con *Naïve Bayes*. A diferencia de los resultados obtenidos en este trabajo podemos indicar que las predicciones con las técnicas de aprendizaje utilizadas fueron mayores (73.95% para *k* vecinos más cercanos y 69.79% para *Naïve Bayes*). Asimismo, Guanin *et al.* (2024) utiliza 6690 datos y una cantidad de variables de 21, a partir de estos datos, obtiene una exactitud del 78.22% con la técnica de árbol de decisión J48 y un 65.49% de exactitud con la técnica *k* vecinos más cercanos. Comparando estos resultados con la exactitud obtenida con la técnica *k* vecinos más cercanos en este trabajo podemos decir que fue superior a un 8.46%, utilizando menos variables. Del mismo modo, Contreras *et al.* (2020) realiza un análisis con 1620 datos obteniendo 66% con la técnica de árbol de decisión utilizando 7 variables. En contraste con los trabajos antes mencionados, en este estudio, se consigue una mayor exactitud de las predicciones, utilizando tan solo 7 variables correspondientes a 96 estudiantes. Además, en este trabajo se realizan predicciones de estudiantes de un curso posterior al semestre donde se recabaron los datos para crear el modelo de predicción.

Conclusiones

En este estudio se desarrolló una metodología que utiliza datos personales de estudiantes para realizar modelos con técnicas de aprendizaje automático, a fin de predecir la aprobación y reprobación de un curso determinado. Para la elaboración de los modelos de predicción se utilizaron las técnicas de aprendizaje automático *Naïve Bayes* y *k* vecinos más cercanos. Los registros utilizados para los modelos de predicción fueron 96 registros de estudiantes de nivel medio superior, así como 7 de sus datos personales. Una de las ventajas que se observó al identificar a los estudiantes con predicciones de reprobación fue que el profesor puede tomar las medidas pertinentes para reducir el número de estudiantes reprobados. Asimismo, el modelo con la técnica *k* vecinos más cercanos fue el que obtuvo mayor exactitud con un 73.95%, la cual fue un 4,16% mayor que la obtenida con la otra técnica de aprendizaje automático *Naïve Bayes*. Cabe destacar, que en este trabajo se comprobó la utilidad del modelo con la predicción de la evaluación final de 10 estudiantes de un curso posterior de la misma

asignatura obteniendo una exactitud de 70%. Esta metodología constituye una base de apoyo para los docentes a fin de poder intervenir con tiempo en la recuperación de estudiantes de cualquier nivel educativo con altas probabilidades de reprobación antes de que el curso inicie.

Futuras líneas de investigación

Cabe hacer mención que, pese a los progresos vistos en este estudio, aún existe mucho trabajo por hacer sobre esta línea de investigación. Un ejemplo podría ser la de aumentar la cantidad de registros de estudiantes, así como la cantidad de datos personales que intervengan en el desarrollo educativo de los estudiantes. De igual forma, cambiar las técnicas de aprendizaje automático y realizar una comparación entre ellas para saber la exactitud de las predicciones. Además, los modelos con las técnicas de aprendizaje automático podrían aplicarse a diferentes cursos de manera simultánea para realizar una comparación en cuanto a su efectividad. Otra alternativa es aplicar este tipo de metodología otras modalidades educativas como a cursos virtuales donde la recopilación de los datos sea automatizada.

Referencias

Álvarez, Y. M., Ponce, V. M., Alban, A. J. y Zambrano, S. L. (2024). Revisión de modelos estadísticos para pronosticar el desempeño académico en estudiantes universitarios. MQRInvestigar, 8(2), 3806–3823. https://doi.org/10.56048/MQR20225.8.2.2024.3806-3823

Barahona, A. D. M., Vega, C. P. A., Moyota, P. A. R. y Porras, R. L. I. (2024). Análisis de modelos estadísticos para predecir el éxito académico en estudiantes universitarios. MQRInvestigar, 8(2), 2951–2969. https://doi.org/10.56048/MQR20225.8.2.2024.2951-2969

Bermúdez, Q. N. C. (2024). Análisis de Big Data en las tecnologías de la información. Revista Científica Arbitrada de Investigación en Comunicación, Marketing Y Empresa REICOMUNICAR. ISSN 2737-6354., 7(14), 676-682. https://doi.org/10.46296/rc.v7i14.0290

Donoso V. M. E. y Calvopiña A. D. M. (2024). Gestión del conocimiento y desempeño académico en instituciones de educación fiscal de Ambato. Estudio de caso: Centro Integral y de Capacitación Blue Up. Arandu UTIC, 11(2), 455–471. https://doi.org/10.69639/arandu.v11i2.278

Fuentes, H. J. y Rivas, E. (2021). Análisis del rendimiento académico mediante técnicas de aprendizaje automático con métodos de ensamble. Revista Boletín Redipe, 10(13), 171-190. https://revista.redipe.org/index.php/1/article/view/1737

Castillo, D. y Martínez, J. J. (2023). Predicción del rendimiento académico en la UNADECA por medio de sistemas de clasificación. Unaciencia Revista De Estudios E Investigaciones, 16(31), 17–35. https://doi.org/10.35997/unaciencia.v16i31.738

Contreras, L. E., Fuentes, H. J. y Rodríguez, J. I. (2020). Academic performance prediction by machine learning as a success/failure indicator for engineering students. Formación Universitaria, 13(5), 233-246. https://doi.org/10.4067/S0718-50062020000500233.

Cruz, A. R., Arafet, Z. Y. y Herrera, L. M. (2024). Estrategias de aprendizaje y rendimiento académico en estudiantes de Ingeniería Informática del Instituto Superior Politécnico Benguela (Angola). MLS Educational Research (MLSER), 8(2). https://doi.org/10.29314/mlser.v9i1.2455

Daza, J., Castro, J. E. y Ávila, H. (2024). Optimizando el aprendizaje de los lenguajes de programación. Un enfoque basado en la analítica de datos para los estudiantes de Ingeniería de Sistemas en la Fundación Universitaria Los Libertadores. Perspectivas, 9(24), 234–256. https://doi.org/10.26620/uniminuto.perspectivas.9.24.2024.234-256

Forero, W. y Negre, F. (2024). Diseño y simulación de un modelo de predicción para la evaluación de la competencia digital docente usando técnicas de Machine Learning. Edutec, Revista Electrónica De Tecnología Educativa, (89), 18–43. https://doi.org/10.21556/edutec.2024.89.3201

Grasso, I. P. (2024). Rendimiento académico: modelo predictivo en universitarios según perspectiva temporal, estrategias de aprendizaje y personalidad. Revista Iberoamericana ConCiencia, 9(2), 1-13. https://doi.org/10.70298/ConCiencia.9-2.1

Guanin, J. H., Guaña, J. y Casillas, J. (2024). Predicting Academic Success of College Students Using Machine Learning Techniques. Data 9(4), 1-27. https://doi.org/10.3390/data9040060

Jahuey, F. J., Magaña, J. G., Segura, J. C., Martínez, J. C., Estrada, R. J. y Parra, G. M. (2024). Predicción genómica de peso vivo con dos métodos de validación cruzada en ganado bovino. Ecosistemas Y Recursos Agropecuarios, 11(1). https://doi.org/10.19136/era.a11n1.3817

Morales, S. R. E., Jiménez, A. J. de J. y Casas, F. A. A. (2023). Nivel de dominio de la competencia digital en el uso y alfabetización tecnológica en docentes de educación superior. Revista Científica Retos De La Ciencia, 7(16(e), 58–77. https://doi.org/10.53877/rc.7.16e.20230915.5

Pérez, E. L. y Quimbayo, J. A. (2024). Predicción del logro académico en clases espejo: análisis sociodemográfico y pedagógico con minería de datos. Revista Científica, 49(1), 79–98. https://doi.org/10.14483/23448350.21820

Proaño, M. P., Ulloa, C. S., Hernández, A. y Gunsha, M. A. (2023). Predicción del rendimiento académico mediante técnicas del análisis multivariado en la asignatura de ecuaciones diferenciales. Tesla Revista Científica, 3(1), e126. https://doi.org/10.55204/trc.v3i1.e126

Quijije, Q. H. B. y Maldonado, Z. K. (2023). Técnica de minería de datos para procesos educativos en estudiantes con necesidades educativas especiales basado en un modelo predictivo. Revista Científica Arbitrada Multidisciplinaria PENTACIENCIAS, 5(5), 205–217. https://doi.org/10.59169/pentaciencias.v5i5.730

Quimbayo, J., García, E., Alarcón, Á., Diaz, F. y Granada R. K. (2024). Predicción del aprendizaje sobre el correo electrónico en la población mayor: un enfoque con aprendizaje automático. CIE Academic Journal, 3(2), 2-53. https://doi.org/10.47300/2953-3015-v3i2-06

Vega, C. L. G., Rueda, V. G. y Niño, R. C. V. (2024). Análisis al Sistema de inventarios de una institución Educativa de la Ciudad de Cúcuta. Eco Matemático, 15(2), 6–12. https://doi.org/10.22463/17948231.4614

Villarreal, H., Ángeles, J., Marín, W. y Cano, J. (2024). Modelo de clasificación para la deserción estudiantil en las universidades públicas del Perú. Revista De Ciencias Sociales, 30(1), 452-469. https://dialnet.unirioja.es/servlet/articulo?codigo=9370050