revista de investigación educativa de la Rediech

• Volumen 13 • 2022 • e1426 • ISSN: 2448-8550 •

Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería

Predictive models of academic performance based on characteristics of engineering students

Andrés Rico Páez
Nora Diana Gaytán Ramírez

Resumen

El objetivo de esta investigación es proponer una metodología para construir modelos predictivos del rendimiento académico mediante características de estudiantes de ingeniería de nuestro país y comparar los modelos utilizando diferentes métricas de evaluación. En este estudio participaron 228 estudiantes que forman parte de una universidad pública en México. Los datos fueron recabados al inicio del curso y, por medio de tres técnicas de aprendizaje automático, se construyeron los modelos predictivos. Se analizaron las características de cada modelo y se consiguió una exactitud de las predicciones de alrededor de 65%. El modelo con la técnica *Naïve Bayes* resultó el más adecuado para la mayoría de las métricas empleadas en el estudio, principalmente, para identificar estudiantes en peligro de reprobación. Además, se encontró que el promedio actual fue la característica más significativa para la predicción del rendimiento académico de los estudiantes participantes en el estudio. La metodología desarrollada puede ser replicada para otros cursos y las características de los estudiantes pueden recabarse al inicio del curso o antes, permitiendo la posibilidad de realizar estrategias de intervención para estudiantes en peligro de reprobación.

Palabras clave: aprendizaje automático, árbol de decisión, exactitud, *k* vecinos más cercanos, *Naïve Bayes*.

Abstract

The aim of this research is to propose a methodology to build predictive models of academic performance through characteristics of engineering students in our country and to compare the models using different evaluation metrics. In this study, 228 students who are part of a public University in Mexico participated. Data were collected at the beginning of the course and, by means of three machine learning techniques, the predictive models were built. The characteristics of each model were analyzed and a prediction accuracy of around 65% was achieved. The model with the Naïve Bayes technique was the most suitable for most of the metrics used in the study, mainly to identify students in danger of failure. In addition, it was found that the current average was the most significant characteristic for the prediction of the academic performance of the students participating in the study. The methodology developed can be replicated for other courses and the characteristics of the students can be collected at the beginning of the course or before, allowing the possibility of carrying out intervention strategies for students in danger of failure.

Keywords: machine learning, decision tree, accuracy, *k* nearest neighbors, Naïve Bayes.

Recibido: octubre 11 de 2021 | *Aprobado:* marzo 26 de 2022 | *Publicado:* abril 20 de 2022
https://doi.org/10.33010/ie_rie_rediech.v13i0.1426

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Introducción

Existe un avance tecnológico sin precedentes en años recientes en diversas áreas (Sánchez, 2009), en particular, en el área educativa. En la literatura se ha observado que el uso de herramientas tecnológicas puede fomentar con prácticas adecuadas la enseñanza y aprendizaje (Torres y Cobo, 2017). Las tecnologías de la información y la comunicación impulsan las actividades cognitivas de los usuarios, específicamente, las herramientas tecnológicas que permiten almacenar, analizar e interpretar diversas características de estudiantes para predecir comportamientos académicos futuros y poder realizar intervenciones de manera oportuna en lugar de esperar hasta que el alumno repruebe alguna actividad y sea necesaria una recuperación académica que lleva más tiempo y es más costosa tanto para el alumno como para la institución educativa.

Uno de los objetivos del análisis de datos educativos es encontrar patrones y predicciones que permitan caracterizar el desarrollo académico de estudiantes, no obstante, se requiere la recopilación de datos de las características de los estudiantes teniendo en cuenta el contexto, para así conseguir una mayor comprensión de los resultados obtenidos. Algunas de estas características son factores socioeconómicos, datos familiares y escolares del estudiante.

Para el análisis de datos se emplean diferentes métodos, técnicas y algoritmos (Peña, 2014). La predicción del rendimiento académico se realiza con diversos propósitos, tales como detectar el riesgo de abandono o la posibilidad de deserción por parte de los estudiantes.

El rendimiento académico es un concepto complejo afectado por diversos agentes como características de la institución educativa, programas de estudio, profesores, estudiantes, entre otros. Montero et al. (2007) lo consideran como un indicador de calidad de la enseñanza que agrupa diversos factores, tales como pedagógicos, institucionales, entre otros. En carreras de ingeniería, cada institución acepta, casi inercialmente, los criterios a evaluar para dar una valoración del rendimiento académico

Andrés Rico Páez. Profesor Titular de la Escuela Superior de Ingeniería Mecánica y Eléctrica Unidad Zacatenco, Instituto Politécnico Nacional en Ciudad de México. Es ingeniero en Comunicaciones y Electrónica, maestro en Ciencias en la Especialidad de Ingeniería Eléctrica con opción en Comunicaciones y doctor en Tecnología Avanzada. Su área de investigación de interés incluye el análisis de datos educativos por medio de técnicas de inteligencia artificial, el uso de herramientas tecnológicas aplicadas a la educación y el estudio de sistemas de comunicaciones móviles inalámbricos. Ha publicado artículos de investigación en revistas nacionales e internacionales con arbitraje estricto. Es miembro del Sistema Nacional de Investigadores. Correo electrónico: aricop.ipn@gmail.com. ID: https://orcid.org/ 0000-0002-6450-318X.

Nora Diana Gaytán Ramírez. Profesora del Centro de Estudios Científicos y Tecnológicos no. 11 "Wilfrido Massieu" del Instituto Politécnico Nacional en Ciudad de México. Es ingeniera en Comunicaciones y Electrónica, maestra en Tecnología Avanzada y doctora en Tecnología Avanzada. Cuenta con varios cursos de actualización docente. Su línea de investigación son los sistemas tutores inteligentes aplicados a la educación mediante inteligencia artificial. Ha publicado artículos de investigación en revistas con arbitraje estricto. Correo electrónico: nora_diana@hotmail.com. ID: https://orcid.org/0000-0002-5159-9194.

(Ridgell y Lounsbury, 2004). Es decir, la interacción de varios factores influye en el rendimiento académico de los estudiantes (Gómez et al., 2011). Por lo tanto, es de interés de instituciones educativas identificar los principales factores que influyen en el rendimiento académico para realizar las acciones necesarias para seleccionar de mejor manera dichos factores en beneficio de los estudiantes (Mendoza y Herrera, 2013).

El rendimiento académico puede medirse en diferentes fases del proceso de formación académica del estudiante, así como también se pueden recopilar diversas variables o características del estudiante asociadas al rendimiento. Esta información puede almacenarse para ser analizada posteriormente para predecir el rendimiento académico del estudiante y tomar decisiones adecuadas de manera oportuna para mejorar los resultados del aprendizaje (Gutiérrez et al., 2021).

En años recientes, la deserción escolar y los altos índices de reprobación son las problemáticas más importantes en instituciones educativas, provocando bajos índices de eficiencia terminal (Martínez et al., 2013; Timarán et al., 2013). Una de las razones para la deserción escolar es que los estudiantes tienen un bajo rendimiento académico en una o varias asignaturas, lo que provoca que con mayor probabilidad reprueben hasta terminar con todas sus oportunidades para aprobar sus asignaturas, dando por resultado la deserción de la escuela. De esta manera, la predicción del rendimiento académico y los factores que contribuyen han sido estudiados por investigadores, principalmente, en el nivel superior, y son de gran interés para la sociedad en general (Tapasco et al., 2020).

Para realizar la predicción se han analizado datos para construir modelos predictivos a partir de técnicas de aprendizaje automático debido a que las herramientas estadísticas clásicas pueden no funcionar adecuadamente con grandes cantidades de datos y con varias características de estudiantes (Romero y Ventura, 2012). Estas técnicas de aprendizaje automático se centran en el uso y manejo de datos para obtener resultados representados como decisiones y son útiles para desarrollar modelos de predicción. Típicamente estas técnicas se utilizan en áreas de tipo comercial o empresarial (Han, 2012), sin embargo, se han comenzado a emplear en el diseño de modelos predictivos del rendimiento académico a partir de factores o características de estudiantes (Romero y Ventura, 2010).

Algunos trabajos que utilizan técnicas de aprendizaje automático en la creación de modelos predictivos: Salal et al. (2019), los cuales construyeron modelos de predicción para predecir el rendimiento académico de estudiantes con algoritmos como *Naïve Bayes*, árbol de decisión, entre otros. Este estudio muestra como ciertas características tienen influencia en el desempeño de estudiantes. Contreras et al. (2020) utilizaron modelos basados en algoritmos de aprendizaje automático para establecer que alumnos de ingeniería interrumpían o continuaban sus estudios. Castrillón et al. (2020) realizaron un estudio en el que se predice el rendimiento académico de estu-

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

diantes de nivel superior utilizando técnicas de árboles de decisión. Recientemente, los factores asociados al rendimiento académico de estudiantes universitarios han sido estudiados mediante regresiones lineales (Gutiérrez et al., 2021).

En la literatura revisada se ha encontrado que las instituciones educativas de nivel superior buscan mejorar el desempeño académico de estudiantes mediante la mejora de la calidad educativa (Kumar y Chadha, 2011). La reprobación en carreras de ingeniería puede ser disminuida si se tiene información oportuna de los estudiantes en peligro de reprobación, y en general, de su desempeño académico. Es por esta razón que se han utilizado técnicas de aprendizaje automático para la creación de modelos predictivos empleados en la predicción del rendimiento académico con el fin de identificar los factores que influyen en el proceso de enseñanza y aprendizaje; no obstante, el uso de técnicas de aprendizaje automático para este propósito es reciente en países de Latinoamérica (Estrada et al., 2016).

En México existe poco desarrollo en la construcción de este tipo de modelos con técnicas de aprendizaje automático debido, principalmente, al desconocimiento de la metodología para llevarlo a cabo. Esto ha ocasionado un cierto rezago en comparación con otros países en cuanto al análisis de datos de estudiantes para lograr beneficios potenciales para la mejora de la calidad educativa.

En este trabajo se plantean las siguientes preguntas de investigación: ¿Cómo desarrollar una metodología para construir modelos predictivos del rendimiento académico por medio de características al inicio de un curso de estudiantes de ingeniería de nuestro país? y ¿Cómo comparar los modelos de predicción del rendimiento académico mediante métricas de evaluación? El objetivo de esta investigación es proponer una metodología para desarrollar modelos predictivos del rendimiento académico a partir de características iniciales de estudiantes de ingeniería de nuestro país y comparar los modelos con distintas métricas de evaluación.

Metodología

En la literatura revisada se observó que existen pocos trabajos en México que utilizan técnicas de aprendizaje automático para el diseño de modelos predictivos del rendimiento académico, a pesar del beneficio potencial que pueden tener en el desempeño académico de estudiantes, específicamente, la predicción del rendimiento académico ofrece la oportunidad de elaborar planes de prevención de reprobación estudiantil mediante la realización de estrategias de intervención en lugar de estrategias de recuperación académica. Es decir, estos modelos permiten a los profesores e instituciones educativas realizar intervenciones desde el principio del curso y no al final cuando es demasiado tarde para realizar alguna acción para evitar la reprobación del estudiante.

En este artículo se propone una metodología basada en recopilar información de estudiantes al inicio, o incluso antes, de un curso de una asignatura, como se muestra

en la Figura 1. Posteriormente, se construyen modelos predictivos que permiten predecir el rendimiento académico que obtendrán futuros estudiantes al finalizar el curso. Finalmente, se evalúan los modelos y se comparan con base en métricas representativas.

Figura 1
Metodología propuesta

Fuente: Construcción personal.

En este trabajo, la construcción y evaluación de los modelos predictivos se hizo con apoyo del *software* de licencia libre Weka (*Waikato Environment for Knowledge Analysis*), el cual contiene varias técnicas de aprendizaje automático y una interfaz para poder visualizar los datos de diferentes formas (Witten et al., 2005). Este *software* permite introducir datos en archivos que tienen varios registros con un cierto número de atributos (Díaz et al., 2021).

Es importante señalar que la información recopilada se realizó al inicio del curso, no obstante, se puede hacer antes de que inicie el curso, por lo que permite predecir la aprobación del estudiante antes de que inicie, dando oportunidad a profesores e instituciones educativas de tener un tiempo razonable para planear o realizar algún tipo de intervención para disminuir los índices de reprobación.

Recopilación de datos

La muestra de datos corresponde a estudiantes inscritos en un curso de ingeniería de una universidad pública de la Ciudad de México. La información recopilada fue la escolaridad de los padres, ingreso familiar, promedio en la escuela anterior, materias reprobadas, promedio actual, preferencia de estudio y de actividades, frecuencia de estudio y su calificación en el curso. Este tipo de características fueron seleccionados debido a que son más simples de recolectar y han sido utilizados en otros estudios de predicción del rendimiento académico (Shahiri et al., 2015). En esta investigación participaron 228 estudiantes, y para cada uno de ellos se recopilaron las características mostradas en la Tabla 1.

La información referente a la aprobación y reprobación de los estudiantes fue proporcionada por los docentes del curso. Las otras características fueron recopiladas mediante una encuesta a los estudiantes en la cual se indicó que la información recabada era para fines estadísticos y de investigación.

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Tabla 1
Atributos con sus respectivos valores posibles

Atributos	Valores posibles
Escolaridad del padre	Primaria y secundaria Media superior Superior o mayor
Escolaridad de la madre	Primaria y secundaria Media superior Superior o mayor
Ingreso familiar	< $5000 $5000 - $10000 > $10000
Promedio de media superior	< 7.5 7.5 – 8.5 > 8.5
Materias reprobadas	0 1 > 2
Promedio actual	< 7.5 7.5 – 8.5 > 8.5
Preferencia de estudio	Solo Con otra persona En grupo
Preferencia al realizar actividades académicas	Solo Con otra persona En grupo
Frecuencia de estudio	Diario Una semana antes del examen Un día antes del examen
Curso	Aprueba Reprueba

Fuente: Construcción personal.

Con los datos recopilados se forma una tabla de 228 renglones (registros de estudiantes) y 10 columnas (atributos o características del estudiante), y servirá para construir los modelos predictivos del rendimiento académico.

Construcción de modelos predictivos

Las técnicas de aprendizaje automático permiten construir modelos específicos a partir de un conjunto de registros para un resultado concreto. Es decir, para construir un

modelo de este tipo se requiere una cierta información (también llamada "datos de entrenamiento") y una técnica de aprendizaje automático. Las técnicas de aprendizaje automático implementan diferentes mecanismos para la construcción de modelos predictivos que pueden consistir en algoritmos, ecuaciones, algún tipo de estructura, entre otros. En esta investigación se utilizan las técnicas *Naïve Bayes*, *k* vecinos más cercanos y árbol de decisión C4.5, las cuales han sido utilizadas en trabajos similares (Salal et al., 2019; Contreras et al., 2020; Castrillón et al., 2020).

Naïve Bayes

La técnica de aprendizaje automático *Naïve Bayes* (Witten et al., 2005) emplea un conjunto de datos organizados en columnas o atributos representados como $\{A_1, ..., A_n\}$ y un atributo de clase o simplemente clase representada como C_i que forma parte de un conjunto $\Omega_C = \{C_1, ..., C_k\}$. Esta técnica utiliza la probabilidad de C_i dado un conjunto de atributos obtenida con el teorema de Bayes de la siguiente forma:

$$P(C_i \mid A_1,...,A_n)=[P(A_1, ..., A_n \mid C_i)P(C_i)]/P(A_1, ..., A_n)$$

El teorema de Bayes se emplea para calcular el valor de la clase con mayor probabilidad, en particular, se emplea la máxima probabilidad *a posteriori* (MPA); para obtener el valor de la clase más probable sería:

$$\begin{aligned} C_{MPA} &= \arg\max_{C_i \in \Omega_C} P(C_i \mid A_1,\ldots,A_n) \\ &= \arg\max_{C_i \in \Omega_C} [P(A_1,\ldots,A_n \mid C_i)P(C_i)]/P(A_1,\ldots,A_n) \\ &= \arg\max_{C_i \in \Omega_C} P(A_1,\ldots,A_n \mid C_i)P(C_i) \end{aligned}$$

La clase obtenida con base en la suposición de independencia con esta técnica es:

$$C_{MAP} = \arg\max_{C_i \in \Omega_C} P(C_i) \prod_{j=1}^{n} P(A_j \mid C_i)$$

Las probabilidades $P(C_i)$ se obtienen dividiendo la cantidad de registros de la clase C_i entre el total de los datos y las probabilidades $P(A_j \mid C_i)$ se calculan dividiendo la cantidad de registros de los casos favorables entre el total de los casos (Hernández et al., 2004). El modelo predictivo se construye por medio del cálculo de estas probabilidades, en donde los atributos $\{A_1, ..., A_n\}$ son los que están representados en la Tabla 1, y del atributo "Curso", el cual representa la clase C_i. En la Tabla 2 se presenta el modelo predictivo con las probabilidades estimadas con la técnica *Naïve Bayes*.

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Tabla 2
Modelo predictivo con la técnica Naïve Bayes

Probabilidades	clase = Aprueba	clase = Reprueba
P(Curso=clase)	0.59	0.41
P(Escolaridad del padre=Primaria y secundaria/Curso=clase)	0.34	0.3
P(Escolaridad del padre=Media superior/Curso=clase)	0.38	0.35
P(Escolaridad del padre=Superior o mayor/Curso=clase)	0.28	0.35
P(Escolaridad del madre=Primaria y secundaria/Curso=clase)	0.41	0.36
P(Escolaridad del madre=Media superior/Curso=clase)	0.4	0.42
P(Escolaridad del madre=Superior o mayor/Curso=clase)	0.19	0.22
P(Ingreso familiar=< $5000/Curso=clase)	0.35	0.33
P(Ingreso familiar=$5000 - $10000/Curso=clase)	0.5	0.52
P(Ingreso familiar=> $10000/Curso=clase)	0.15	0.15
P(Promedio de media superior=< 7.5/Curso=clase)	0.28	0.39
P(Promedio de media superior=7.5 – 8.5/Curso=clase)	0.55	0.45
P(Promedio de media superior=> 8.5/Curso=clase)	0.17	0.16
P(Materias reprobadas=0/Curso=clase)	0.51	0.25
P(Materias reprobadas=1/Curso=clase)	0.34	0.34
P(Materias reprobadas=>2/Curso=clase)	0.15	0.41
P(Promedio actual=< 7.5/Curso=clase)	0.45	0.76
P(Promedio actual=7.5 – 8.5/Curso=clase)	0.42	0.19
P(Promedio actual=> 8.5/Curso=clase)	0.13	0.05
P(Preferencia de estudio=Solo/Curso=clase)	0.49	0.49
P(Preferencia de estudio=Con otra persona/Curso=clase)	0.25	0.27
P(Preferencia de estudio=En grupo/Curso=clase)	0.26	0.24
P(Preferencia al realizar actividades académicas=Solo/Curso=clase)	0.21	0.27
P(Preferencia al realizar actividades académicas=Con otra persona/Curso=clase)	0.39	0.33
P(Preferencia al realizar actividades académicas=En grupo/Curso=clase)	0.4	0.4
P(Frecuencia de estudio=Diario/Curso=clase)	0.3	0.32
P(Frecuencia de estudio=Una semana antes del examen/Curso=clase)	0.5	0.55
P(Frecuencia de estudio=Un día antes del examen/Curso=clase)	0.2	0.13

Fuente: Construcción personal.

En la Tabla 2, las probabilidades $P(C_i)$ representan el primer renglón y los demás renglones representan las probabilidades condicionales $P(A_j \mid C_i)$. Para predecir un nuevo registro se asocian sus atributos con las probabilidades de la tabla 2 y se multiplican como indica la fórmula anterior para predecir si el estudiante aprueba o reprueba el curso.

De esta manera, las probabilidades condicionales se multiplican para predecir el rendimiento académico. Por lo tanto, las probabilidades que tienen valores mayores son las que más influyen en la aprobación y en la reprobación. De la Tabla 2 se puede observar que, para este modelo y tomando en cuenta solo las probabilidades condicionales, la característica que más influye para la aprobación es que el estudiante haya tenido en el nivel medio superior un promedio entre 7.5 y 8.5 debido a que tiene la mayor probabilidad condicional, la cual es de 0.55. No obstante, esta probabilidad no se aleja mucho de las demás, tal es el caso de la probabilidad de que el estudiante haya reprobado cero materias actualmente, la cual es de 0.5. Sin embargo, se puede observar que para reprobación la mayor característica que influye es que el promedio actual del estudiante sea menor a 7.5 debido a que su probabilidad condicional es de 0.76. De hecho, esta es la mayor probabilidad de todo el modelo predictivo, por lo que los estudiantes que tengan un promedio menor a 7.5 es probable que reprueben.

k vecinos más cercanos

La técnica de aprendizaje automático *k* vecinos más cercanos (Cover y Hart, 1967; Hernández et al., 2004) clasifica el registro nuevo mediante la asignación de la clase más frecuente de entre los *k* registros más cercanos conforme a una cierta métrica, típicamente, la distancia euclidiana. La distancia euclidiana entre registros X_i y X_j representados con vectores $[x_{i1}, x_{i2}, ..., x_{in}]$ y $[x_{j1}, x_{j2}, ..., x_{jn}]$ se calcula como:

$$d(X_i, X_j) = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2}$$

Para cada registro de un estudiante a predecir su rendimiento académico se calcula su distancia a cada uno de los otros registros de estudiantes con la fórmula anterior. Posteriormente, se seleccionan los *k* registros más cercanos al registro a predecir. Por último, se le asigna al registro a predecir la clase que más se repita entre los *k* registros elegidos.

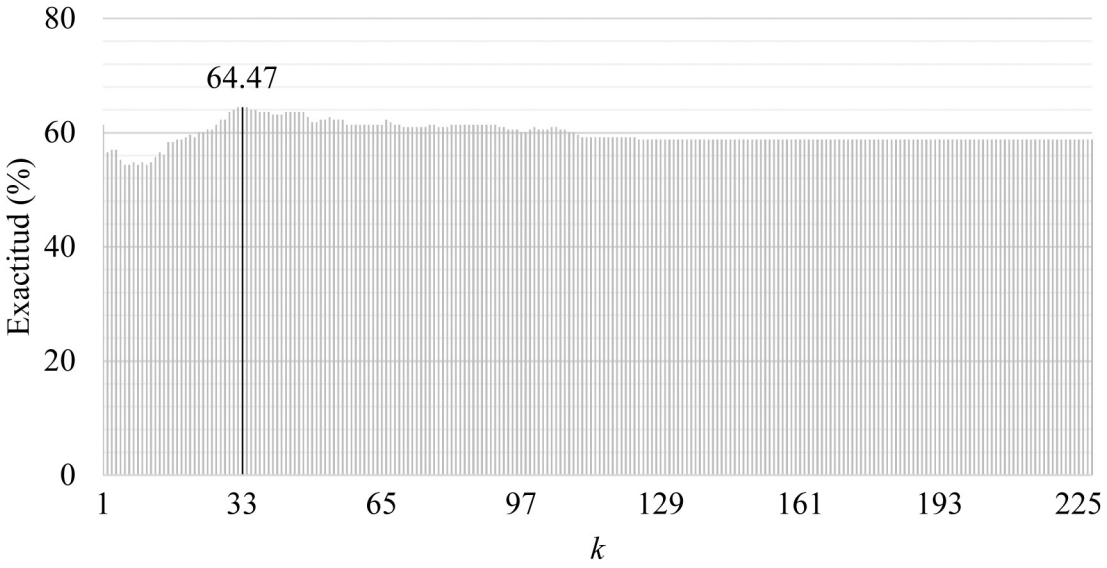
En el modelo predictivo con la técnica *k* vecinos más cercanos es necesario elegir el valor de *k*. Una manera de seleccionar este valor es eligiendo aquel valor que consiga un mayor valor en la exactitud de las predicciones. La exactitud se define como las predicciones correctas entre el total de las predicciones (Durairaj y Vijitha, 2014). En este artículo, la exactitud es obtenida por medio de un análisis de datos conocido como "validación cruzada", la cual consiste en dividir aleatoriamente los datos de entrenamiento en una cantidad fija de grupos o particiones. Se reserva una partición para realizar las predicciones con la técnica de aprendizaje automático y con las restantes se construye el modelo predictivo, este proceso se repite dejando

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

una partición diferente para hacer las predicciones. En este trabajo se utiliza una validación cruzada con 10 particiones debido a que es un valor que ha sido utilizado en otros trabajos similares (Márquez et al., 2012; Mueen et al., 2016).

Con este procedimiento se calcula la exactitud para distintos valores de *k* desde 1 hasta el número total de registros (228 estudiantes), es decir, *k* = 1, 2, 3, …, 228. En la Figura 2 se presenta la gráfica de la exactitud con validación cruzada con la técnica *k* vecinos más cercanos y se puede observar que la exactitud mayor se obtiene para *k* = 33 y es de 64.47%, de esta manera, este es el valor de *k* utilizado en la construcción del modelo predictivo.

Figura 2
Exactitud con la técnica k vecinos más cercanos con validación cruzada



Fuente: Construcción personal.

Árbol de decisión C4.5

Un árbol de decisión es un grupo de reglas estructuradas de forma jerárquica y siguiendo estas reglas desde la raíz del árbol hasta alguna de las hojas se llega a una decisión final, es decir, se puede analizar un problema y siguiendo la estructura del árbol de manera adecuada se llega a una sola decisión. Los elementos de un árbol se conocen como nodos y cada uno representa una regla, de tal manera que siguiendo un nodo a otro nodo descendiente se llega, eventualmente, a un nodo terminal donde se realiza la predicción.

La técnica de aprendizaje automático árbol de decisión C4.5 fue presentada por Quinlan (1993). Esta técnica representa a A_j como el j-ésimo atributo y S_v como el subconjunto de ejemplos de un conjunto S en donde el atributo A_j tiene el valor de v, y la entropía del conjunto S se calcula como:

$$H(S) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

En donde P_i es la probabilidad de la clase i dentro del conjunto de datos. La ganancia de información de cada atributo A_j con respecto a la clase es:

$$\text{Ganancia}\left(S,A_j\right) = H(S) - \sum_{v \in \text{Valores}(A_j)} \frac{|S_v|}{|S|} H(S_v)$$

En donde los valores (A_j) es el conjunto de posibles valores del atributo A_j, $|S_v|$ es el número de ejemplos en S etiquetados con v, $|S|$ es el número total de ejemplos y $H(S_v)$ es la entropía de los ejemplos etiquetados con v. La información de la partición del conjunto S con respecto a los valores del atributo A_j se calcula como:

$$I\left(S,A_j\right) = -\sum_{v \in \text{Valores}(A_j)} \frac{|S_v|}{|S|} \log_2 \frac{|S_v|}{|S|}$$

La técnica de árbol de decisión C4.5 utiliza la proporción de la ganancia para seleccionar la partición, la cual se calcula como:

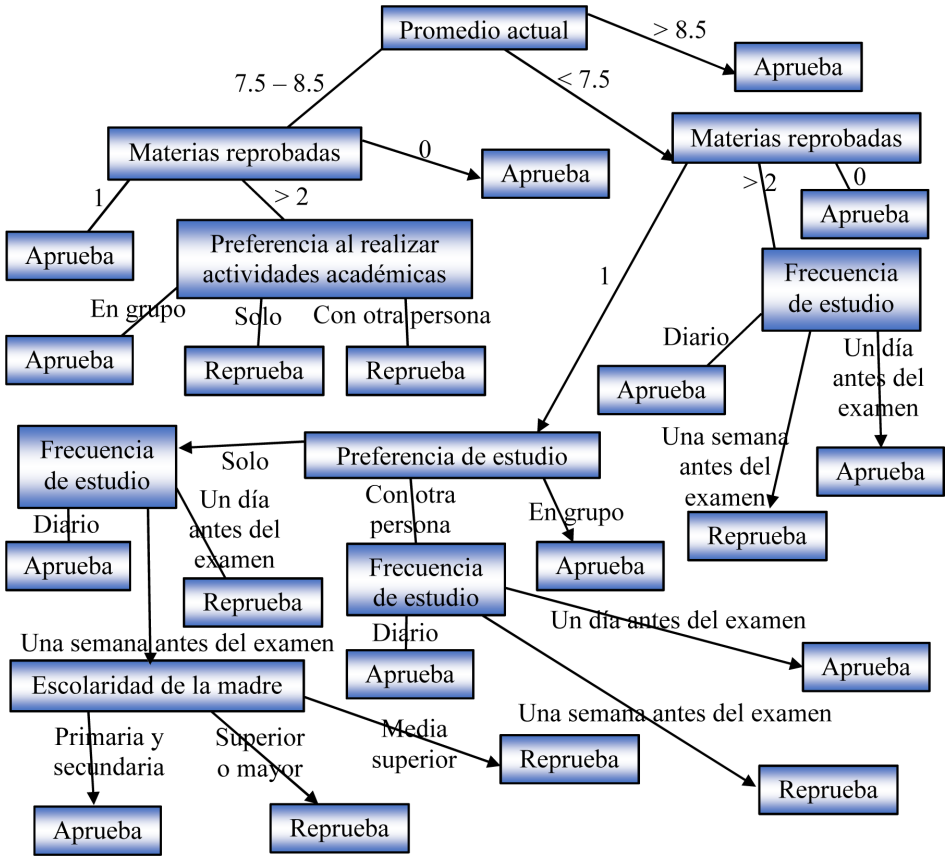
$$\text{Proporción de la ganancia}\left(S,A_j\right) = \frac{\text{Ganancia}\left(S,A_j\right)}{I\left(S,A_j\right)}$$

La técnica va calculando para cada nodo la proporción de la ganancia y eligiendo la partición con el valor mayor. Con esto va elaborando el árbol de decisiones hasta que llega a un nodo terminal, con lo que queda construido el modelo predictivo. Para predecir el resultado de un registro considera sus atributos y con base en ellos sigue las reglas del árbol desde nodo el raíz hasta un nodo terminal del árbol en el que se encuentra la predicción obtenida para ese registro. De esta manera, empleando los datos de entrenamiento y las fórmulas anteriores implementadas en el *software* Weka se construye el árbol de decisión mostrado en la Figura 3.

En la Figura 3 se observa que el nodo raíz del árbol es el atributo "Promedio actual". Este atributo representa el que más influye en las predicciones porque a partir de él se siguen distintas trayectorias para obtener la predicción. Además, se puede observar que el atributo "Promedio actual" también es el que más influye en el modelo con la técnica *Naïve Bayes*. Específicamente, con el modelo *Naïve Bayes*, los estudiantes que tengan un promedio menor a 7.5 es muy probable que reprueben, sin embargo, para el modelo de árbol de decisión se requiere, además de esta condición, de los valores de otros atributos para predecir que el estudiante reprobará. Una de

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Figura 3
Modelo predictivo con la técnica árbol de decisión C4.5



Fuente: Construcción personal.

las ramas más fáciles de visualizar es que el estudiante aprueba cuando su promedio actual sea mayor a 8.5 independientemente de los demás valores de sus atributos.

Resultados

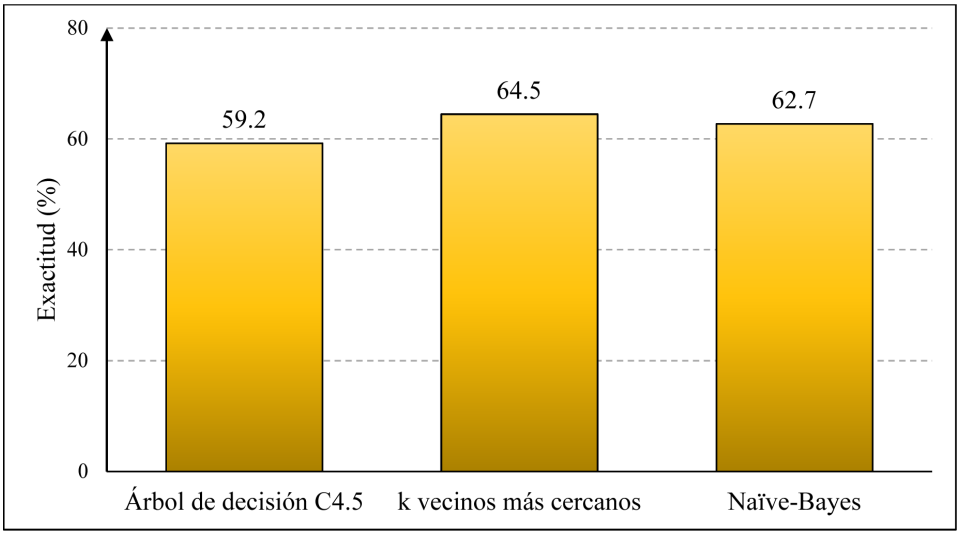
La construcción de modelos predictivos requiere evaluarlos, es decir, comprobar que arroje resultados suficientemente satisfactorios, por lo que se requiere el planteamiento de métricas de evaluación. En este trabajo se utiliza la exactitud, tasa de verdaderos positivos y la tasa de verdaderos negativos (Durairaj y Vijitha, 2014). La exactitud fue definida en la sección anterior como la cantidad de predicciones correctas con respecto al total de las predicciones. La tasa de verdaderos positivos es una métrica que se calcula dividiendo el número de registros predichos como positivos entre el total de registros positivos, en este trabajo, representaría las predicciones correctas de estudiantes aprobados con respecto al total de aprobados. De manera similar, la

tasa de verdaderos negativos es el número de registros predichos como reprobados con respecto al total de registros de estudiantes que reprobaron.

Para el cálculo de la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos de los modelos predictivos con las técnicas *Naïve Bayes*, *k* vecinos más cercanos y árbol de decisión C4.5 se utiliza la validación cruzada con 10 particiones (Mueen et al., 2016), es decir, se dividen aleatoriamente los datos de entrenamiento en diez particiones y con nueve particiones se construye el modelo predictivo para predecir la aprobación de la partición restante para calcular la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos; este proceso se repite dejando una partición diferente para hacer las predicciones y se toma el promedio de las métricas de evaluación. Este proceso se repite para cada una de las tres técnicas de aprendizaje automático. En las figuras 4, 5 y 6 se presenta la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos, respectivamente, de los modelos construidos con las técnicas de aprendizaje automático.

En la Figura 4 se observa que la exactitud es muy similar en los tres modelos, sin embargo, el modelo con la técnica *k* vecinos más cercanos es el que presenta mayor exactitud en las predicciones.

Figura 4
Exactitud de los modelos predictivos del rendimiento académico



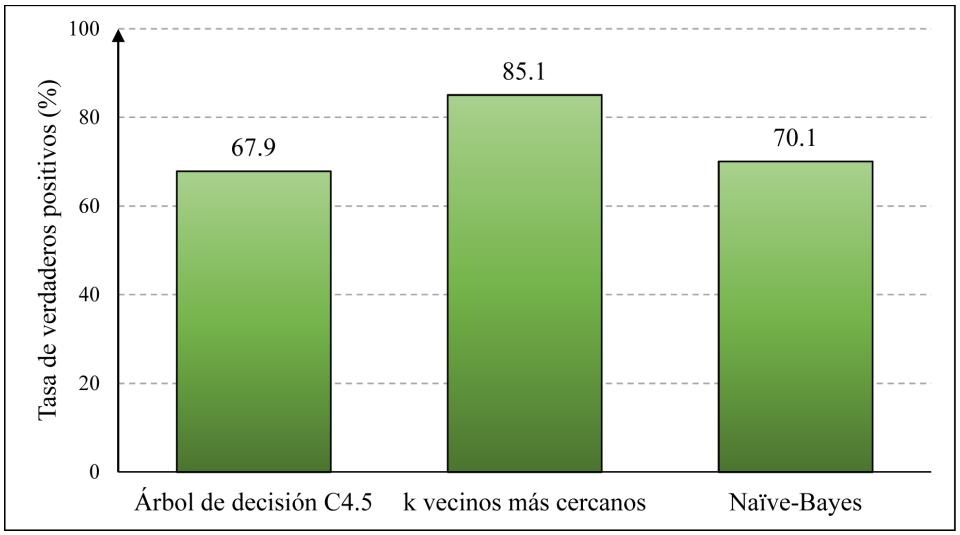
Fuente: Construcción personal.

En la Figura 5 se observa que el mayor valor de la tasa de verdaderos positivos se obtiene con la técnica *k* vecinos más cercanos, además se observan valores similares de esta métrica con las técnicas de árbol de decisión y *Naïve Bayes*.

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Figura 5

Tasa de verdaderos positivos de los modelos predictivos del rendimiento académico

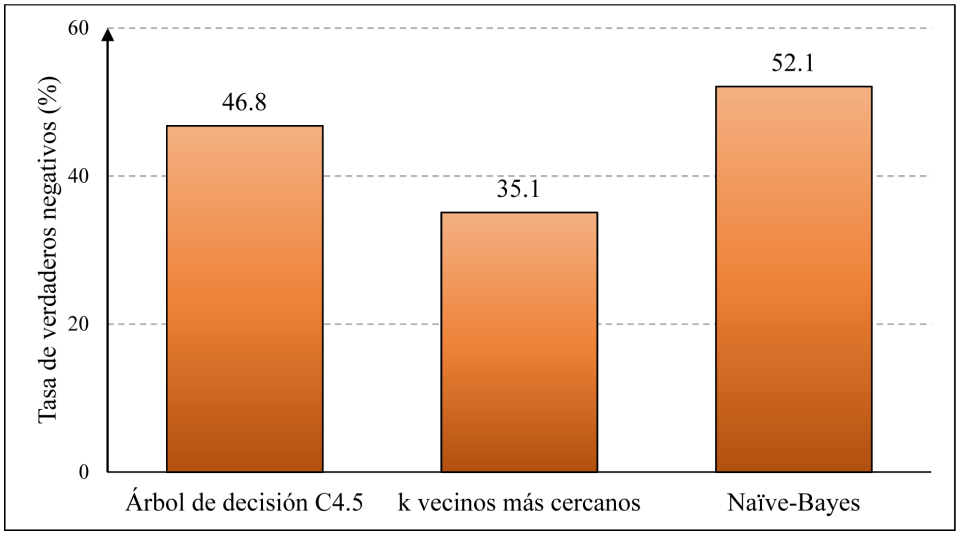


Fuente: Construcción personal.

En la Figura 6 se observa que la tasa de verdaderos negativos es mayor para el modelo predictivo con la técnica *Naïve Bayes* y el menor valor se obtiene con el modelo de *k* vecinos más cercanos, lo cual es opuesto a lo obtenido con las métricas de evaluación anteriores.

Figura 6

Tasa de verdaderos negativos de los modelos predictivos del rendimiento académico



Fuente: Construcción personal.

Discusión y conclusiones

En el presente trabajo se mostró que, empleando algunas características al inicio de un curso de ingeniería, se pueden construir modelos de predicción del rendimiento académico de estudiantes. Para evaluar los modelos se emplearon la exactitud, la tasa de verdaderos positivos y la tasa de verdaderos negativos. Se utilizaron diferentes técnicas de aprendizaje automático para construir los modelos, obteniendo distintos valores para las métricas de evaluación, es decir, se compararon los modelos con distintas métricas de evaluación.

En la construcción de los modelos predictivos de árbol de decisión y de *Naïve Bayes* se observó que el atributo más significativo para la predicción del rendimiento académico es el promedio actual del estudiante. En el modelo *Naïve Bayes* la probabilidad de que el promedio actual del estudiante sea menor a 7.5 es la más alta en el modelo, por lo que es la que más influye en las predicciones de este. En el modelo de árbol de decisión, el nodo raíz esta en función del atributo promedio actual, en específico, si este promedio es mayor a 8.5 el estudiante aprueba independientemente de cualquier valor de otro atributo. Esto concuerda con lo reportado por Shahiri et al., (2015), que indican que este atributo o característica del estudiante es uno de los más utilizados en la literatura para la predicción del rendimiento académico.

En la exactitud de los modelos predictivos del rendimiento académico se observa que es mayor con la técnica *k* vecinos más cercanos por casi 2% con respecto al modelo con *Naïve Bayes* y alrededor de 5% con respecto al modelo árbol de decisión, es decir, que las diferencias no son tan significativas con esta métrica. Sin embargo, en la tasa de verdaderos positivos se observa una diferencia más marcada de 15% con respecto al de *Naïve Bayes* y 17% con respecto al de árbol de decisión. La tasa de verdaderos positivos permite identificar las predicciones correctas de aprobación con respecto al total de aprobados, es decir, el modelo con la técnica *k* vecinos más cercanos sería el más adecuado para el conjunto de datos empleados, si lo que se desea es predecir con mayor certidumbre los estudiantes que aprueban el curso con respecto a todos los que realmente aprueban, además de ofrecer una mayor exactitud en las predicciones.

No obstante, con la tasa de verdaderos negativos se obtiene un valor mayor con el modelo con *Naïve Bayes*, obteniendo una diferencia de alrededor de 5% con el modelo de árbol de decisión y una diferencia de alrededor de 17% con respecto al modelo con *k* vecinos más cercanos. De esta manera, para el conjunto de datos empleados en el análisis, si lo que se desea es predecir con mayor certidumbre los estudiantes que reprueban con respecto al total que realmente reprobaron, el modelo con *Naïve Bayes* resultaría el más adecuado manteniendo un cierto compromiso con la exactitud.

En la literatura existen trabajos que utilizan técnicas de aprendizaje similares a los utilizados en la presente investigación. Juárez et al. (2014) realizaron un estudio

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

en el que participaron 104 estudiantes y el valor más alto de exactitud fue de 80%. Salal et al. (2019) emplearon 649 registros de estudiantes y el valor mayor de exactitud obtenido fue 76.7%. Para ambos trabajos se requirieron alrededor de 30 o más atributos correspondientes a datos personales, de domicilio, entre otros, además, solo se calculó la exactitud como métrica de evaluación.

Cabe mencionar que, si bien la exactitud de las predicciones en estos trabajos fue mayor a la obtenida en el presente artículo, en esta investigación se emplearon solamente nueve características de estudiantes y se calculó no solo la exactitud sino también la tasa de verdaderos positivos y negativos, las cuales tienen la potencialidad de brindar información adicional acerca de los modelos predictivos. Se debe notar que recabar menos características de los estudiantes facilita la recolección y análisis de datos.

Castrillón et al. (2020) utilizaron únicamente la técnica de aprendizaje automático árbol de decisión y 22 atributos. Se emplearon 460 registros de estudiantes para el modelo y se obtuvo una exactitud del 91%, sin embargo, una de las razones de este valor puede deberse a que la exactitud fue calculada con métodos que introducen poca aleatoriedad a la evaluación, entre los cuales fueron la validación cruzada con dos particiones y otras evaluaciones que consistieron en dividir los datos en una partición para crear el modelo y otra partición para realizar las predicciones. Estos métodos de evaluación con poca aleatoriedad tienden a favorecer al modelo que se ajusta más al conjunto de datos de entrenamiento (Hernández et al., 2004).

A diferencia del artículo mencionado, en este trabajo se utilizó la validación cruzada con diez particiones, la cual introduce mayor aleatoriedad en la evaluación, brindando resultados más confiables en las métricas de los modelos; también se utilizó una menor cantidad de atributos y se emplearon más métricas de evaluación para caracterizar de mejor manera los modelos de predicción. Es necesario puntualizar que entre mayor sea la exactitud de las predicciones más útil es el modelo en ambientes reales para predecir el rendimiento académico.

Finalmente, se debe notar que las características empleadas en este artículo pueden ser utilizadas en la construcción de modelos predictivos del rendimiento académico para otros cursos. Es decir, la metodología desarrollada puede ser replicada para otros cursos y los atributos empleados pueden ser recabados al inicio del curso o incluso antes de que el curso haya iniciado, lo cual ofrece un tiempo razonable para la elaboración de estrategias de intervención para estudiantes en peligro de reprobación.

Referencias

Castrillón, O. D., Sarache, W., y Ruiz-Herrera, S. (2020). Prediction of academic performance using artificial intelligence techniques. *Formación Universitaria*, *13*(1), 93-102. https://doi.org/10.4067/S0718-50062020000100093

Contreras, L. E., Fuentes, H. J., y Rodríguez, I. (2020). Academic interruption model using automatic learning algorithms. *International Journal of Mechanical and Production Engineering Research and Development*

(IJMPERD), *10*(3), 16075-16086. http://www.tjprc.org/publishpapers/2-67-1602700574-IJMPERD-JUN20201525.pdf

Cover, T., y Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, *13*(1), 21-27. http://dx.doi.org/10.1109/TIT.1967.1053964

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

Durairaj, M., y Vijitha, C. (2014). Educational data mining for prediction of student performance using clustering algorithms. *International Journal of Computer Science and Information Technologies*, *5*(4), 5987-5991. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.567.8824&rep=rep1&type=pdf

Estrada, R. I., Zamarripa, R. A., Zúñiga, P. G., y Martínez, I. (2016). Aportaciones desde la minería de datos al proceso de captación de matrícula en instituciones de educación superior particulares. *Revista Electrónica Educare*, *20*(3), 1-21. http://www.redalyc.org/articulo.oa?id=194146862011

Gómez, D., Oviedo, R., y Martínez, E. (2011). Factores que influyen en el rendimiento académico del estudiante universitario. *Educación y Humanidades*, *5*(2), 90-97. http://tecnociencia.uach.mx/numeros/v5n2/data/Factores_que_influyen_en_el_rendimiento_academico_del_estudiante_universitario.pdf

Gutiérrez, J. A., Garzón, J. y Segura, A. M. (2021). Factores asociados al rendimiento académico en estudiantes universitarios. *Formación Universitaria*, *14*(1), 13-24. http://dx.doi.org/10.4067/S0718-50062021000100013

Han, J. (2012). *Data mining: Concepts and techniques*. Morgan Kaufmann Publishers.

Hernández, J., Ramírez, M., y Ferri, C. (2004). *Introducción a la minería de datos*. Pearson.

Juárez, A., Cortés, J., y Coronilla, U. (2014). Aplicación de la inteligencia artificial en la sistematización de procesos educativos. Caso: sistema de detección de riesgo escolar en ESCOM. *Revista Iberoamericana de Producción Académica y Gestión Educativa*, *1*(1), 140-163. https://pag.org.mx/index.php/PAG/article/view/92/140

Kumar, V., y Chadha, A. (2011). An empirical study of the applications of data mining techniques in higher education. *International Journal of Advanced Computer Science and Applications*, *2*(3), 80-84. http://dx.doi.org/10.14569/IJACSA.2011.020314

Márquez, C., Cano, A., Romero, C., y Ventura, S. (2012). Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Applied Intelligence*, *38*(3), 315-330. http://dx.doi.org/10.1007/s10489-012-0374-8

Martínez, A., Hernández, L. I., Carillo, D., Romualdo, Z., y Hernández, C. P. (2013). Factores asociados a la reprobación estudiantil en la Universidad de la Sierra Sur, Oaxaca. *Temas de Ciencia y Tecnología*, *17*(51), 25-33. https://www.utm.mx/edi_anteriores/temas51/T51_1Ensayo3-FactAsocReprobacion.pdf

Mendoza, A. A., y Herrera, R. J. (2013). *Propuesta para la predicción del rendimiento académico de los estudiantes de la universidad del Atlántico, basado en la aplicación del análisis discriminante* [Ponencia]. Encuentro Internacional de Educación en Ingeniería, Cartagena de Indias, Colombia. https://acofipapers.org/index.php/eiei/article/view/1442

Montero, E., Villalobos, J., y Valverde, A. (2007). Factores institucionales, pedagógicos, psicosociales y sociodemográficos asociados al rendimiento académico en la Universidad de Costa Rica: un análisis multinivel. *Revista Electrónica de Investigación y Evaluación Educativa*, *13*(2), 215-234. https://www.redalyc.org/articulo.oa?id=91613205

Mueen, A., Zafar, B., y Manzoor, U. (2016). Modeling and predicting students' academic performance using data mining techniques. *International Journal of Modern Education and Computer Science*, *8*(11), 36-42. http://dx.doi.org/10.5815/ijmecs.2016.11.05

Peña, A. (2014). Review: Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, *41*(4),1432-1462. https://doi.org/10.1016/j.eswa.2013.08.042

Quinlan, J. (1993). *C4.5: Programs for machine learning*. Morgan Kaufmann.

ie revista de investigación educativa de la Rediech
vol. 13 • 2022 • e1426 • ISSN: 2448-8550

Ridgell, S. D., y Lounsbury, J. W. (2004). Predicting academic success: general intelligence, "big five" personality traits, and work drive. *College Student Journal*, *38*(4), 607-618. https://psycnet.apa.org/record/2004-22470-015

Romero, C., y Ventura, S. (2010). Educational data mining: A review of the state of the art. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (applications and reviews)*, *40*(6), 601-618. https://doi.org/10.1109/TSMCC.2010.2053532

Romero, C., y Ventura, S. (2012). Data mining in education. *Wiley Interdisciplinary Reviews: Data mining and knowledge discovery*, *3*(1), 12-27. https://doi.org/10.1002/widm.1075

Salal, Y. K., Abdullaev, S. M., y Kumar, M. (2019). Educational data mining: Student performance prediction in academic. *International Journal of Engineering and Advanced Technology*, 8(4C), 54-59. https://www.researchgate.net/publication/332369964_Educational_Data_Mining_Student_Performance_Prediction_in_Academic

Sánchez, D. (2009). *Agentes inteligentes: diseño e implementación para la enseñanza de la física* [Tesis de doctorado]. Centro de Investigación en Ciencia Aplicada y Tecnología Avanzada, IPN. Repositorio DSpace Tesis IPN. https://tesis.ipn.mx/bitstream/handle/123456789/8108/AGEINTEL.pdf?sequence=1&isAllowed=y

Shahiri, A. M., Husain, W., y Rashid, N. (2015). A review on predicting student's performance using data mining techniques. *Procedia Computer Science*, *72*, 414-422. https://doi.org/10.1016/j.procs.2015.12.157

Tapasco, O. A., Ruiz, F. J., Osorio, D., y Ramírez, D. (2020). El historial académico de secundaria como factor predictor del rendimiento universitario. Caso de estudio. *Revista Colombiana de Educación*, *1*(81), 147-170. https://doi.org/10.17227/rce.num81-753

Timarán, R., Calderón, A., y Jiménez, J. (2013). Descubrimiento de perfiles de deserción estudiantil con técnicas de minería de datos. *Revista Vínculos*, *10*(1), 373-383. https://doi.org/10.14483/2322939X.4687

Torres, P. C., y Cobo, J. K. (2017). Tecnología educativa y su papel en el logro de los fines de la educación. *Educere*, *21*(68), 31-40. https://dialnet.unirioja.es/servlet/articulo?codigo=6560961

Witten, I., Frank, E., y Hall, M. (2005). *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Cómo citar este artículo:

Rico Páez, A., y Gaytán Ramírez, N. D. (2022). Modelos predictivos del rendimiento académico a partir de características de estudiantes de ingeniería. *IE Revista de Investigación Educativa de la REDIECH*, 13, e1426. https://doi.org/10.33010/ie_rie_rediech.v13i0.1426