REVISTA ELECTRÓNICA SOBRE TECNOLOGÍA, EDUCACIÓN Y SOCIEDAD
ISSN: 2448 - 6493
CTES

Artículos científicos

Predicción del rendimiento académico mediante selección de características de estudiantes universitarios

Prediction of academic performance through selection of characteristics of university students

Andrés Rico Páez
Instituto Politécnico Nacional, México
aricop.ipn@gmail.com
https://orcid.org/0000-0002-6450-318X

Resumen

El propósito de esta investigación fue elaborar modelos de predicción del rendimiento académico de estudiantes universitarios de México con técnicas de aprendizaje automático, considerando la selección de atributos más significativos. En este trabajo, se recabaron datos académicos y demográficos de 51 estudiantes universitarios para elaborar modelos que predigan su rendimiento académico al final del curso. Se utilizaron las técnicas de aprendizaje automático *Naïve Bayes*, k vecinos más cercanos y árbol de decisión C4.5. Se observó en los modelos una mayor exactitud cuando se realizó una selección de los atributos más significativos en comparación con cuando se utilizaron todos los atributos. Se obtuvo una exactitud del 78.43% cuando se emplearon cinco atributos más significativos y la técnica *Naïve Bayes*. La metodología mostrada puede ser aplicada a diferentes tipos de cursos o modalidades. Los resultados muestran que, con la selección de los atributos más significativos, se puede obtener una mejora en la exactitud de las predicciones, brindando mayor certeza a las instituciones educativas para la identificación de estudiantes en riesgo de reprobación.

Palabras clave: rendimiento académico, aprendizaje automático, selección de atributos, modelo de predicción.

REVISTA ELECTRÓNICA SOBRE TECNOLOGÍA, EDUCACIÓN Y SOCIEDAD
ISSN: 2448 - 6493
CTES

Abstract

The purpose of this research was to develop models for predicting the academic performance of university students in Mexico with machine learning techniques considering the selection of the most significant attributes. In this work, academic and demographic data of 51 university students were collected to develop models to predict their academic performance at the end of the course. The Naïve Bayes machine learning techniques, k nearest neighbors and C4.5 decision tree were used. A greater accuracy was observed in the models when a selection of the most significant attributes was made compared to when all the attributes are used. An accuracy of 78.43% was obtained when the five most significant attributes and the Naïve Bayes technique were used. The methodology shown can be applied to different types of courses or modalities. The results show that with the selection of the most significant attributes, an improvement in the accuracy of the predictions can be obtained, giving greater certainty to educational institutions for the identification of students at risk of failure.

Keywords: machine learning, mathematical model, prevention, academic performance.

Fecha Recepción: Julio 2022 **Fecha Aceptación:** Enero 2023

Introducción

El desarrollo tecnológico ha provocado una gran cantidad de actividades digitales, tales como transacciones bancarias, uso de redes sociales, entre muchos otros. Esto ha originado un incremento exponencial en la cantidad de datos almacenados y en el análisis de datos. Las técnicas de aprendizaje automático están siendo utilizadas por las empresas para analizar datos y tomar decisiones que les puedan ofrecer algún tipo de ventaja o beneficio competitivo. Es decir, las técnicas de aprendizaje automático generan predicciones que permiten tomar decisiones para llevar a cabo determinadas actividades a partir de la experiencia de los datos previos. El área educativa no ha sido ajena a la aplicación de este tipo de técnicas para analizar datos, y en años recientes, las instituciones educativas están analizando sus datos con este tipo de técnicas con el propósito de comprender los procesos de aprendizaje y tratar de mejorar, en la medida de lo posible, el desempeño académico de los estudiantes (Czibula et al., 2019).

El rendimiento académico involucra una gran cantidad de variables, es decir, es la combinación de distintos procesos que ocurren alrededor del estudiante (Grimaldo y Manzanares, 2023). Además, el rendimiento académico es uno de los más importantes

indicadores acerca de la calidad educativa de una institución (Sánchez, 2022). De esta manera, las instituciones educativas están en la búsqueda de métodos que permitan mejorar la calidad de sus procesos educativos y han encontrado en el análisis de datos educativos una manera de hacerlo. Actualmente, existen varios trabajos que analizan datos de entornos educativos para la predicción del rendimiento académico de estudiantes (Incio, Capuñay y Estela, 2022; Vásconez, 2023; Contreras, Nieves y González, 2022). La predicción del rendimiento académico puede aplicarse a diferentes acciones de formación, como predecir el resultado de actividades o tareas, exámenes parciales o finales, o incluso cursos académicos (Asif et al., 2017). De cualquier manera, la predicción del rendimiento académico ayuda a los profesores a adaptar su método de enseñanza a las características de los estudiantes. De igual forma, la predicción del rendimiento académico ofrece la posibilidad de involucrar a los estudiantes en su proceso de aprendizaje para que tengan una mayor comprensión de sus habilidades y tomen las medidas necesarias para ajustarlas de acuerdo con sus necesidades.

En años recientes, para realizar las predicciones del rendimiento académico, se construyen modelos predictivos por medio de técnicas de aprendizaje automático a partir de datos de entornos educativos (Jiménez y Toapanta, 2023). Usman et al. (2019) utilizaron distintas técnicas de aprendizaje automático para predecir el rendimiento académico a partir de la interacción que tienen con una plataforma de internet educativa. Díaz, Meleán y Marín (2021) predijeron el rendimiento académico de estudiantes mediante la técnica de árbol de decisión J48, considerando factores educativos, familiares, entre otros. Contreras, Fuentes y Rodríguez (2020) proponen la selección de ciertas variables que pueden influir en la predicción del rendimiento académico de estudiantes de ingeniería industrial, implementando diferentes técnicas de aprendizaje automático. Se debe notar que la selección de atributos más significativos puede mejorar los modelos para aumentar, en cierta medida, la exactitud de las predicciones. De esta manera, existe una necesidad de elaborar metodologías de predicción del rendimiento académico que consideren no solo distintas técnicas de aprendizaje automático, sino también la selección de atributos más significativos para mejorar la exactitud de los modelos predictivos.

En este trabajo se plantean las siguientes preguntas de investigación: ¿Cómo se puede realizar la selección de los atributos más significativos para la predicción del rendimiento académico de estudiantes de una universidad de México? ¿Cómo construir los modelos para la predicción del rendimiento académico con distintas técnicas de aprendizaje automático? y ¿cómo influye la selección de atributos más significativos en la exactitud de las predicciones

de los modelos realizados? Por lo tanto, el propósito de este artículo es elaborar una metodología para construir modelos de predicción del rendimiento académico de estudiantes con distintas técnicas de aprendizaje automático, considerando la selección de los atributos más significativos.

Metodología

En este artículo se utiliza una metodología basada en el proceso conocido como descubrimiento de conocimiento en bases de datos (Hernández, Ramírez y Ferri, 2004). Primeramente, se recopila información de estudiantes al iniciar un curso. Después, se elabora un análisis para la selección de los atributos o características más significativas del conjunto de datos de estudiantes para los modelos de predicción. Posteriormente, se desarrollan y evalúan modelos predictivos con distintas técnicas de aprendizaje automático, tomando en cuenta distintos atributos seleccionados con el propósito de identificar el modelo que obtenga mejor exactitud en las predicciones. La metodología descrita se presenta en la Figura 1.

Figura 1. Metodología utilizada para el desarrollo de modelos de predicción del rendimiento académico mediante selección de atributos de estudiantes.

Fuente: Elaboración propia

En este artículo participaron 51 estudiantes de un curso de tercer semestre de ingeniería de una universidad pública de México. Los datos fueron recopilados a través de una encuesta en la que se avisó a los estudiantes que era para fines de investigación. La información recabada fue la aprobación y reprobación del curso, así como características tales como el nivel educativo de los padres o el ingreso familiar. Algunos atributos, como el promedio, se transformaron a valores nominales con el propósito de adecuarse a las técnicas

de aprendizaje automático utilizadas en este trabajo. Los atributos específicos recabados y los distintos valores que pueden adquirir se ilustran en la Tabla 1.

Tabla 1. Atributos de estudiantes con sus respectivos valores.

Atributo	Valores
Escolaridad del padre	Básica, Media superior, Superior o posgrado
Escolaridad de la madre	Básica, Media superior, Superior o posgrado
Ingreso familiar	<$5000, $5000 - $10000, >$10000
Promedio de media superior	<7.5, 7.5-8.4, >8.4
Cursos reprobados actualmente	0, 1, >1
Promedio	<7.5, 7.5-8.4, >8.4
Forma de estudio	Solo, En pareja, En grupo
Modo para hacer actividades	Solo, En pareja, En grupo
Preferencia de estudio para un examen	Continuamente, Una semana antes del examen, Un día antes del examen
Edad	18 -19 años, 20 - 21 años, > 22 años
Lugar de nacimiento	Ciudad de México, Estado de México, Otro
Tipo de bachillerato	Pertenece al Instituto Politécnico Nacional (IPN), Pertenece a la Universidad Nacional Autónoma de México (UNAM), Otro
Tiempo de traslado a la escuela	<1 hr., 1 – 1.5 hrs., >1.5hrs.
Dominio de inglés	Básico, Intermedio, Avanzado
Apoyo familiar	Regular, Bueno, Excelente
Aprobación del curso	SI, NO

Fuente: Elaboración propia

El atributo "Aprobación del curso" determina la etiqueta de la clase y es la característica a predecir en futuros registros. Se debe notar que características o atributos de estudiantes similares a los mostrados en la Tabla 1 han sido empleados en trabajos semejantes de predicción del rendimiento académico (Shahiri, Husain y Rashid, 2015).

Después de recabar datos, se elaboró una tabla con 51 registros (estudiantes) y 16 columnas (atributos o características de los estudiantes); una parte de esta se ilustra en la Tabla 2.

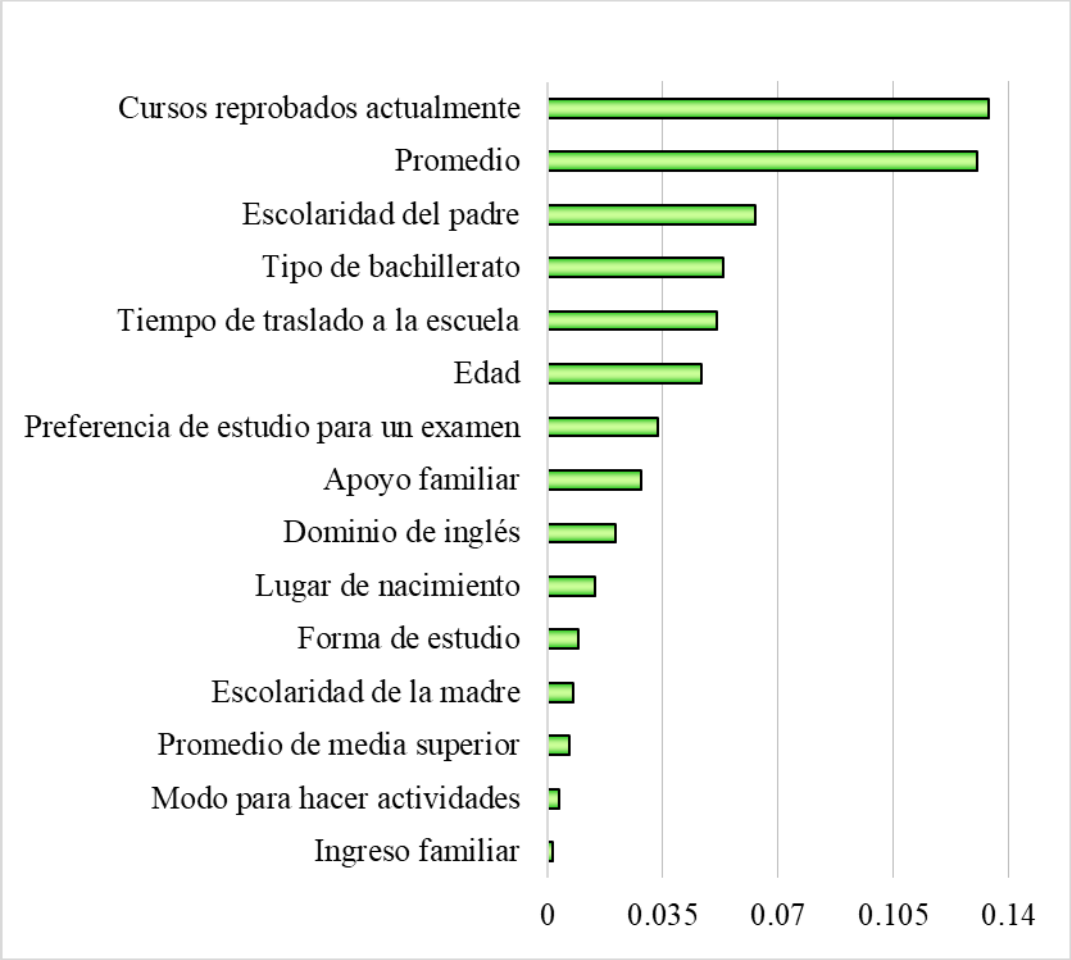
Tabla 2. Muestra de la tabla de datos recabada de estudiantes.

Escolaridad del padre	Escolaridad de la madre	Ingreso familiar	Promedio de media superior
Básica	Básica	$5000 - $10000	7.5-8.4
Media superior	Básica	$5000 - $10000	7.5-8.4
Superior o posgrado	Básica	<$5000	7.5-8.4
Superior o posgrado	Básica	$5000 - $10000	<7.5
Superior o posgrado	Superior o posgrado	$5000 - $10000	7.5-8.4
Media superior	Media superior	$5000 - $10000	<7.5
Superior o posgrado	Superior o posgrado	>$10000	>8.4

Fuente: Elaboración propia

A continuación, se realiza un análisis acerca de la selección de atributos que mayor impacto tienen en la predicción del rendimiento académico para diferentes técnicas de aprendizaje automático. Es decir, este estudio consiste en determinar qué atributos son los que más influyen en la predicción del atributo "Aprobación del curso", es decir, la clase a predecir del conjunto de datos. Para realizar la selección de atributos, se utiliza la relación de la ganancia de información con respecto a la clase basada en la teoría de la información (Hernández et al., 2018; Mosquera, Castrillón y Parra, 2018). Entre mayor sea la relación de la ganancia de clase de un atributo, entonces, más grande será su impacto en la clase a predecir. De esta manera, se pueden ordenar los atributos en función de dicha relación y seleccionar los que tengan valores mayores. En este trabajo, se calculó la relación de la ganancia de clase de los 15 atributos por medio del algoritmo *GainRatioAttributeEval*, que es parte del software de código abierto WEKA (*Waikato Environment for Knowledge Analysis*) (Abuhaija et al., 2023). Los valores de cada atributo se presentan en la Figura 2.

Figura 2. Atributos ordenados de acuerdo con su relación de ganancia de clase.



Fuente: Elaboración propia

En la siguiente sección, se detallan los resultados obtenidos a partir de las pruebas realizadas tomando en cuenta el orden de los atributos en función de la relación de ganancia de clase.

Resultados

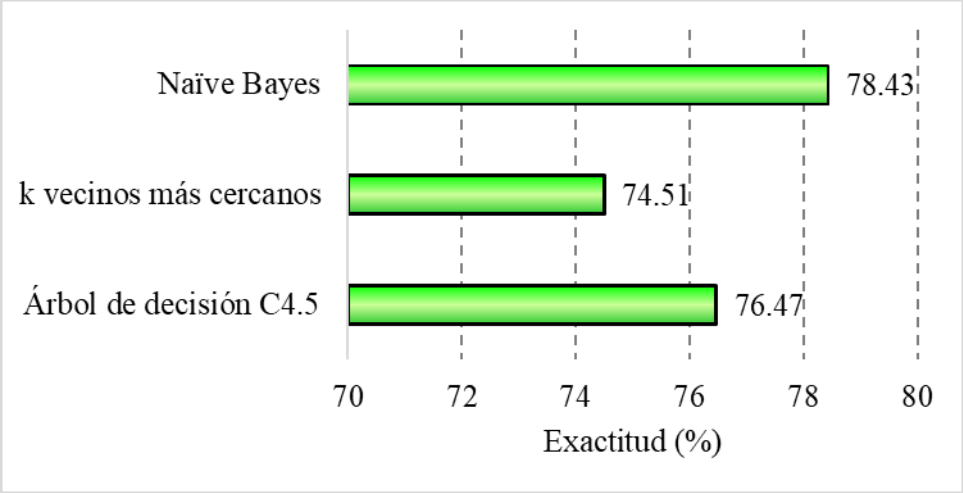
En esta sección, se realizan una serie de pruebas a partir de la selección de los atributos más significativos obtenidos en la sección anterior. En este artículo, se emplean técnicas de aprendizaje automático que han sido empleadas en la literatura para la predicción del rendimiento académico, las cuales son *Naïve Bayes* (Aziz et al., 2015), k vecinos más cercanos (Nugroho et al., 2020) y árbol de decisión C4.5 (Wang, Zhou y Xu, 2019). De igual manera, para evaluar los modelos construidos, se usa como métrica la exactitud de las predicciones, es decir, el porcentaje de los registros con predicciones correctas divididas entre el total de registros (Bedoya, Guarín y Agudelo, 2023). La exactitud se calcula

utilizando el método conocido como validación cruzada con 10 particiones (Postiguillo, Ribal y Blasco, 2018). Este método consiste en particionar aleatoriamente los datos en 10 grupos, uno de ellos se emplea para efectuar las predicciones y los demás para elaborar el modelo; esto se repite apartando un grupo diferente para realizar las predicciones. La exactitud se lleva a cabo realizando el promedio de las 10 exactitudes obtenidas con cada grupo de datos.

En el caso de la técnica de k vecinos más cercanos, se requiere especificar el parámetro k; para elegirlo, se hizo variar desde 1 hasta 51, que es la cantidad total de registros, calculando la exactitud de las predicciones con validación cruzada de 10 particiones. Se obtuvo que uno de los valores de mayor exactitud (72.549%) fue de k=6, por lo que es el valor que se utilizará en las pruebas.

En la primera prueba, para realizar los modelos predictivos, solo se consideran los datos de los primeros cinco atributos de mayor relación de ganancia de clase de acuerdo con la Figura 2, es decir, los cinco atributos más significativos. Estos atributos son "Cursos reprobados actualmente", "Promedio", "Escolaridad del padre", "Tiempo de traslado a la escuela" y "Edad". A partir de estos atributos, se construyen los modelos de predicción con las tres técnicas de aprendizaje automático. En la Figura 3 se muestra la exactitud de las predicciones con estas técnicas y se puede observar que la técnica *Naïve Bayes* es la que consigue un mayor valor de exactitud, aunque no tan alejado del valor obtenido por la técnica de árbol de decisión.

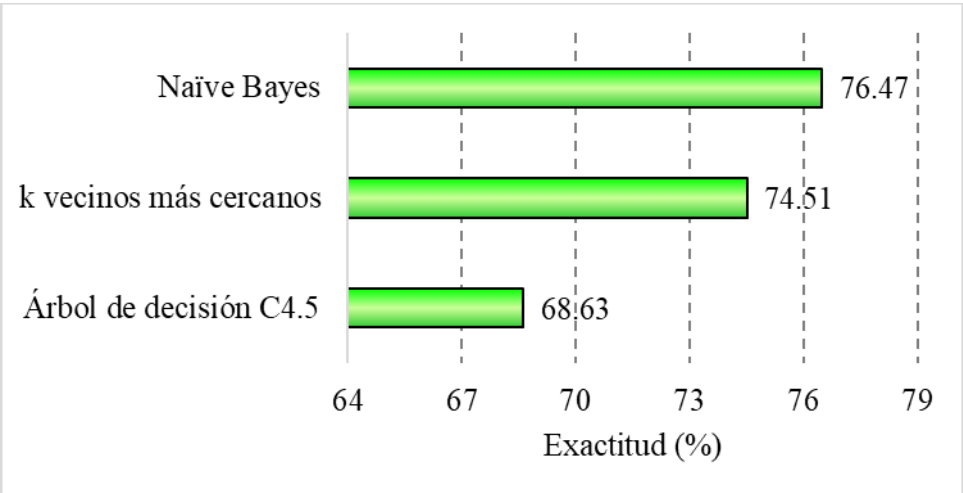
Figura 3. Exactitud de las técnicas de aprendizaje automático utilizando los cinco atributos más significativos.



Fuente: Elaboración propia

En la siguiente prueba, se toman los 10 atributos más significativos de los datos para hacer los modelos predictivos con las tres técnicas de aprendizaje automático. La exactitud de las predicciones se presenta en la Figura 4, en donde se puede notar que la exactitud con la técnica árbol de decisión es la que tiene un valor menor de exactitud.

Figura 4. Exactitud de las técnicas de aprendizaje automático utilizando los 10 atributos más significativos.

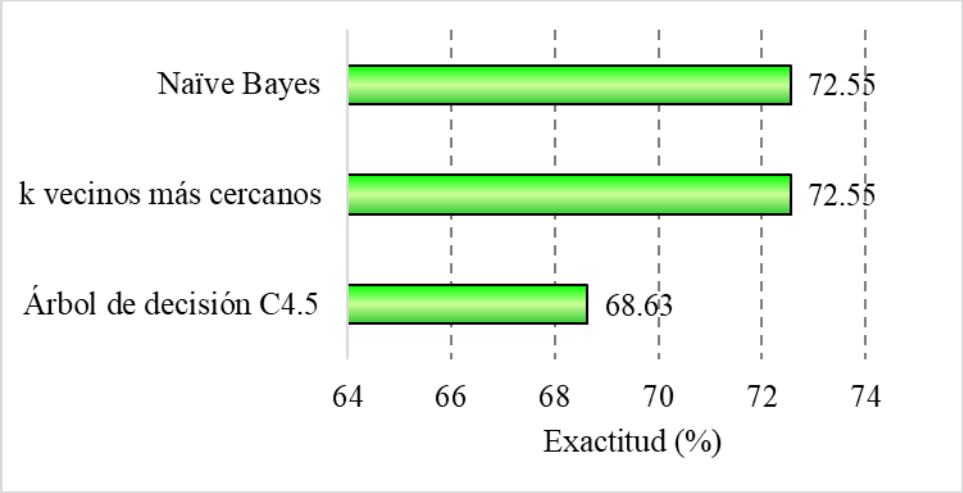


Fuente: Elaboración propia

A continuación, se construyen los modelos predictivos tomando todos los atributos de los datos, es decir, los 15 atributos para calcular la exactitud con las tres técnicas de

aprendizaje automático mostrada en la Figura 5, en donde nuevamente se puede observar que la exactitud con la técnica árbol de decisión es la que tiene un valor más pequeño de exactitud.

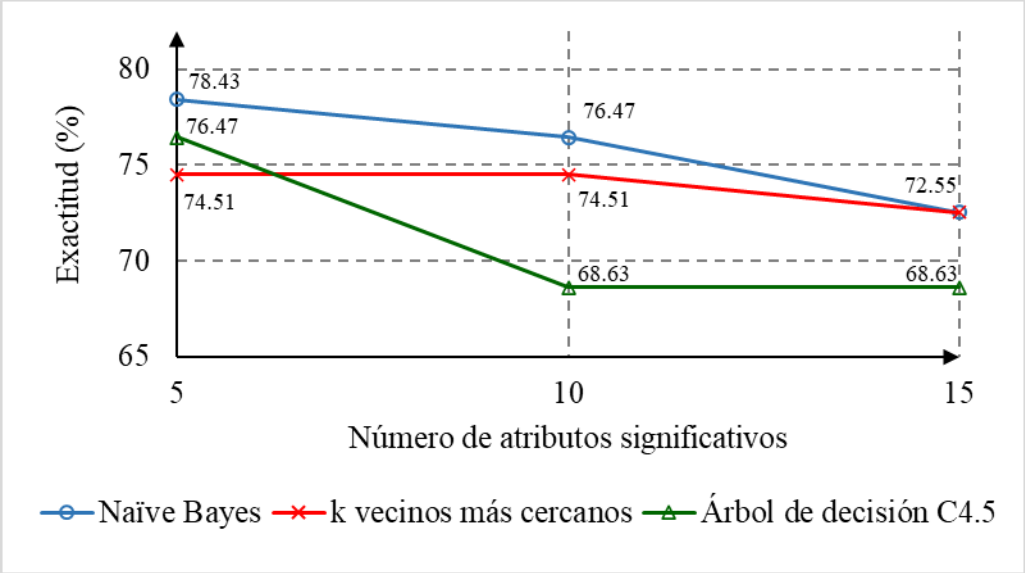
Figura 5. Exactitud de las técnicas de aprendizaje automático utilizando todos los atributos.



Fuente: Elaboración propia

En las figuras anteriores, se observa que la técnica *Naïve Bayes* tiene los valores mayores de exactitud en las pruebas anteriores, aunque en la última prueba tiene el mismo valor que la técnica de k vecinos más cercanos. Para visualizar el comportamiento de la exactitud de las predicciones, en la Figura 6, se grafica esta métrica con respecto a la técnica de aprendizaje automático y al número de atributos más significativos utilizados en las figuras anteriores.

Figura 6. Exactitud de las técnicas de aprendizaje automático con respecto al número de atributos más significativos.



Fuente: Elaboración propia

En la figura anterior, se observa que la exactitud con el valor menor es con la técnica de árbol de decisión cuando se utilizan 10 atributos más significativos y cuando se emplean todos los atributos (68.63%). Sin embargo, el caso que es de interés es cuando la exactitud tiene el valor más alto de exactitud, el cual se obtiene con la técnica *Naïve Bayes* y con los cinco atributos más significativos (78.43%).

El modelo predictivo construido con la técnica *Naïve Bayes* consiste en un conjunto de probabilidades del atributo de clase y en probabilidades condicionales que utilizan los demás atributos (Bustomi et al., 2023). Estas probabilidades se calculan con ayuda del software WEKA para los cinco atributos más significativos y se muestran en la Tabla 3.

Tabla 3. Probabilidades de la técnica *Naïve Bayes* considerando los cinco atributos más significativos.

Probabilidades	clase = SI	clase=NO
P(Aprobación del curso=clase)	0.7	0.3
P(Escolaridad del padre=Básica/Aprobación del curso=clase)	0.49	0.17
P(Escolaridad del padre= Media superior /Aprobación del curso=clase)	0.2	0.28
P(Escolaridad del padre= Superior o posgrado Aprobación del curso=clase)	0.31	0.55
P(Tipo de bachillerato=Pertenece al IPN/Aprobación del curso=clase)	0.67	0.39
P(Tipo de bachillerato=Pertenece a la UNAM/Aprobación del curso=clase)	0.15	0.17
P(Tipo de bachillerato=Otro/Aprobación del curso=clase)	0.18	0.44
P(Tiempo de traslado a la escuela =<1 hr./Aprobación del curso=clase)	0.26	0.28
P(Tiempo de traslado a la escuela=1 – 1.5 hrs./Aprobación del curso=clase)	0.43	0.17
P(Tiempo de traslado a la escuela=>1.5hrs./Aprobación del curso=clase)	0.31	0.55
P(Cursos reprobados actualmente=0/Aprobación del curso=clase)	0.41	0.22
P(Cursos reprobados actualmente=1/Aprobación del curso=clase)	0.51	0.28
P(Cursos reprobados actualmente=>1/Aprobación del curso=clase)	0.08	0.5
P(Promedio=<7.5/Aprobación del curso=clase)	0.38	0.78
P(Promedio=7.5-8.4/Aprobación del curso=clase)	0.54	0.17
P(Promedio=>8.4/Aprobación del curso=clase)	0.08	0.05

Fuente: Elaboración propia

Para observar el beneficio del modelo predictivo de la Tabla 3, se aplicó a 10 estudiantes del mismo curso de la misma universidad en la cual se recabaron los datos, pero de un semestre posterior (datos de prueba). En la técnica *Naïve Bayes*, para predecir un nuevo registro, se asocian sus atributos con las probabilidades de la Tabla 3, se realizan las multiplicaciones de las probabilidades para cuando el atributo "Aprobación del curso" es igual a "SÍ" y luego todas las multiplicaciones para cuando es igual a "NO". Finalmente, la predicción del nuevo registro es la clase asociada al valor del producto más alto. De esta manera, se realizaron las predicciones y, posteriormente, se contrastaron con los resultados de aprobación logrados por los estudiantes al terminar el curso. Los registros de estos estudiantes, sus predicciones y resultados al final del curso se presentan en la Tabla 4.

Tabla 4. Predicciones de datos de prueba de estudiantes.

Escolaridad del padre	Tipo de bachillerato	Tiempo de traslado a la escuela	Cursos reprobados actualmente	Promedio	Predicción	Aprobación verdadera
Superior o posgrado	IPN	<1 hr.	1	<7.5	SI	NO
Básica	IPN	1 – 1.5 hrs.	0	7.5-8.4	SI	SI
Básica	UNAM	>1.5hrs.	0	>8.4	SI	SI
Superior o posgrado	UNAM	>1.5hrs.	>1	<7.5	NO	NO
Superior o posgrado	IPN	>1.5hrs.	>1	7.5-8.4	NO	NO
Media superior	IPN	1 – 1.5 hrs.	1	<7.5	SI	SI
Media superior	IPN	1 – 1.5 hrs.	1	7.5-8.4	SI	SI
Media superior	IPN	<1 hr.	0	<7.5	SI	NO
Básica	IPN	<1 hr.	1	7.5-8.4	SI	SI
Superior o posgrado	IPN	<1 hr.	1	<7.5	SI	SI

Fuente: Elaboración propia

En la Tabla 4 se muestran las características de los estudiantes de los datos de prueba. Además, se puede observar que solo ocho predicciones coinciden con la aprobación verdadera obtenida de los estudiantes al terminar el curso, es decir, se obtiene una exactitud de las predicciones de 80%.

Discusión

En la sección anterior, se observó la exactitud de las predicciones con modelos predictivos realizados con distintas técnicas de aprendizaje automático y diferentes cantidades de atributos significativos. De las técnicas de aprendizaje automático, Naïve Bayes es la que muestra mejores valores de exactitud con distintos valores de mejores atributos. Esto concuerda con lo observado en otros trabajos en los que esta técnica tiene mejor exactitud en comparación a otras cuando se utilizan cantidades de datos similares a los empleados en este trabajo (Osmanbegović y Suljić, 2012; Mueen, Zafar y Manzoor, 2016). De igual manera, se observa en la Figura 6, que seleccionando los atributos más significativos se tiene mejor exactitud en comparación a cuando se emplean todos los atributos. Esto coincide con lo observado en Contreras et al. (2020), en el cual la exactitud con las técnicas de aprendizaje automático empleadas a partir de atributos seleccionados en este estudio es superior a la exactitud en comparación a cuando se emplean todos los atributos.

El modelo predictivo con la mayor exactitud fue con la técnica Naïve Bayes y con los cinco atributos más significativos. En esta técnica se llevan a cabo multiplicaciones de las probabilidades para cuando el atributo "Aprobación del curso" es igual a "SÍ" y luego todas las multiplicaciones para cuando es igual a "NO". De esta manera, las probabilidades con valor mayor son las que influyen en las multiplicaciones y en las predicciones. Entre más grande sea la diferencia entre las mismas probabilidades, pero con diferente valor del atributo "Aprobación del curso", mayor será su influencia en la predicción. Esto ocurre con las probabilidades del atributo P(Cursos reprobados actualmente>=1/Aprobación del curso=SÍ) = 0.08 y P(Cursos reprobados actualmente>=1/Aprobación del curso=NO) = 0.5, cuya diferencia absoluta es de 0.42 y es la mayor de toda la Tabla 3. Esto significa que si el estudiante tiene más de un curso reprobado, las multiplicaciones de la clase "SÍ" disminuyen su valor al ser multiplicadas por 0.08, y al mismo tiempo, las multiplicaciones de la clase "NO" aumentan su valor al ser multiplicadas por 0.5, por lo que, en este caso, se aumenta la estimación de que la predicción sea que el alumno repruebe el curso. Esto mismo concuerda con lo presentado en la Figura 3, en donde el atributo más significativo es el de "Cursos reprobados actualmente".

En años recientes, se han observado trabajos acerca de la predicción del rendimiento académico empleando técnicas de aprendizaje automático. Díaz et al. (2021) elaboraron un estudio en el cual se tuvo una muestra de 237 estudiantes con 26 atributos y se utilizó la técnica de árbol de decisión J48, obteniendo una exactitud del 62.45%. Contreras et al. (2020) utilizaron cuatro técnicas de aprendizaje automático y emplearon 1571 registros con 30 atributos, de los cuales realizaron una selección de los mejores atributos y utilizaron los 10 mejores; la exactitud más alta obtenida fue de 66.24%. A diferencia de estos trabajos, en este artículo se obtuvo una exactitud máxima de 78.43% con tan solo 5 atributos de los estudiantes. Además, se probó el modelo en datos de prueba y se obtuvo una exactitud del 80%.

Conclusiones

En este artículo se realizó una metodología que elabora un análisis para la selección de atributos más significativos para la predicción del rendimiento académico de estudiantes. Para realizar la selección de estos atributos, se utilizó la relación de la ganancia de información con respecto a la clase a predecir, de tal manera que, entre mayor sea esta relación, mayor es su influencia en el atributo a predecir, lo cual permite ordenar los atributos

de acuerdo con su impacto en la predicción del rendimiento académico. Para construir los modelos predictivos del rendimiento académico de estudiantes, se utilizaron 51 registros con 15 atributos más su aprobación en un curso de ingeniería. Además, se emplearon las técnicas de aprendizaje automático *Naïve Bayes*, k vecinos más cercanos y árbol de decisión C4.5. Se observó que la exactitud de las predicciones con estas técnicas es mayor cuando se utilizaron los atributos más significativos en comparación con cuando se utilizan todos los atributos, teniendo en cuenta que esto es válido para el conjunto de datos utilizado en este estudio, llegando a obtener una exactitud de 78.43% con los cinco atributos más significativos y la técnica *Naïve Bayes*. Además, cuando se aplicó este modelo a un conjunto de datos de prueba, se obtuvo una exactitud de 80%. La metodología utilizada en este estudio puede ser aplicada a otro tipo de cursos o modalidades, como en la educación a distancia, en donde la recopilación de datos puede ser automatizada mediante plataformas digitales. También es importante resaltar que, realizando una selección adecuada de atributos, se pueden mejorar la exactitud de las predicciones de los modelos, ofreciendo una mayor confiabilidad en la detección de estudiantes que vayan a reprobar por parte de la institución educativa y realizar algún tipo de acción para prevenirla.

Futuras líneas de investigación

Es importante mencionar que, a pesar de los avances mostrados, se pueden realizar varios estudios en esta misma línea de investigación. En primer lugar, se pueden utilizar una mayor cantidad de registros de estudiantes. De igual manera, se pueden emplear otro tipo de técnicas de aprendizaje automático. Además, se pueden utilizar otros atributos demográficos, personales, motivacionales, entre otros, que puedan tener alguna influencia en el rendimiento académico. También se puede aplicar a otros contextos educativos, como en la educación a distancia, donde la recopilación de datos sería más rápida por medio de plataformas en línea.

Referencias

Abuhaija, B., Alloubani, A., Almatari, M., Jaradat, G. M., Abdalla, H. B., Abualkishik, A. M. (2023). "A comprehensive study of machine learning for predicting cardiovascular disease using Weka and SPSS". *International Journal of Electrical and Computer Engineering, 13*(2), 1891-1902. http://doi.org/10.11591/ijece.v13i2.pp1891-1902

Aziz, A. A., Ismail, N. H., Ahmad, F., y Hassan, H. (2015). A Framework for Students' Academic Performance Analysis using Naïve Bayes Classifier. *Jurnal Teknologi, 75*(3), 13-19. https://doi.org/10.11113/jt.v75.5037

Asif, R., Merceron, A., Ali, S. A. y Haider, N. G. (2017). Analyzing undergraduate students' performance using educational data mining. *Computers & Education, 113*, 177-194. https://doi.org/10.1016/j.compedu.2017.05.007.

Bedoya O. Guarín H. S. y Agudelo, J. (2023). Aplicación de técnicas de inteligencia artificial para la detección de tuberculosis pulmonar en Colombia. *Revista EIA, 20*(39). 1-23. https://doi.org/10.24050/reia.v20i39.1617

Bustomi, Y., Nugraha, A., Juliane, C. y Rahayu, S. (2023). Data Mining Selection of Prospective Government Employees with Employment Agreements using Naive Bayes Classifier. *Sinkron: Jurnal Dan Penelitian Teknik Informatika, 8*(1), 1-8. https://doi.org/10.33395/sinkron.v8i1.11968

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.

Contreras, L. E., Nieves, N. y Gonzalez, K. (2022). Prediction of University-Level Academic Performance through Machine Learning Mechanisms and Supervised Methods. Ingeniería, 28(1), e19514. https://doi.org/10.14483/23448393.19514

Czibula, G., Mihai, A. y Crivei, L. M. (2019). S PRAR: A novel relational association rule mining classification model applied for academic performance prediction. Procedia Computer Science, 159, 20–29. https://doi.org/10.1016/j.procs.2019.09.156

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

Grimaldo, M. y Manzanares, E. (2023). Variables intervinientes en el rendimiento académico en ingresantes de una universidad privada de Lima. *Revista Electrónica Educare*, *27*(1), 1-14. https://doi.org/10.15359/ree.27-1.6

Hernández, E. J., Quintero, D. P., Escobar, J. C., Ramírez, J. S. y Duque, N. D. (2018). Educational data mining for the analysis of student desertion. *Learning Analytics for Latin America 2018*(2231), 51-60. https://ceur-ws.org/Vol-2231/LALA_2018_paper_8.pdf

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

Incio, F., Capuñay, D. y Estela, R. (2022). Artificial Neural Network Model to Predict Academic Results in Mathematics II. *Revista Electrónica Educare, 27*(1), 1-19. https://doi.org/10.15359/ree.27-1.14516

Jiménez, S. A. y Toapanta, A. E. M. (2023). Modelos de Aprendizaje Automático basados en CRISP-DM para el Análisis de los niveles de Depresión en los estudiantes de la Escuela Politécnica Nacional. *Latin-American Journal of Computing, 10*(1), 22-43.

Mosquera, R., Castrillón, O. D. y Parra, L. (2018). Predicción de riesgos psicosociales en docentes de colegios públicos colombianos utilizando técnicas de Inteligencia Artificial. *Información tecnológica, 29*(4), 267-280. http://dx.doi.org/10.4067/S0718-07642018000400267

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. https://doi.org/10.5815/ijmecs.2016.11.05

Nugroho, A., Riady, O. R., Calvin, A. y Suhartono, D. (2020). Identification of Student Academic Performance using the KNN Algorithm. *Engineering, Mathematics and Computer Science (EMACS) Journal, 2* (3), 115-122. https://doi.org/10.21512/emacsjournal.v2i3.6537

Osmanbegović, E. y Suljić, M. (2012). Data Mining Approach for Predicting Student Performance. *Journal of Economics and Business, 10*(1), 3-12. http://hdl.handle.net/10419/193806

Postiguillo, D., Ribal, J. y Blasco, A. (2018). Caso de estudio: Modelización de la vida útil de maquinaria destinada a proyectos de obra pública a través de la aplicación de

regresiones mínimo cuadráticas y validación cruzada. *Finance, Markets and Valuation, 4*(1), 57-79. http://hdl.handle.net/10251/122884

Sánchez, I. (2022). Creatividad y rendimiento académico: Dos variables inseparables y controvertidas en futuros maestros de educación infantil. *Revista Internacional De Pedagogía E Innovación Educativa, 3*(1), 11–30. https://doi.org/10.51660/ripie.v3i1.121

Shahiri, A., W. Husain y N. Rashid (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

Usman, U. I., Salisu, A., Barroon, A. I. y Yusuf, A. U. (2019). A Comparative Study of Base Classifiers in Predicting Students' Performance based on Interaction with LMS Platform. *FUDMA Journal of Sciences, 3*(1), 231-239.

Vásconez, G. E. (2023). Modelo de predicción de deserción escolar en los estudiantes de la unidad educativa Los Andes por impacto de la pandemia. *Ciencia Latina Revista Científica Multidisciplinar, 7*(1), 3038-3052. https://doi.org/10.37811/cl_rcm.v7i1.4640

Wang, X., Zhou, C. y Xu X. (2019). Application of C4.5 decision tree for scholarship evaluations. *Procedia Computer Science, 151* (2019). 179-184. https://doi.org/10.1016/j.procs.2019.04.027