Revista Internacional de Investigación e Innovación Tecnológica

Página principal: www.riiit.com.mx

Aplicación de Técnicas de Minería de Datos para Clasificación: Un Caso de Estudio en la Educación Superior

Application of Data Mining Techniques for Classification: A Case Study in Higher Education

Chávez-Vega, N.B.[1], Pérez-Olguín, I.J.C.[2]*, Luviano-Cruz, D.[3], Portillo-Escobedo, A.[4]

[1] Departamento de Mecatrónica; Universidad Tecnológica de Chihuahua; C.P.31216; Chihuahua, México. https://orcid.org/0000-0001-8868-4387.
[2] Instituto de Ingeniería y Tecnología; Universidad Autónoma de Ciudad Juárez; C.P.32315; Chihuahua, México. https://orcid.org/0000-0003-2445-0500.
[3] Instituto de Ingeniería y Tecnología; Universidad Autónoma de Ciudad Juárez; C.P.32315; Chihuahua, México. https://orcid.org/0000-0002-4778-8873.
[4] División de Estudios de Posgrado e Investigación; Instituto Tecnológico de Chihuahua; C.P.31200; Chihuahua, México. https://orcid.org/0000-0003-3309-4720.
nchavez@utch.edu.mx; ivan.perez@uacj.mx*; david.luviano@uacj.mx; alberto.ep@chihuahua.tecnm.mx

Innovación tecnológica: Clasificación de alumnos vulnerables con técnicas de minería de datos.

Área de aplicación industrial: Educación.

Recibido: 03 marzo 2023
Aceptado: 11 diciembre 2023

Abstract

In Mexico, the EXANI_II assessment instrument has been designed to comprehensively assess academic skills and specific knowledge of applicants to enter higher education. The application of data mining techniques, such as decision trees for classification, can support the detection of vulnerable students. This research compares two decision trees: The J-48 algorithm and the random tree algorithm using the Weka software for its implementation in the EXANI-II database for the evaluation of applicants to enter the Technological University of Chihuahua in 2021. The results are reviewed regarding the accuracy in the classification obtained in each of the algorithms, with the J-48 algorithm having a better performance.

Keywords: Educative data mining, J-48 classification algorithm, random tree, Weka software application.

Resumen

En México, el instrumento de evaluación EXANI_II ha sido diseñado para evaluar integralmente habilidades académicas y conocimientos específicos de los aspirantes a ingresar a la educación superior. La aplicación de técnicas de minería de datos, como lo son los árboles de decisión para clasificación puedan apoyar en la detección de alumnos vulnerables. Esta investigación compara dos árboles de decisión: El algoritmo J-48 y el algoritmo de árbol aleatorio utilizando el software Weka para su implementación en la base de datos EXANI-II generada de la evaluación de aspirantes a ingresar a la Universidad Tecnológica de Chihuahua en el 2021. Se revisan los resultados en cuánto a la exactitud en la clasificación obtenida en cada uno de los algoritmos, teniendo un mejor desempeño el algoritmo J-48.

Palabras clave: Minería de datos educativos, algoritmo de clasificación J-48, árbol aleatorio, aplicación informática Weka.

1. Introducción

Las instituciones educativas tienen como finalidades la transmisión de valores, creencias, conocimientos y el desarrollo de habilidades técnicas y para la vida en sus estudiantes (Alfonzo y Pedagógica Experimental Libertador Venezuela, 2018). La efectividad con que se cumplen dichos propósitos es un factor determinante de la calidad educativa, la transformación y el progreso académico, lo que implica la adaptación de los centros educativos a la realidad social, sus necesidades presentes y, sobre todo, futuras (Jiménez-Cruz, 2019). En consonancia con lo anterior, la identificación precisa de fortalezas y deficiencias en el proceso enseñanza-aprendizaje es fundamental ya que, además, mejora las probabilidades de que los estudiantes finalicen su programa académico, y disminuya la deserción escolar (Herrera Rivas y Roque Hernández, 2019). El campo de la minería de datos educativos permite la exploración y análisis de grandes conjuntos de datos generados en entornos escolares, dando pauta a la toma de decisiones con respecto a la gestión de la participación en el aula, los métodos de enseñanza implementados y la predicción de estudiantes en riesgo de reprobación, entre otras. Actualmente, en México ya existen investigaciones direccionadas en este sentido, como lo es el trabajo de (Ayala et al., 2021), en donde se busca predecir a estudiantes vulnerables académicamente a partir del análisis de variables cualitativas y cuantitativas del estudiante con apoyo de clasificadores con árboles de decisión y de esta manera lograr proponer estrategias de intervención educativa oportuna.

El estudio propuesto en este articulo tiene como objetivo comparar la exactitud en la clasificación y análisis entre el algoritmo J-48 y el árbol de decisión aleatorio analizando a alumnos de nuevo ingreso de la Universidad Tecnológica de Chihuahua con la finalidad de determinar estrategias didácticas como lo es el uso de tutores inteligentes para disminuir los índices de reprobación en matemáticas.

Estos algoritmos fueron seleccionados dado que ya se han encontrado estudios similares en la búsqueda de literatura, (Aslam et al., 2021; Hamoud et al., 2018) mostrando estabilidad, precisión, velocidad y facilidad en la interpretabilidad de los resultados.

Un medio para identificar el nivel de dominio logrado sobre determinados conocimientos y

habilidades, respecto a un estándar y/o grado académico, son las pruebas de logro académico (Wild Santamaría et al., 2015), instrumentos cuyo fin es caracterizar a una población estudiantil, comparar el rendimiento entre instituciones educativas y determinar el ingreso, o no, a un nuevo nivel de estudios. El Centro Nacional de Evaluación para la Educación Superior (CENEVAL), asociación civil, que diseña y aplica los Exámenes Nacionales de Ingreso (EXANI) para evaluar "…conocimientos, habilidades y competencias; así como el análisis y la difusión de sus resultados". (González-Marrón et al., 2017; Perfil Institucional - Ceneval, 2022), al respecto, dichos instrumentos son empleados por instituciones de educación superior, como criterio de admisión a sus planes y programas, no obstante, un porcentaje significativo de estudiantes es rechazado debido a que "…la calidad educativa en el nivel medio superior podría no estar generando las competencias requeridas para que las personas continúen estudios posobligatorios". En la tabla 1 se observan los resultados del área de matemáticas mostrando un porcentaje elevado de estudiantes con resultados de desempeño insatisfactorios (CENEVAL, 2018); aunado a este factor de bajo rendimiento académico, existe también una problemática con la disponibilidad y la accesibilidad a la educación superior que propicia que solo seis de cada diez jóvenes se matriculen en dichas instituciones (Consejo Nacional de Evaluación de la Política de Desarrollo Social, 2018, pág. 25).

Tabla 1. Extracto de tabla de resultados del examen nacional de ingreso del año 2018, Ceneval.

	Variable	Categoría	Población	% población sustentante	Matemáticas		
					Sin dictamen %	Nivel desempeño insatisfactorio %	Nivel desempeño satisfactorio %
Nacional	Sexo	Hombres	9485	57.34	0.03	29.31	28.00
		Mujeres	7046	42.59	0.03	19.77	22.79
		Respuesta no válida	11	0.07	0.00	0.05	2.00
	Régimen	Público	13905	84.06	0.05	41.17	42.84
		Privado	2604	15.74	0.01	7.86	787.00
		Respuesta no válida	33	0.20	0.00	0.10	0.10
	Modalidad	Bachillerato general	11079	66.97	0.04	31.99	3495.00
		Bachillerato tecnológico	3974	24.02	0.01	12.12	11.90
		Profesional técnico	1171	7.08	0.01	4.02	305.00
		Bachillerato intercultural	16	0.10	0.00	0.04	0.05
		Bachillerato internacional	94	0.57	0.00	0.13	0.44
		Telebachillerato	173	1.05	0.00	0.73	32.00
		Respuesta no válida	35	0.21	0.00	0.11	0.10
	Promedio	6-6.9	413	250.00	0.00	1.52	98.00
		7-7.9	4130	24.97	0.02	14.76	10.19
		8-8.9	8282	50.07	0.03	24.33	257.00
		9-9.9	3400	20.55	0.01	7.51	13.03
		10	87	0.53	0.00	0.14	0.39
		Respuesta no válida	230	1.39	0.00	0.88	0.51
	Total		16542	100.00	0.06	49.14	50.80

La minería de datos (*Data Mining*) es un campo de la estadística, asistido por computadora, que estudia la extracción de información en fuentes masivas de datos, que permite identificar patrones, correlaciones y anomalías, para la toma de decisiones e incluso, para predecir resultados (Oviedo Carrascal y Jiménez Giraldo, 2019). Dicho proceso se puede aplicar en la detección de patrones en respuestas de estudiantes (*V.gr*. áreas de conocimiento vulnerables), para fortalecer los procesos de enseñanza-aprendizaje. La aplicación de la minería de datos en el ámbito educativo es la disciplina que busca desarrollar nuevos métodos de exploración de la información que se genera al interior de las instituciones educativas, para establecer la forma en que los estudiantes aprenden, y con ello, mejorar el proceso educativo (Alveiro et al., 2019).

En minería de datos, los árboles de decisión son una técnica, o modelo informático analítico, de clasificación, segmentación y predicción, basada en aprendizaje supervisado no paramétrico (Uvidia Fassler et al., 2018), que puede ser empleada para el diagnóstico de evaluaciones de rendimiento académico y así, beneficiar la propuesta de estrategias de mejora. De acuerdo con Kai *et al.* (2017), el uso de los modelos J-48 y J-Rip para generar árboles de decisión es pertinente para realizar predicciones de comportamiento de estudiantes en cursos virtuales, por ejemplo, a través de factores como la cantidad y/o frecuencia de revisión de calificaciones, y el total de vistas de mensajes en línea o de foros de discusión.

La predicción temprana del fracaso estudiantil es necesaria para la toma de decisiones y la generación de estrategias pedagógicas de apoyo a la retención de alumnos. En un estudio realizado por Mishra *et al*., (2014), en la educación superior, se consideró el clasificador J-48 y el árbol de decisión aleatoria, para el análisis de 25 atributos que incluyen variables sociales, académicas y emocionales de estudiantes matriculados en el programa de Maestría de Computación Aplicada de la Universidad Guru Gobind Singh Indraprastha. Se aplicó un cuestionario a una muestra de 250 estudiantes y se evaluó su desempeño académico de tercer semestre y sus habilidades emocionales a través del estándar de evaluación de habilidades emocionales (ESAP, por sus siglas en inglés). Los resultados evidenciaron el rendimiento de los algoritmos: el árbol de decisión aleatoria con una exactitud del 94,418%, en comparación con el modelo J-48 con 88,372% de exactitud. Al igual que en esta investigación, el trabajo que se presenta en este artículo trata de demostrar el desempeño de los árboles de decisión como clasificadores en la minería de datos educativos.

Con la minería de datos es posible identificar patrones que ayuden a los estudiantes a elegir el programa académico adecuado, según su perfil y habilidades. En este tema particular, estudios como el de Patacsil (2020) muestran que, el promedio académico, en programas de matemáticas e ingeniería es una variable predictora dominante, a diferencia de programas de ciencias blandas como el inglés, en donde el promedio académico no tiene un aporte significativo para la predicción de patrones. En el resultado obtenido de la evaluación del desempeño en exactitud del algoritmo J-48, el bosque aleatorio y el árbol de decisión aleatorio, el primero de los tres tuvo mayor exactitud.

2. Materiales y métodos

La metodología consta de varias etapas que se desarrollaron con apoyo del software Weka versión 3.8.6, y van desde la selección de variables académicas definidas en la base de datos EXANI-II, preprocesamiento de datos utilizando filtros para eliminar valores atípicos y balanceo de datos; aplicación del

algoritmo J-48 y modelo de árbol aleatorio para clasificación y análisis de resultados (Ver figura 1).

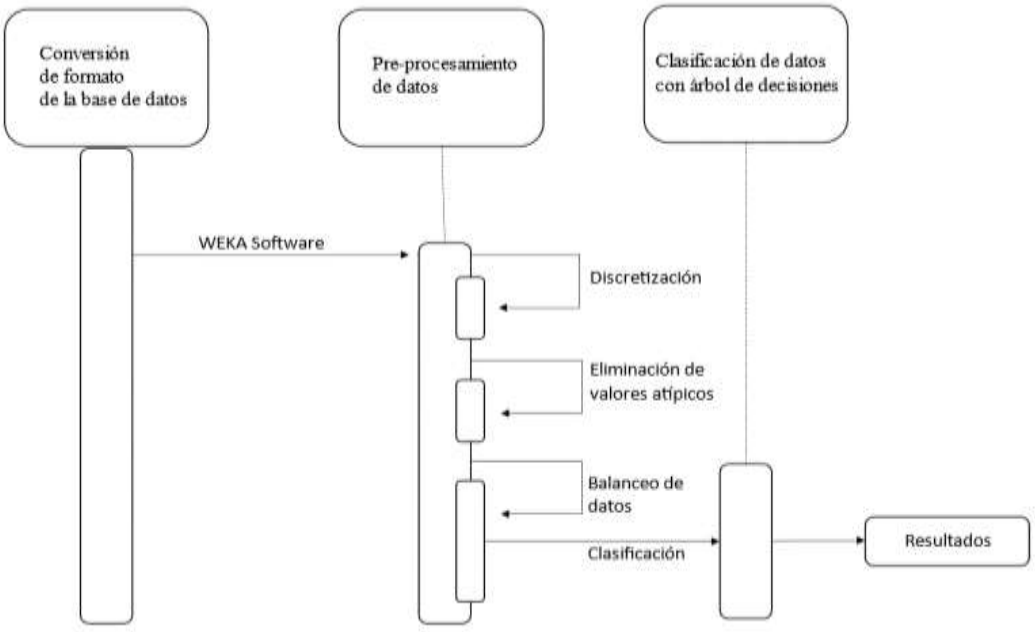

Figura 1. Diagrama secuencial para clasificación de alumnos con técnicas de minería de datos.

2.1. Análisis y selección de variables

El rendimiento académico de los estudiantes depende de factores de diversa índole, como variables académicas, psicológicas, económicas y sociales. En esta investigación se seleccionan índices de puntuación establecidos en el examen EXANI-II donde, a través de un conjunto de preguntas, se evalúan áreas de física, cálculo, comprensión lectora y razonamiento lógico-matemático.

La Tabla 2 muestra el rango de valores de cada variable académica, así como una breve descripción de lo que evalúa.

Tabla 2. Variables académicas obtenidas del EXANI-II.

Variable	Descripción	Rango de valor
ICNE	Puntuación en el índice de la prueba de admisión.	700-1300
PORCENTAJE	Percentil de la prueba de admisión.	0-100
IMOD1	Puntuación en el módulo de Cálculo.	700-1300
IMOD2	Puntuación en el módulo de Física.	700-1300
ICLE	Calificación de comprensión lectora.	700-1300
IRIN	Calificación de escritura indirecta.	700-1300
IPMA	Puntuación de pensamiento matemático.	700-1300
CMPPMA	Puntuación de comprensión matemática en porcentaje de respuestas correctas.	0-100
PPMA MA	Puntaje de Matemáticas en Porcentaje de Respuestas Correctas.	0-100

Elaboración de los autores con información de la Universidad Tecnológica de Chihuahua, 2021.

2.2. Ajuste del conjunto de datos

Para analizar los datos se empleó el software especializado en aprendizaje automático; Weka, una colección de algoritmos de aprendizaje automático para tareas de minería de datos, creada por la Universidad de Waikato en Nueva Zelanda bajo una Licencia Pública General (Markov y Russell, 2006).

Para leer los datos del software (ver Tabla 3), fue necesario convertir el formato de valores separados por comas (CSV), a un formato de tipo de texto ASCII que describe una lista de instancias con atributos comunes (ARFF).

Tabla 3. Muestra de la base de datos de evaluación EXANI II.

Viewer

Relation: datadat

No.	1: SEXO Nominal	2: MOD_BAC Nominal	3: ICNE Numeric	4: PERCEN Numeric	5: PPMA_CM Numeric	6: PPMA_MA Numeric	7: IMOD1 Numeric	8: IMOD2 Numeric	9: ICLE Numeric	10: IRIN Numeric	11: IPMA Numeric
1	Mujer	bachillerato_general	1151.0	100.0	83.33	72.73	1150.0	1050.0	1140.0	1220.0	1176.0
2	Mujer	bachillerato_tecnologico	1138.0	99.69	83.33	54.55	1075.0	1075.0	1180.0	1200.0	1134.0
3	Hombre	bachillerato_tecnologico	1134.0	99.38	77.78	45.45	1200.0	1150.0	1100.0	1140.0	1093.0
4	Hombre	bachillerato_general	1120.0	99.06	55.56	36.36	1175.0	1100.0	1140.0	1200.0	990.0
5	Hombre	bachillerato_tecnologico	1116.0	98.75	55.56	63.64	1175.0	1025.0	1120.0	1200.0	1052.0
6	Mujer	bachillerato_tecnologico	1116.0	98.75	77.78	45.45	1025.0	1025.0	1180.0	1220.0	1093.0
7	Mujer	bachillerato_tecnologico	1112.0	98.13	72.22	72.73	1150.0	1025.0	1160.0	1080.0	1134.0
8	Hombre	bachillerato_tecnologico	1112.0	98.13	77.78	45.45	1100.0	1125.0	1120.0	1120.0	1093.0
9	Mujer	bachillerato_general	1107.0	97.5	61.11	54.55	1150.0	1050.0	1140.0	1140.0	1052.0
10	Hombre	bachillerato_general	1107.0	97.5	72.22	45.45	1125.0	1000.0	1140.0	1180.0	1072.0
11	Mujer	bachillerato_general	1107.0	97.5	55.56	36.36	1100.0	1025.0	1220.0	1180.0	990.0
12	Hombre	bachillerato_general	1103.0	96.56	66.67	63.64	1125.0	1025.0	1120.0	1140.0	1093.0

2.3. Preparación para el análisis de datos

Trabajar con datos reales la mayoría de las veces implica operaciones de preprocesamiento, como limpieza, discretización y balanceo de datos. La etapa de discretización reduce el número de valores de un atributo continuo al dividir el rango del atributo en intervalos (Rajalakshmi et al., 2016). Esto conduce a un nivel de conocimiento conciso y fácil de usar y se puede aplicar antes o después de la extracción de datos. En este caso, se optó por aplicarlo antes para preparar la información para la aplicación de las siguientes etapas de acondicionamiento. Para discretizar los datos con el parámetro del mismo ancho del intervalo, se aplica la Ecuación 1.

$$\delta = \frac{X_{Max} - X_{Min}}{k} \quad \textbf{(Ec. 1)}$$

Dónde:
k: Indica el número de valores observados.
X_{Max}: Indica el valor máximo de las observaciones o datos.
X_{Min}: Indica el valor mínimo de las observaciones o datos.
δ: Indica el ancho del intervalo.

Los valores atípicos son patrones que no están en el rango de comportamiento normal de los datos, y es un problema concurrente cuando se aplican técnicas de minería de datos. A este respecto, para trabajar solo con valores dentro del rango normal, se utiliza un filtro de atributos no supervisado llamado Rango Intercuartil. Este tipo de filtro ubica valores atípicos en una determinada sección de datos mediante la creación de una ventana de tamaño "*k*". Una vez que se localizan los valores atípicos, se pueden eliminar para obtener mejores resultados en el proceso de aprendizaje automático (Vinutha et al., 2018).

El filtro de rango intercuartil integrado en el software Weka se utiliza para realizar mediciones de variabilidad en los datos. La metodología consiste en ordenar de forma ascendente los datos y luego dividirlos en cuatro secciones de las mismas dimensiones. Los valores que seccionan la información se denominan cuartiles (Q_1, Q_2, Q_3).

En la Ecuación 2, el rango intercuartílico se define como:

$$I_{QR} = Q_3 - Q_1 \quad \textbf{(Ec. 2)}$$

Las Ecuaciones 3 y 4 calculan los valores atípicos.

$$Q_3 + O_f * I_{QR} < x \leq Q_3 + E_{vf} * I_{QR} \quad \textbf{(Ec. 3)}$$

$$Q_1 - E_{vf} * I_{QR} < x \leq Q_1 - O_f * I_{QR} \quad \textbf{(Ec. 4)}$$

Aplicando las ecuaciones 5 y 6 que están implementadas en el diseño del filtro no supervisado "*interquartilerange*" que provee el software Weka, se obtienen los valores extremos.

$$x > Q_3 + E_{vf} * I_{QR} \quad \textbf{(Ec. 5)}$$

$$x < Q_1 - E_{vf} * I_{QR} \quad \textbf{(Ec. 6)}$$

Dónde:

Q_1: percentil 25 de los datos.
Q_2: percentil 50 de los datos.
Q_3: percentil 75 de los datos.
I_{QR}: Rango intercuartil.
O_f: Factor atípico.
E_{vf}: Factor de valor extremo.
X: Conjunto de datos.

El desequilibrio de datos genera problemas por la disparidad de clases en las variables, provocando un sesgo en el clasificador hacia las clases mayoritarias, ya que el algoritmo tiende a aprender más sobre las clases que tienen más ejemplos en el conjunto de datos. La elección de la estrategia específica para resolver este conflicto puede depender del tamaño del conjunto de datos, la naturaleza del problema y las características particulares de las clases minoritarias. Para corregir este problema, se aplica un filtro supervisado por instancia llamado Técnica de Sobremuestreo de Minorías Sintéticas (SMOTE), para crear nuevas muestras de las clases minoritarias o categorías que tienen menos instancias en comparación con otras en el conjunto de datos y, de esta manera, compensarlas y distribuirlas más uniformemente (Katore, 2015).

El filtro selecciona uno de sus *k* vecinos más cercanos a *X*, que es parte de la clase minoritaria, luego encuentra un nuevo patrón *Z* (ecuación 7) en un punto aleatorio en el segmento de línea que conecta el patrón y el vecino seleccionado (Elreedy y Atiya, 2019).

$$Z = X_0 + w(X - X_0) \quad \textbf{(Ec. 7)}$$

Dónde:
w: Es una variable aleatoria uniforme en el rango [0,1].
X_0: Representa los datos de la clase minoritaria.
X: Uno de los *k* vecinos más cercanos.

2.4. Clasificación de datos con algoritmo J-48 y árbol aleatorio

El algoritmo J-48 se utiliza para la creación de árboles de decisión que luego pueden encontrar nuevo conocimiento de patrones. El proceso de trabajo del algoritmo comienza con la conversión de un dato en un árbol de decisión soportado por reglas de decisión.

Para construir el árbol de decisión es necesario tener una base de datos de entrenamiento. Una vez definida la base de datos de entrenamiento, se selecciona un atributo para iniciar la primera división y a partir de esta operación se crea una rama para cada uno de los valores. Este proceso se

realiza recursivamente en orden descendente (Masrur et al., 2019).

La evaluación del desempeño de un modelo de clasificación se basa en el número de predicciones correctas e incorrectas. Esta información se registra en la matriz de confusión (ver Tabla 4). Para problemas de clasificación binaria, las predicciones correctas se obtienen de la suma de la posición de los elementos en F_{11} y F_{00}, y las predicciones incorrectas de la suma de F_{10} y F_{01} (Visa et al., 2011).

Tabla 4. Matriz de confusión.

	Predecido Positivo	**Predecido Negativo**
Actual Positivo	F_{00}	F_{01}
Actual Negativo	F_{10}	F_{11}

Para calcular la exactitud, se define en la Ecuación 8, como la suma de los elementos de la diagonal principal dividida por todos los elementos de la matriz de confusión.

$$\text{Exactitud}=\frac{F_{11}+F_{00}}{F_{11}+F_{00}+F_{10}+F_{01}} \quad \textbf{(Ec. 8)}$$

3. Discusión de Resultados

Se aplicaron los algoritmos J-48 y árbol aleatorio en 2228 instancias obtenidas del EXANI-II, utilizando una validación cruzada de diez iteraciones (ver tabla 5). Los resultados estadísticos obtenidos en comparación de la exactitud esperada con la exactitud observada (kappa) muestran valores de 0,45 y 0,41, respectivamente, que según la escala propuesta por Landis y Koch (Landis & Koch, 1977), los coeficientes indican concordancia moderada.

Tabla 5. Comparativa de rendimiento de los algoritmos de árbol aleatorio y algoritmo J-48.

Criterio	**Árbol aleatorio**	**Algoritmo J-48**
Instancias clasificadas correctamente	1689/2228 (75.8078%)	1724/2228 (77.3788%)
Instancias clasificadas incorrectamente	539/2228 (24.1921%)	504/2228 (22.6212%)
Valor estadístico kappa	0.4132	0.4515
Error absoluto medio	0.2432	0.2672
Error raíz cuadrático medio	0.4643	0.4081
Error relativo absoluto	57.5727	63.2486
Error raíz cuadrático relativo	101.0343	88.7963

El clasificador categoriza correctamente el 77,37% de las instancias con el modelo J-48, y el 75,80% con el árbol aleatorio. En este caso de estudio, los algoritmos generaron resultados satisfactorios con muy poca variabilidad entre ellos, por lo que se considera que ambos son pertinentes para clasificar datos afines a los atributos académicos que ofrece el EXANI-II.

Los árboles de decisión son algoritmos estadísticos o técnicas de aprendizaje automático, que construyen modelos predictivos de análisis de datos para grandes cantidades de información, en función de su clasificación, según determinadas características o propiedades.

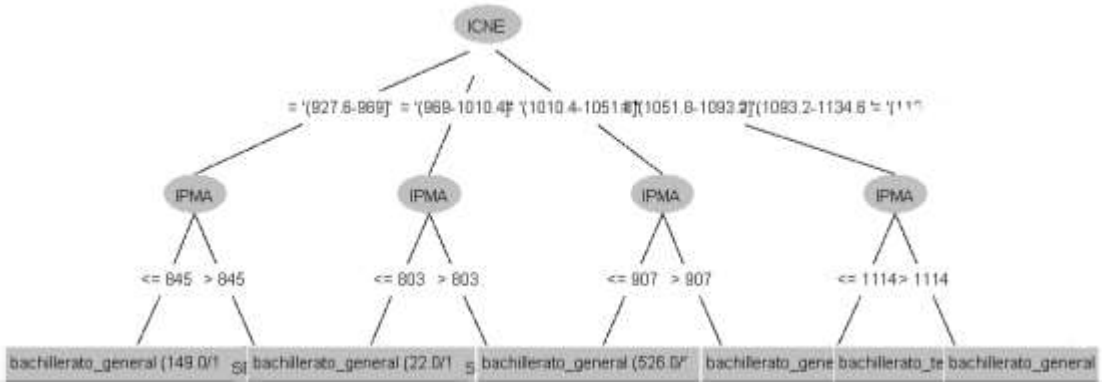

Figura 2. Extracto de diagrama de árbol generado con el algoritmo J-48.

En la figura 2 se visualiza una sección del árbol de decisión generado con el algoritmo J-48, en donde sigues el camino desde el nodo raíz (nodo principal superior), hasta una hoja, tomando decisiones en cada nodo interno según los valores que describen la variable académica evaluada para la toma de decisión y posterior ramificación.

Comparando los resultados obtenidos en esta investigación con los de otros autores (Kai et al., 2017; Mishra et al., 2014; Patacsil, 2020) los árboles de decisión analizados muestran un buen desempeño como clasificadores de datos.

En la tabla 6, se despliegan las métricas calculadas con el software Weka para validar exactitud, precisión, sensibilidad, tasas de verdaderos y falsos positivos, tanto de estudiantes que provienen de bachillerato general como para alumnos que egresaron de bachillerato técnico.

Tabla 6. Comparativa de métricas obtenidas en los clasificadores.

Árbol de decisión	Clases	Tasa de verdaderos positivos	Tasa de falsos positivos	Precisión	Sensibilidad	Medida F
Algoritmo J-48	Bachillerato General	0.856	0.415	0.826	0.856	0.841
	Bachillerato Técnico	0.585	0.144	0.638	0.585	0.611
Árbol aleatorio	Bachillerato General	0.845	0.441	0.815	0.845	0.83
	Bachillerato Técnico	0.559	0.155	0.61	0.559	0.583

4. Conclusiones

La reprobación de alumnos en la educación superior es una gran problemática para mantener la calidad de los programas educativos. La predicción temprana de alumnos vulnerables genera oportunidades para aplicar estrategias didácticas que apoyen a mejorar el desempeño académico en los estudiantes de bajo rendimiento.

Este trabajo se centró únicamente en variables académicas consideradas por el EXANI-II y se aplicó a estudiantes de nuevo ingreso en la Universidad Tecnológica de Chihuahua. El algoritmo J-48 dio una mayor exactitud que el árbol de decisión aleatorio. El uso de técnicas de minería de datos como lo son los árboles de decisión beneficia con la generación de retroalimentación que se puede considerar para el diseño de tutores inteligentes en cuanto a patrones en los procesos de aprendizaje y caracterización de

alumnos en función de su desempeño académico, influyendo en la toma de decisiones para selección de estrategias de enseñanza que fortalezcan las áreas débiles de cada alumno.

En investigaciones futuras se pretende incluir parámetros emocionales para analizar qué otros factores afectan el desempeño del estudiante y, posteriormente, desarrollar un sistema tutor inteligente para apoyo en la enseñanza de las matemáticas que incida en la reducción de índices de reprobación.

5. Agradecimientos

Esta investigación no se hubiera realizado sin el apoyo de la Universidad Tecnológica de Chihuahua, por permitir el acceso a la base de datos EXANI II de aspirantes a ingreso a su universidad.

6. Referencias

Alfonzo, M., & Pedagógica Experimental Libertador Venezuela, U. (2018). *Gestión del Conocimiento e Instituciones Educativas*. *IV*.

Alveiro, M., Rosado Gómez, A., Alejandra, E., & Ibáñez, V. (2019). *DATA MINING APPLICATION IN THE VIRTUAL EDUCATION*.

Aslam, S. M., Jilani, A. K., Sultana, J., & Almutairi, L. (2021). Feature Evaluation of Emerging E-Learning Systems Using Machine Learning: An Extensive Survey. In *IEEE Access* (Vol. 9, pp. 69573–69587). Institute of Electrical and Electronics Engineers Inc. https://doi.org/10.1109/ACCESS.2021.3077663

CENEVAL. (2018). *Resultados del Examen Nacional de Ingreso a la Educación Superior en el año 2018*.

Elreedy, D., & Atiya, A. F. (2019). A Comprehensive Analysis of Synthetic Minority Oversampling Technique (SMOTE) for handling class imbalance. *Information Sciences*, *505*, 32–64. https://doi.org/10.1016/j.ins.2019.07.070

Franco, E. A., López Martínez, R. E., Hugo, V., & Domínguez, M. (n.d.). Predictive models of academic risk in computing careers with educational data mining. *Revista de Educación a Distancia. Núm*, *66*, 30–2021. https://doi.org/10.6018/red

Gonzalez-Marron, D., Enciso-Gonzalez, A., Karen Hernandez-Gonzalez, A., Gutierrez-Franco, D., Guizar-Barrera, B., & Marquez-Callejas, A. (2017). Evaluación de parámetros de encuesta de ingreso del CENEVAL para alumnos candidatos a ingresar al nivel superior, caso de estudio ITP Evaluation of CENEVAL Admission Surveyed Parameters for Students that are Candidates to Enter the Higher Education, ITP Study Case. In *Research in Computing Science* (Vol. 139).

Hamoud, A. K., Hashim, A. S., & Awadh, W. A. (2018). Predicting Student Performance in Higher Education Institutions Using Decision Tree Analysis. *International Journal of Interactive Multimedia and Artificial Intelligence*, *5*(2), 26. https://doi.org/10.9781/ijimai.2018.02.004

Herrera Rivas, H., & Roque Hernández, R. V. (2019). Brecha digital, idioma inglés y su vínculo con la comprensión lectora en español. *RIDE Revista Iberoamericana Para La Investigación y El Desarrollo Educativo*, *10*(19). https://doi.org/10.23913/ride.v10i19.555

Jimenez-Cruz, J. (2019). Transformando la educación desde la gestión educativa: hacia un cambio de mentalidad. *Praxis*, *15*(2), 223–235. https://doi.org/10.21676/23897856.2646

Kai, S., Miguel, J., Andres, L., Paquette, L., Baker, R. S., Molnar, K., Watkins, H., & Moore, M. (2017). *Predicting Student*

Retention from Behavior in an Online Orientation Course.

Katore, L. S. (2015). Comparative Study of Recommendation Algorithms and Systems using WEKA. In *International Journal of Computer Applications* (Vol. 110, Issue 3).

Landis, J. R., & Koch, G. G. (1977). *This content downloaded from 128.230.234.162 on Fri* (Vol. 33, Issue 1).

Markov, Z., & Russell, I. (2006). *An Introduction to the WEKA Data Mining System*. https://doi.org/10.1145/1140123.1140127

Masrur, A., Riyanarto, S., & Kelly, R. S. (2019). *2019 International Seminar on Application for Technology of Information and Communication (iSemantic)*. IEEE.

Mishra, T., Kumar, D., & Gupta, S. (2014). Mining students' data for prediction performance. *International Conference on Advanced Computing and Communication Technologies, ACCT*, 255–262. https://doi.org/10.1109/ACCT.2014.105

Oviedo Carrascal, A. I., & Jiménez Giraldo, J. (2019). Minería de datos educativos: Análisis del desempeño de estudiantes de ingeniería en las pruebas SABER-PRO. *Revista Politécnica*, *15*(29), 128–140. https://doi.org/10.33571/rpolitec.v15n29a10

Patacsil, F. F. (2020). Predicting University Students' Academic Success Using Different Tree Classifiers and Ensemble Approaches to Suggest Suitable Program. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9*, 2. www.ijstr.org

Perfil institucional - Ceneval. (n.d.). Retrieved December 6, 2023, from https://ceneval.edu.mx/sobre_el_ceneval-perfil_institucional/

Rajalakshmi, A., Vinodhini, R., Fathima Bibi, K., & Research Scholar, M. P. (2016). *Data Discretization Technique Using WEKA Tool*. www.ijcset.net

Uvidia Fassler, M., Cisneros Barahona, A., Naranjo, P. M., & Villa Yánez, H. (2018). Minería de datos para la toma de decisiones en la unidad de nivelación y admisión universitaria ecuatoriana Data mining for decision-making in the ecuadorian university leveling and admission unit. In *REVISTA CIENTÍFICA Revista Cumbres* (Vol. 4). http://investigacion.utmachala.edu.ec/revistas/index.php/Cumbres

Vinutha, H. P., Poornima, B., & Sagar, B. M. (2018). Detection of outliers using interquartile range technique from intrusion dataset. *Advances in Intelligent Systems and Computing*, *701*, 511–518. https://doi.org/10.1007/978-981-10-7563-6_53

Visa, S., Ramsay, B., Ralescu, A., & van der Knaap, E. (2011). *Confusion Matrix-based Feature Selection*.

Wild Santamaría, G., Barrios Mendoza, S., Berlanga Reséndiz, K., & Hernández Castillo, L. (2015). *Estudio analítico del impacto de la información del cuestionario de contexto del Exani-II en el índice CENEVAL*. https://www.eumed.net/rev/tectzapic/2015/01/ceneval.html