Revista Amazónica de Ciencias Básicas y Aplicadas
Vol. 1 Núm. 2: e190 (2022)
https://doi.org/10.55873/racba.v1i1.190
e-ISSN: 2810-8736
Universidad Nacional Amazónica de Madre de Dios

Artículo original / Original article

Técnicas supervisadas de minería de datos para el análisis del rendimiento académico de estudiantes universitarios en la Amazonia peruana

Supervised data mining techniques for the analysis of academic performance of university students in the Peruvian Amazon

Luis Alberto Holgado-Apaza [1*]; Nelly Jacqueline Ulloa-Gallardo [1]; Yban Vílchez-Navarro [1]; Diego Raúl Quispe-Barra [1]

[1]Universidad Nacional Amazónica de Madre de Dios, Puerto Maldonado, Perú

Recibido: 08/04/2022
Aceptado: 21/06/2022
Publicado: 25/07/2022

*Autor de correspondencia: nulloa@unamad.edu.pe

Resumen: El presente estudio tuvo como propósito identificar la técnica supervisada de minería de datos con mejor desempeño para el análisis del rendimiento académico de estudiantes universitarios. Se optó por el diseño no experimental de corte transversal. El conjunto de datos inicial para los experimentos estuvo conformado por 17771 registros de procesos académicos, tras el preprocesamiento se obtuvo un conjunto de datos final de 17035 registros. La metodología de minería de datos empleada fue Knowledge Discovery in Databases (KDD). Se emplearon las técnicas de regresión logística binaria, Classification and Regression Trees (CART), C4.5, Máquinas de soporte vectorial, K-vecinos más cercanos. Los resultados demuestran que el algoritmo C5.0 obtiene una exactitud del 93%, área bajo la curva (AUC) del 0,98 y un tiempo de entrenamiento de 0,87 segundos, resultando ser el más eficiente en relación con los demás algoritmos comparados.

Palabras clave: algoritmos supervisados; aprendizaje automático; KDD; rendimiento académico; técnicas predictivas

Abstract: The purpose of this study was to identify the supervised data mining technique with the best performance for the analysis of the academic performance of university students. The non-experimental cross-sectional design was chosen. The initial data set for the experiments consisted of 17,771 records of academic processes, after preprocessing a final data set of 17,035 records was obtained. The data mining methodology used was Knowledge Discovery in Databases (KDD). Binary logistic regression techniques, Classification and Regression Trees (CART), C4.5, Support Vector Machines, K-nearest neighbors were used. The results show that the C5.0 algorithm obtains an accuracy of 93%, an area under the curve (AUC) of 0.98 and a training time of 0.87 seconds, turning out to be the most efficient in relation to the other algorithms compared.

Keywords: machine learning; supervised algorithms; KDD; academic performance; predictive techniques

Cómo citar / Citation: Holgado-Apaza, L. A., Ulloa-Gallardo, N. J., Vílchez-Navarro, Y. & Quispe Barra, D. R. (2022). Técnicas supervisadas de minería de datos para el análisis del rendimiento académico de estudiantes universitarios en la Amazonia peruana. *Revista Amazónica de Ciencias Básicas y Aplicadas*, 1(2), e190.
https://doi.org/10.55873/racba.v1i1.190

1. Introducción

El rendimiento académico ha sido por décadas un tema de estudio, actualmente constituye uno de los temas importantes y trascendentales en la investigación educativa que haciendo uso de las herramientas y técnicas tradicionales será difícil comprenderlos y analizarlo en su real magnitud, en este sentido es necesario un análisis empleando técnicas cercanas al aprendizaje automático y minería de datos (Cano Celestino & Robles Rivera, 2018). En relación a ello Norabuena Penadillo (2011), manifiesta que en una sociedad de la información como la que estamos atravesando, uno de los grandes desafíos de la educación en todos sus niveles, es transformar la gran cantidad de información disponible en conocimiento con fines de mejorar la toma de decisiones.

La minería de datos se define como el conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de forma automática o semiautomática, con el objetivo de encontrar patrones, tendencias o reglas repetitivas que expliquen el comportamiento de los datos en un contexto dado (Enke & Thawornwong, 2005; Trakunphutthirak & Lee, 2022; Khor, 2022).

La minería de datos en el sector educativo o minería de datos educativos es un tema emergente debido a la gran cantidad de datos que se generan diariamente en las instituciones de educación básica y superior públicas o privadas de cualquier país. La minería de datos educativos se centra en el descubrimiento de conocimientos de todas las bases de datos educativas generadas por individuos y grupos de individuos apoyados en marcos institucionales (Lemay et al., 2021; Nabil et al., 2022). Los últimos avances en la minería de datos permiten la extracción de conocimiento con fines de mejorar la calidad del proceso educativo (Asif et al., 2017).

Según Han et al. (2012) las técnicas de minería de datos se clasifican en: técnicas predictivas o supervisadas y técnicas descriptivas. Los algoritmos predictivos o supervisados permiten predecir el valor de un atributo (etiqueta) de un conjunto de datos, conociendo otros atributos (atributos descriptivos). A partir de los datos cuya etiqueta se conoce, se obtiene una relación entre esa etiqueta y otro conjunto de atributos (Han et al., 2012).

Estas relaciones se utilizan para hacer la predicción en datos cuya etiqueta se desconoce. Según Rosado Gómez & Verjel Ibáñez (2015) las técnicas predictivas tienen las tareas de clasificación y regresión. Las tareas de regresión buscan obtener un modelo que permita predecir el valor numérico de alguna variable, mientras que la tarea de clasificación tiene una respuesta categórica (Valcárcel Asencios, 2014). Las técnicas supervisadas o predictivas incluyen los métodos de Análisis de Regresión Logística, Redes Neuronales Artificiales, Árboles de Decisión, Bootstrap, Bagging, CART, Random Forest, C5.0 y Support Vector Machines.

Por otro lado, en las técnicas no supervisadas o descriptivas no se asigna un objetivo predeterminado a las variables. Se supone que no existen variables dependientes o independientes, ni se supone la existencia de un modelo previo para los datos. Los modelos se crean automáticamente a partir del reconocimiento de patrones. Esta técnica incluye métodos de agrupamiento y segmentación, métodos de asociación y reducción de dimensiones y escalado multidimensional. Tanto las técnicas predictivas como las descriptivas se centran en el descubrimiento de conocimiento integrado en los datos.

El rendimiento de los estudiantes es una parte esencial en las instituciones de enseñanza superior, esto se debe a que uno de los criterios de una universidad de alta calidad se basa en su excelente historial de logros académicos (Shahiri et al., 2015). En este contexto el análisis y estudio del rendimiento académico mediante técnicas de minería de datos en centros de educación superior cobra importancia, con fines de entender de mejor manera el rendimiento académico y poder valorar la calidad de los aprendizajes, como lo afirma Cano Celestino & Robles Rivera (2018) textualmente: "el rendimiento académico constituye un indicador importante a la hora de valorar la calidad educativa en la educación superior". La minería de datos y el campo de la educación se combinan en lo que se denomina minería de datos educativos, que ayuda a identificar las características y la información de los estudiantes (Amjad et al., 2022). Asimismo, la minería de datos es una de las técnicas más populares para analizar el rendimiento de los estudiantes. Así,

se denomina minería de datos educativos. La minería de datos educativos es un proceso utilizado para extraer información útil y patrones de una enorme base de datos educativa. La información y los patrones útiles pueden utilizarse para predecir el rendimiento de los alumnos. Como resultado, ayudaría a los educadores a proporcionar un enfoque de enseñanza eficaz (Shahiri et al., 2015).

El presente estudio propone la identificación de la técnica supervisada de minería de datos que brinda mejor desempeño para el análisis del rendimiento académico de estudiantes universitarios, de una universidad peruana. El objetivo del presente estudio fue identificar la técnica supervisada de minería de datos con mejor desempeño para el análisis del rendimiento académico de estudiantes de la Universidad Nacional Amazónica de Madre de Dios, Amazonía peruana.

2. Materiales y métodos

Para el logro del objetivo planteado se empleó una adaptación de la metodología denominada Knowledge Discovery in Databases (KDD), propuesta por Brodley et al. (1999), las fases de la metodología propuesta en el presente estudio muestran en la Figura 1:

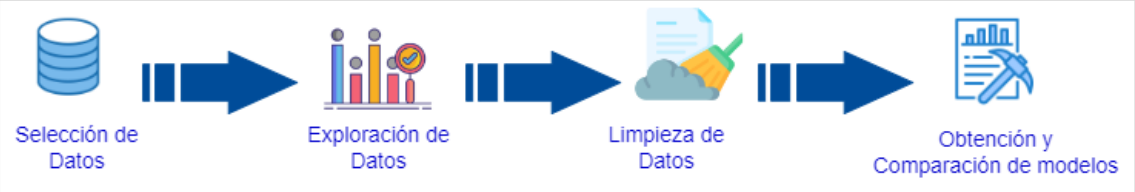


Figura 1. Metodología empleada para la aplicación de técnicas supervisadas de minería de datos en el análisis del rendimiento académico

Fase de selección de datos. La selección de los datos se realizó con la con la ayuda del personal autorizado de la Dirección Universitaria de Asuntos Académicos de la Universidad Nacional Amazónica de Madre de Dios, siendo resultado de ello un reporte de datos anonimizados en formato Excel 2016, correspondiente al periodo semestral (2001-I y 2019-II). Anonimizar los datos implica convertir los datos de forma que no se pueda identificar a los individuos. El conjunto de conjunto de datos original estuvo conformado por 17771 filas y 22 columnas. La Tabla 1 muestra la descripción de las variables del conjunto de datos.

Tabla 1. Descripción de campos del conjunto de datos a analizar

Columna	Formato inicial	Descripción
id	Carácter	Número correlativo para cada registro.
sexo	Carácter	Sexo del estudiante, (0: Femenino, 1: Masculino)
ubigeo_departamento	Carácter	Código de ubicación geográfica del departamento de procedencia del estudiante.
ubigeo_provincia	Carácter	Código de ubicación geográfica de la provincia de procedencia del estudiante.
ubigeo_distrito	Carácter	Código de ubicación geográfica del distrito de procedencia del estudiante.
edad_ingreso	Carácter	Edad de ingreso a la carrera del estudiante.
deudor	Carácter	Deuda con la universidad (0: No posee deuda, 1: Posee deuda).
codigo_carrera	Carácter	Código de carrera al que pertenece el estudiante.
gestion_colegio	Carácter	Gestión del colegio de egreso del estudiante.
modalidad	Carácter	Modalidad del colegio de egreso del estudiante.

escuela_ubigeo_departamento	Carácter	Código de ubicación geográfica-departamento de la escuela de egreso del estudiante.
escuela_ubigeo_provincia	Carácter	Código de ubicación geográfica-provincia de la escuela de egreso del estudiante.
escuela_ubigeo_distrito	Carácter	Código de ubicación geográfica-distrito de la escuela de egreso del estudiante.
codigo_modalidad_ingreso	Carácter	Código de modalidad de ingreso a la universidad.
modalidad_ingreso	Carácter	Modalidad de ingreso a la universidad.
tipo_matricula	Carácter	Tipo de matrícula.
cant_creditos_matriculados	Carácter	Cantidad de créditos matriculados.
cant_creditos_desaprobados	Carácter	Cantidad de créditos aprobados.
cant_creditos_aprobados	Carácter	Cantidad de créditos aprobados.
promedio_acumulado	Carácter	Promedio ponderado acumulado.

Fase de exploración de datos. En esta fase se realizó un primer acercamiento al conjunto de datos, para entender la naturaleza de los datos. El resumen del conjunto de datos se muestra en la Tabla 2, en esta se detalla la cantidad de filas, columnas, cantidad de variables por tipo, así como la completitud del conjunto de datos.

Tabla 2. Resumen del conjunto de datos

Nombre	**Valor**
Filas	17,771
Columnas	19
Columnas discretas	14
Columnas continuas	5
Todas las columnas que faltan	0
Observaciones faltantes	0
Filas completas	17,771
Observaciones totales	337,649
Asignación de memoria	2,7 Mb

El resumen del conjunto de datos muestra que no existen valores faltantes en ninguna de las variables. A continuación, se procedió con establecer el tipo de datos adecuado a cada variable del conjunto de datos. Obsérvese los resultados de esta tarea:

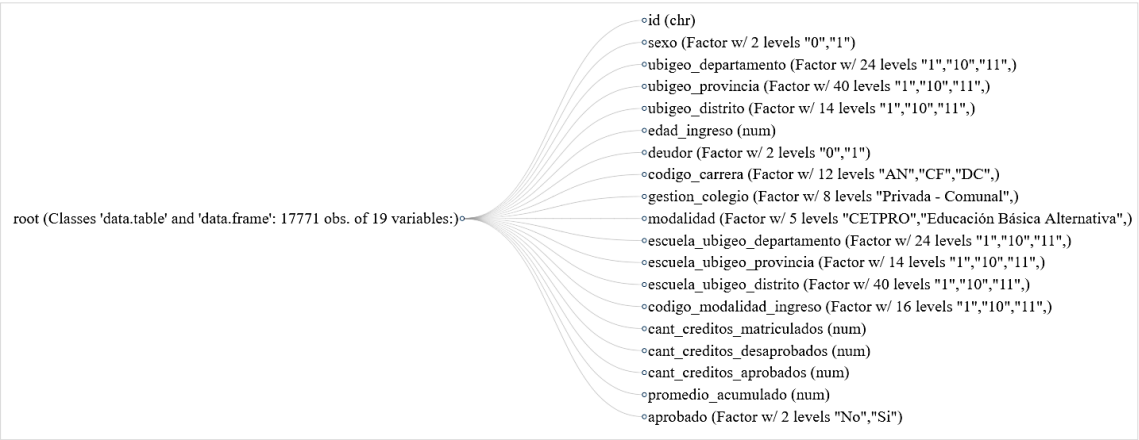


Figura 2. Resumen de las variables del conjunto de datos

Como última tarea de esta fase se procede a verificar la distribución, simetría y normalidad de los datos, los gráficos de histograma y densidad de las variables numéricas, este procedimiento se realiza con la finalidad de verificar el grado de simetría o asimetría de las mismas
Carbono orgánico en el suelo**.**

Fase limpieza de datos. Dado que el conjunto de datos posee variables categóricas y numéricas, se procede a verificar la existencia de valores atípicos en las variables numéricas. La Figura 3 muestra los diagramas de cajas y bigotes de esta tarea.

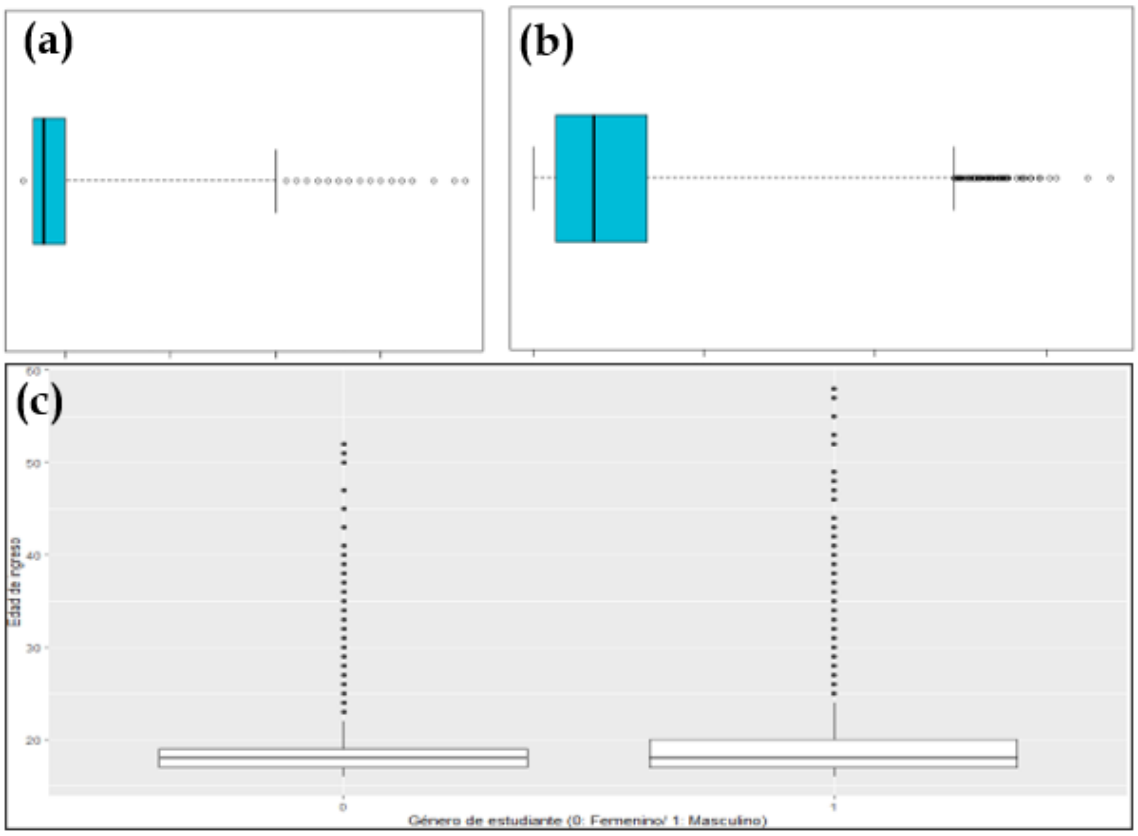


Figura 3. Distribución de las variables numéricas

(a) Edad de ingreso.

(b) Cantidad de créditos acumulados.

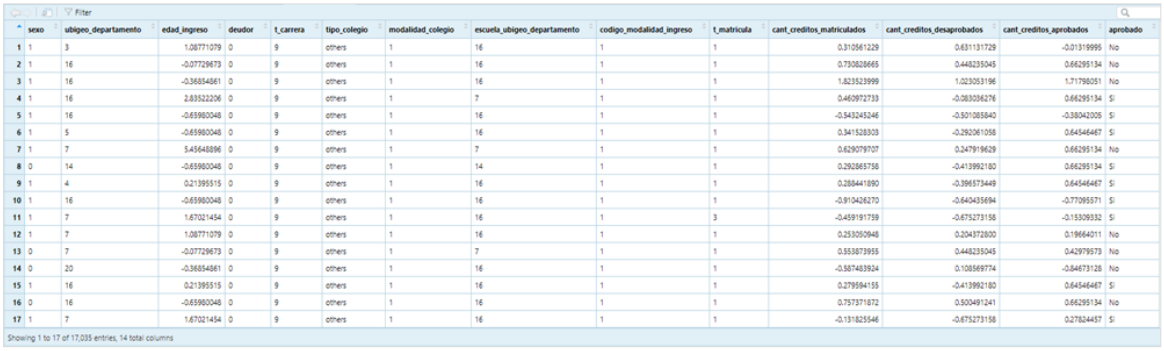
(c) Edad de ingreso por género.

En la Figura 3, se puede observar valores atípicos y algunos valores con errores al momento de la digitación, lo que nos indica que se debe eliminar estos valores para mejorar el rendimiento de los algoritmos de predicción que se emplearán.

Para corregir esto se aplicó a cada una de estas variables la imputación por los valores de los percentiles 5 y 95, se debe aclarar que se consideró valor extremo aquellos que están por encima de (Q3+1,5 x IQR) o por debajo de (Q1-1,5 x IQR), siendo Q1 y Q3 los cuartiles 1 y 3 respectivamente (IQR= Q3-Q1). Tras la limpieza de los valores atípicos el conjunto de datos se redujo a 17035 registros.

Como última tarea de esta fase se realizó la estandarización de variables mediante la combinación de las transformaciones 'scale' y 'center', mediante este proceso los atributos numéricos obtendrán un valor medio de 0 y una desviación estándar de 1, método center resta la media de los valores, mientras que scale divide los valores por la desviación estándar. Luego de esta tarea

se obtiene la vista minable con 17,035 filas y un total de 14 columnas que se muestra en la Figura 4.



	sexo	ubigeo_departamento	edad_ingreso	deudor	t_carrera	tipo_colegio	modalidad_colegio	escuela_ubigeo_departamento	codigo_modalidad_ingreso	t_matricula	cant_creditos_matriculados	cant_creditos_desaprobados	cant_creditos_aprobados	aprobado
1	1	3	1.08771079	0	9	others	1	16	1	1	0.310561229	0.631131729	-0.01319995	No
2	1	16	-0.07729673	0	9	others	1	16	1	1	0.730828665	0.448235045	0.66295134	No
3	1	16	-0.36854861	0	9	others	1	16	1	1	1.823523999	1.023053196	1.71796051	No
4	1	16	2.83522206	0	9	others	1	7	1	1	0.469972733	-0.083036276	0.66295134	Si
5	1	16	-0.65980048	0	9	others	1	16	1	1	-0.543245246	-0.501085840	-0.38042005	Si
6	1	5	-0.65980048	0	9	others	1	16	1	1	0.341528303	-0.292061058	0.64546467	Si
7	1	7	5.45648896	0	9	others	1	7	1	1	0.629079707	0.247919629	0.66295134	No
8	0	14	-0.65980048	0	9	others	1	14	1	1	0.292865758	-0.413992180	0.66295134	Si
9	1	4	0.21395515	0	9	others	1	16	1	1	0.288441890	-0.396573449	0.64546467	Si
10	1	16	-0.65980048	0	9	others	1	16	1	1	-0.910426270	-0.640435694	-0.77095571	Si
11	1	7	1.67021454	0	9	others	1	16	1	3	-0.459191759	-0.675273158	-0.13309332	Si
12	1	7	1.08771079	0	9	others	1	16	1	1	0.253050948	0.204372800	0.19664011	No
13	0	7	-0.07729673	0	9	others	1	7	1	1	0.553873955	0.448235045	0.42979573	No
14	0	20	-0.36854861	0	9	others	1	16	1	1	-0.587483924	0.108569774	-0.84673128	No
15	1	16	0.21395515	0	9	others	1	16	1	1	0.279594155	-0.413992180	0.64546467	Si
16	0	16	-0.65980048	0	9	others	1	16	1	1	0.757371872	0.500491241	0.66295134	No
17	1	7	1.67021454	0	9	others	1	16	1	1	-0.131825546	-0.675273158	0.27824457	Si

Figura 4. Conjunto de datos preprocesado

Fase de exploración de datos. Esta fase se inicia con la verificación del balance de la variable objetivo "aprueba", este procedimiento es importante para ver si el conjunto de datos se encuentra balanceada.

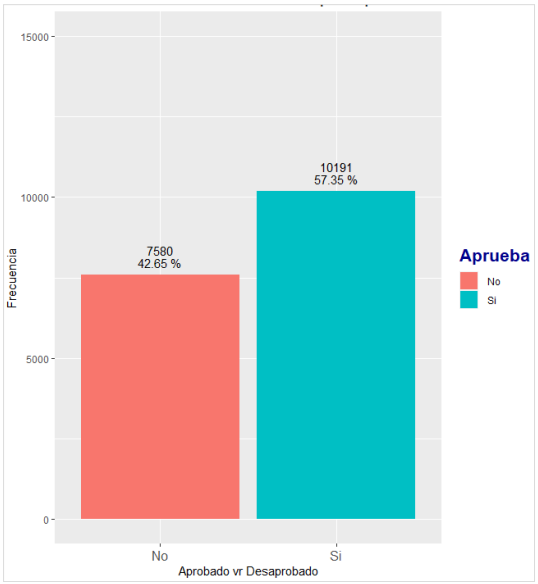


Figura 5. Distribución de Estudiantes de acuerdo con la variable objetivo

De acuerdo con la Figura 5 se observa que la variable objetivo o variable de clase "aprueba", se encuentra relativamente balanceada, por lo que no será necesario ejecutar alguna técnica de balanceo.

Una segunda tarea en esta fase fue dividir el conjunto de datos para el entrenamiento y prueba de los modelos, para los experimentos del presente estudio se consideró un 70% de los datos para el entrenamiento y un 30% para las pruebas. Obteniendo 11924 filas para el entrenamiento y 5111 filas para la prueba de los modelos.

Para la obtención de los modelos de minería de datos se emplearon los algoritmos de regresión logística binaria (GLM), árbol de decisión CART, árbol de decisión C5.0, máquina de vector de soporte (SVM), k vecinos más cercanos (KNN).

3. Resultados y discusión

Los resultados experimentales de la aplicación de las técnicas supervisadas de minería de datos para el análisis del rendimiento académico se obtienen del conjunto de datos de prueba.

La Tabla 3 muestra la matriz de confusión, exactitud y coeficiente de Kappa obtenidos por cada uno de los algoritmos de aprendizaje supervisado.

Tabla 3. Métrica de matriz de confusión de los modelos empleados

			Predicción		**Exactitud**	**Kappa de Cohen**
			1	0		
GLM	Observación	1	2796	129	91%	82%
		0	313	1873		
CART	Observación	1	2632	293	91%	81%
		0	177	2009		
C5.0	Observación	1	2733	192	**93%**	**85%**
		0	173	2013		
SVM	Observación	1	2797	128	91%	82%
		0	324	1862		
KNN	Observación	1	2659	266	85%	69%
		0	494	1692		

En la Tabla 3, se observa que el algoritmo de árbol de decisión C5.0 obtiene una exactitud o porcentaje de aciertos del 93%, un coeficiente de kappa del 0,85 siendo estos los más altos en comparación a los demás algoritmos, un valor del coeficiente de kappa del 0,85 de acuerdo con (Landis & Koch, 1977) , este valor indica una concordancia casi perfecta entre la clasificación real y la clasificación predicha por el modelo de construido por el algoritmo C5.0.

Estos resultados superan a los reportados por Yağcı (2022) donde los algoritmos de Random Forest y Redes Neuronales obtienen solo un 74,6% de exactitud y un área bajo la curva ROC AUC de 0,860 y 0,863 respectivamente en la predicción del rendimiento académico de estudiantes.

La Figura 6 muestra las gráficas de las áreas bajo la curva ROC, para cada uno de los modelos construidos en el conjunto de datos de prueba.

En la Figura 6c se observa la curva ROC correspondiente al modelo C5.0, obtiene un área bajo la curva ROC (AUC) de 0,9797 siendo este el valor más alto en comparación a los demás modelos, esto indica que la capacidad discriminativa del modelo de árbol de decisión C5.0 para las clases aprobado y desaprobado es muy bueno, afirmamos ello de acuerdo con la escala de valoración propuesta por Roig et al. (2017).

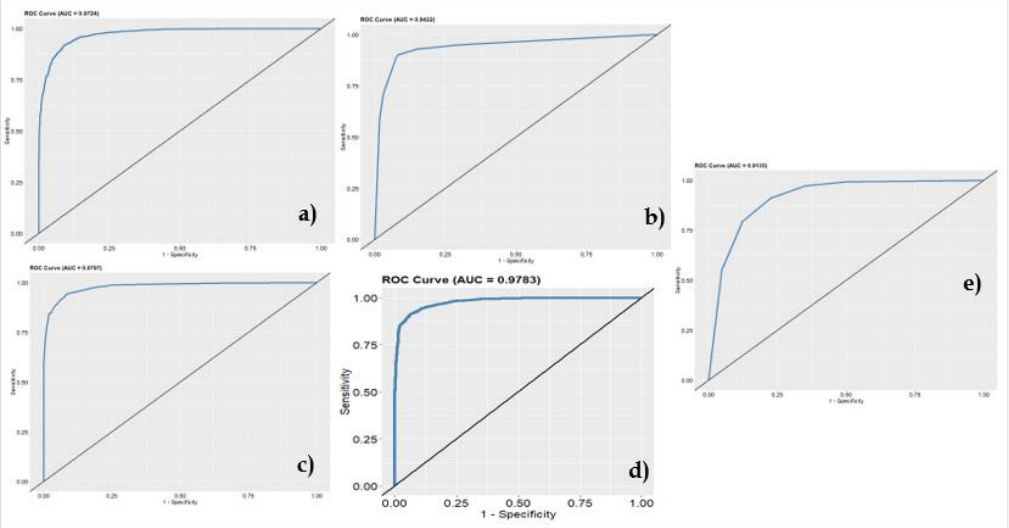


Figura 6. Área bajo la curva ROC de los modelos empleados

(a) GLM

(b) CART

(c) C5.0

(d) SVM

(e) KNN

La Tabla 4, muestra las métricas obtenidas por los algoritmos en el conjunto de datos de prueba, las dos últimas columnas muestran el tiempo en segundos consumido en CPU. Los resultados demuestran que el algoritmo de árbol de decisión C5.0 obtuvo una exactitud del 93% y un valor de coeficiente de kappa del 0,85 siendo los más altos en comparación de los demás algoritmos. Similares resultados reportan (Agaoglu, 2016), en la predicción del rendimiento de los profesores de educación superior, donde el mencionado algoritmo obtiene un valor exactitud del 92,3%, superando a algoritmos como: CART, máquina de soporte vectorial, redes neuronales y análisis discriminante. En otro estudio presentado por Shanmugarajeshwari & Lawrance (2016), donde realizan el análisis del desempeño estudiantil mediante técnicas de clasificación, el algoritmo C5.0 obtiene una exactitud del 100%. Finalmente, Sathe & Adamuthe (2021), concluyen que en la predicción del rendimiento académico los algoritmo de Random forest y C5.0 son mejores que J48, CART, Naive Bayes, K vecinos más cercanos y máquinas de soporte vectorial, obteniendo así exactitudes del 99,75% y 98,49% respectivamente.

En relación al área bajo la curva ROC (AUC), los valores obtenidos son similares a los obtenidos por los algoritmos de máquina de soporte vectorial y regresión logística. Finalmente, aunque el algoritmo C5.0 no posee los tiempos mínimos en la fase en la etapa de entrenamiento y predicción estos valores son aceptables y estables en ambas etapas.

Tabla 4. Métricas obtenidas en el conjunto de datos de prueba

Algoritmo	Accuracy	Kappa	AUC	Tiempo CPU en segundos Entrenamiento	Tiempo CPU en segundos Predicción
LR	91%	0,82	0,972	2,86	0,04
CART	91%	0,81	0,942	0,38	0,05
C5.0	93%	0,85	0,98	0,87	0,37
SVM	91%	0,82	0,978	24,96	2,97
KNN	85%	0,69	0,914	0,07	40,36

4. Conclusiones

Al comparar el desempeño de las técnicas supervisadas de minería de datos para el análisis del rendimiento académico de los estudiantes universitarios, se logró identificar al algoritmo C5.0 como el más eficiente, obteniendo una exactitud del 93%, AUC del 0,9797 y un tiempo de entrenamiento de 0,87 segundos.

Financiamiento

La presente investigación fue financiada por la Universidad Nacional Amazónica de Madre de Dios, aprobada mediante Resolución Nº 247-2019-UNAMAD-VRI.

Conflicto de intereses

Los autores declaran que no incurren en conflicto de intereses.

Contribución de autores

H-A, L.: Metodología, análisis formal, investigación, escritura (preparación del borrador final).

U-G, N. J. y Q-B, D. R.: Conceptualización, investigación, escritura (revisión y edición).

U-G, N. J. y V-N, Y.: Curación de datos, visualización.

Referencias bibliográficas

Agaoglu, M. (2016). Predicting Instructor Performance Using Data Mining Techniques in Higher Education. *IEEE Access*, *4*, 2379–2387. https://doi.org/10.1109/ACCESS.2016.2568756

Amjad, S., Younas, M., Anwar, M., Shaheen, Q., Shiraz, M., & Gani, A. (2022). Data Mining Techniques to Analyze the Impact of Social Media on Academic Performance of High School Students. *Wireless Communications and Mobile Computing*, *2022*, 1–11. https://doi.org/10.1155/2022/9299115

Asif, R., Merceron, A., Ali, S. A., & 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

Brodley, C., Lane, T., & Stough, T. (1999). Knowledge Discovery and Data Mining. *American Scientist*, *87*(1), 54. https://doi.org/10.1511/1999.16.807

Cano Celestino, M. A., & Robles Rivera, R. (2018). Factores asociados al rendimiento académico en estudiantes universitarios. *Revista Mexicana de Orientación Educativa*, 1–25. https://doi.org/10.31206/rmdo072018

Enke, D., & Thawornwong, S. (2005). The use of data mining and neural networks for forecasting stock market returns. *Expert Systems with Applications*, *29*(4), 927–940. https://doi.org/10.1016/j.eswa.2005.06.024

Han, J., Kamber, M., & Pe, J. (2012). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier. https://doi.org/10.1016/C2009-0-61819-5

Khor, E. T. (2022). A data mining approach using machine learning algorithms for early detection of low-performing students. *The International Journal of Information and Learning Technology*, *39*(2), 122–132. https://doi.org/10.1108/IJILT-09-2021-0144

Landis, J. R., & Koch, G. G. (1977). An Application of Hierarchical Kappa-type Statistics in the

Assessment of Majority Agreement among Multiple Observers. *Biometrics*, *33*(2), 363. https://doi.org/10.2307/2529786

Lemay, D. J., Baek, C., & Doleck, T. (2021). Comparison of learning analytics and educational data mining: A topic modeling approach. *Computers and Education: Artificial Intelligence*, *2*, 100016. https://doi.org/10.1016/j.caeai.2021.100016

Nabil, A., Seyam, M., & Elfetouh, A. A. (2022). Predicting students' academic performance using machine learning techniques: a literature review. *International Journal of Business Intelligence and Data Mining*, *20*(4), 456. https://doi.org/10.1504/IJBIDM.2022.123214

Norabuena Penadillo, R. M. (2011). *Relación entre el aprendizaje autorregulado y rendimiento académico en estudiantes de enfermería y obstetricia de la Universidad Nacional " Santiago Antúnez de Mayolo " - Huaraz* [Universidad Nacional Mayor de San Marcos]. https://hdl.handle.net/20.500.12672/2904

Roig, J. G., Roma, J. C., Minguillón, J., & Caihuelas Quiles, R. (2017). *Minería de datos modelos y algoritmos* (1st ed.). Editorial UOC.

Rosado Gómez, A. A., & Verjel Ibáñez, A. (2015). Minería de datos aplicada a la demanda del transporte aéreo en Ocaña, Norte de Santander. *Revista Tecnura*, *19*(45), 101. https://doi.org/10.14483/udistrital.jour.tecnura.2015.3.a08

Sathe, M. T., & Adamuthe, A. C. (2021). Comparative Study of Supervised Algorithms for Prediction of Students' Performance. *International Journal of Modern Education and Computer Science*, *13*(1), 1–21. https://doi.org/10.5815/ijmecs.2021.01.01

Shahiri, A. M., Husain, W., & Rashid, N. A. (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

Shanmugarajeshwari, V., & Lawrance, R. (2016). Analysis of students' performance evaluation using classification techniques. *2016 International Conference on Computing Technologies and Intelligent Data Engineering (ICCTIDE'16)*, 1–7. https://doi.org/10.1109/ICCTIDE.2016.7725375

Trakunphutthirak, R., & Lee, V. C. S. (2022). Application of Educational Data Mining Approach for Student Academic Performance Prediction Using Progressive Temporal Data. *Journal of Educational Computing Research*, *60*(3), 742–776. https://doi.org/10.1177/07356331211048777

Valcárcel Asencios, V. (2014). Data Mining y el descubrimiento del conocimiento. *Industrial Data*, *7*(2), 083. https://doi.org/10.15381/idata.v7i2.6140

Yağcı, M. (2022). Educational data mining: prediction of students' academic performance using machine learning algorithms. *Smart Learning Environments*, *9*(1), 11. https://doi.org/10.1186/s40561-022-00192-z