Rev. innova educ. (2025). Vol. 7 Núm. 2 págs. 88-105

Revista Innova Educación
ISSN 2664-1496 / ISSN-L 2664-1488
https://revistainnovaeducacion.com
Editada por: Instituto Universitario de Innovación Ciencia y Tecnología Inudi Perú

ARTICULO ORIGINAL

Rendimiento académico y condiciones socioculturales de estudiantes de una institución educativa rural en Perú

Academic performance and sociocultural conditions of students from a rural educational institution in Peru

Desempenho acadêmico e condições socioculturais de estudantes de uma instituição educacional rural do Peru

Donia Ruelas[1]
Universidad Nacional del Altiplano, Puno - Puno, Perú
https://orcid.org/0000-0003-1698-1715
druelas@unap.edu.pe (correspondencia)

Elio Ruelas
Universidad Nacional del Altiplano, Puno - Puno, Perú
https://orcid.org/0000-0003-3735-7030
druelas@unap.edu.pe

Marta Carrera
Universidad Nacional del Altiplano, Puno - Puno, Perú
https://orcid.org/0000-0002-5847-6154
druelas@unap.edu.pe

Miguel Aceituno
Universidad Nacional del Altiplano, Puno - Puno, Perú
https://orcid.org/0000-0002-9298-2579
maceituno@unap.edu.pe

DOI: https://doi.org/10.35622/j.rie.2025.02.005

Recibido: 30/03/2025 Aceptado: 11/05/2025 Publicado: 30/05/2025

PALABRAS CLAVE

analítica del aprendizaje, aprendizaje automático, estudiantes de secundaria, rendimiento académico, zona rural.

RESUMEN. Las instituciones educativas recopilan y gestionan información de sus estudiantes con el fin de fundamentar la toma de decisiones orientadas a mejorar el rendimiento académico y el desempeño de su comunidad. El objetivo de esta investigación fue identificar hallazgos sobre el rendimiento académico considerando factores socioculturales a lo largo de la formación secundaria. El estudio siguió el proceso KDD (Descubrimiento de Conocimiento en Bases de Datos) y aplicó algoritmos de aprendizaje automático para la minería de datos. Como resultado, se identificaron dos hallazgos principales: en primer lugar, el entorno familiar y el nivel de responsabilidad influyeron directamente en el rendimiento académico; en segundo lugar, se detectaron cinco perfiles de estudiantes basados en estos factores: (1) entorno familiar favorable y alto nivel de

[1] Ingeniera de Sistemas, Universidad Nacional del Altiplano - Puno, Perú.

responsabilidad, asociados a un rendimiento sobresaliente; (2) entorno favorable y baja responsabilidad, con desempeño limitado; (3) entorno desfavorable y alta responsabilidad, que permite compensar desventajas familiares; (4) entorno desfavorable y baja responsabilidad, vinculado al rendimiento más bajo; y (5) perfiles intermedios con combinaciones mixtas de condiciones familiares y responsabilidad. Estos hallazgos, obtenidos mediante técnicas de aprendizaje automático, brindan un conocimiento sobre la situación de la institución educativa, lo que facilita decisiones más asertivas para la mejora continua del rendimiento académico en el sector rural del Perú. Estos hallazgos derivan de datos de una institución rural específica; por tanto, podrían variar en investigaciones realizadas en contextos urbanos u otros países.

KEYWORDS

learning analytics, machine learning, secondary school students, academic performance, rural area.

ABSTRACT. Educational institutions collect and manage information about their students in order to support decision-making aimed at improving academic performance and the overall achievement of their community. The objective of this research was to identify findings on academic performance considering sociocultural factors throughout secondary education. The study followed the KDD process (Knowledge Discovery in Databases) and applied machine learning algorithms for data mining. As a result, two main findings were identified: first, the family environment and the level of student responsibility directly influenced academic performance; second, five student profiles were detected based on these factors: (1) favorable family environment and high level of responsibility, associated with outstanding performance; (2) favorable environment and low responsibility, with limited performance; (3) unfavorable environment and high responsibility, which partially compensates for family disadvantages; (4) unfavorable environment and low responsibility, linked to the lowest performance; and (5) intermediate profiles with mixed combinations of family conditions and responsibility. These findings, obtained through machine learning techniques, provide knowledge about the situation of the educational institution, facilitating more assertive decisions for the continuous improvement of academic performance in the rural sector of Peru. These findings derive from data collected in a specific rural institution; therefore, they may vary in studies conducted in urban contexts or other countries.

PALAVRAS-CHAVE

analítica da aprendizagem, aprendizado de máquina, estudantes do ensino secundário, desempenho acadêmico, zona rural.

RESUMO. As instituições de ensino coletam e gerenciam informações de seus estudantes com o objetivo de fundamentar a tomada de decisões voltadas para a melhoria do desempenho acadêmico e do rendimento de sua comunidade. O objetivo desta pesquisa foi identificar achados sobre o desempenho acadêmico considerando fatores socioculturais ao longo da formação secundária. O estudo seguiu o processo KDD (Descoberta de Conhecimento em Bases de Dados) e aplicou algoritmos de aprendizado de máquina para a mineração de dados. Como resultado, foram identificados dois achados principais: em primeiro lugar, o ambiente familiar e o nível de responsabilidade influenciaram diretamente o desempenho acadêmico; em segundo lugar, foram detectados cinco perfis de estudantes com base nesses fatores: (1) ambiente familiar favorável e alto nível de responsabilidade, associados a desempenho destacado; (2) ambiente favorável e baixa responsabilidade, com desempenho limitado; (3) ambiente desfavorável e alta responsabilidade, que permite compensar parcialmente as desvantagens familiares; (4) ambiente desfavorável e baixa responsabilidade, vinculado ao menor desempenho; e (5) perfis intermediários com combinações mistas de condições familiares e responsabilidade. Esses achados, obtidos por meio de técnicas de aprendizado de máquina, fornecem conhecimento sobre a situação da instituição de ensino, facilitando decisões mais assertivas para a melhoria contínua do desempenho acadêmico no setor rural do Peru. Esses resultados derivam de dados de uma instituição rural específica; portanto, podem variar em pesquisas realizadas em contextos urbanos ou em outros países.

1. INTRODUCCIÓN

El aprendizaje automático es una rama de la inteligencia artificial que estudia la capacidad de aprendizaje que pueden tener las computadoras a partir de los datos de un contexto determinado. El aprendizaje automático se ha convertido en una herramienta clave en entornos educativos digitalizados, donde grandes volúmenes de información son generados continuamente por sistemas de información educativos plataformas virtuales y sistemas de gestión del aprendizaje (Bhatt et al., 2019). Entre las principales características del aprendizaje

automático destacan su capacidad para encontrar patrones a partir de datos, la automatización de tareas analíticas complejas sin requerir programación explícita, y adaptarse a contextos diversos y dinámicos, como el ámbito educativo (Hilbert et al., 2021).

La aplicación del aprendizaje automático facilita y potencia la analítica de datos en el ámbito educativo para una mejor toma de decisiones (Contreras Bravo et al., 2018). Actualmente se tiene la minería de datos educativos y la analítica de aprendizaje como dos campos centrados en los datos (Lemay et al., 2021) que aprovechan las técnicas del aprendizaje automático para optimizar los procesos de enseñanza y aprendizaje (Baek & Doleck, 2020). Una de las técnicas más utilizadas es el empleo de algoritmos de agrupamiento, como K-means, para segmentar a los estudiantes en distintos perfiles y proporcionar las bases para desarrollar estrategias efectivas de gestión de estudiantes (Liu, 2022).

El campo de la analítica de aprendizaje apoya el proceso de enseñanza y aprendizaje, y su mayor reto es descubrir qué factores impactan en el logro de aprendizaje (Viberg et al., 2018), puesto que se tiene poca evidencia disponible sobre mejoras de aprendizaje en un contexto determinado (Knobbout & Van Der Stappen, 2020).

El rendimiento académico es un concepto complejo y multidimensional que no puede ser abordado desde una perspectiva única, ya que involucra una interacción dinámica de diversos factores que influyen en el desempeño de los estudiantes. Investigaciones recientes han destacado que el rendimiento académico está determinado por una combinación de elementos biológicos, psicológicos, económicos y sociológicos, los cuales interactúan de manera significativa en el proceso de aprendizaje (Tacilla et al., 2020). Esta multidimensionalidad sugiere que no existe un único factor determinante, sino una red de variables que, en conjunto, explican el éxito o el fracaso académico.

Por un lado, se han identificado factores externos que juegan un papel crucial en el rendimiento académico. Entre estos, destacan los aspectos contextuales y sociales, como el entorno familiar, el nivel socioeconómico y las condiciones del entorno comunitario (Grasso, 2020). Asimismo, la asistencia regular a clases y la participación activa en el proceso educativo han demostrado ser predictores significativos del éxito académico (Kassarnig et al., 2018). Por otro lado, los factores internos también han sido ampliamente estudiados. La ansiedad, por ejemplo, ha sido identificada como un obstáculo frecuente que afecta negativamente el rendimiento académico, especialmente en situaciones de evaluación o alta exigencia (Hernández et al., 2022). De igual manera, los rasgos de personalidad, como la perseverancia, la autodisciplina y la capacidad de autorregulación, han demostrado tener una influencia directa en el rendimiento académico (Mateus et al., 2021). Estos factores internos interactúan con las condiciones externas, creando un escenario complejo que define el rendimiento de los estudiantes. Además, otros estudios han explorado cómo variables como la motivación intrínseca y extrínseca, el entorno social inmediato, la zona geográfica en la que reside el estudiante y su edad impactan de manera significativa en su desempeño académico (Quílez et al., 2021).

En este contexto, resulta evidente que el rendimiento académico es el resultado de una interacción compleja entre múltiples factores, tanto internos como externos. Comprender esta dinámica es fundamental para diseñar estrategias efectivas que permitan mejorar el desempeño académico de los estudiantes, especialmente en contextos donde las condiciones adversas pueden exacerbar las dificultades. En la institución educativa de estudio los estudiantes obtuvieron bajos niveles de logro académico en la Evaluación Censal Educativa del año 2019, que son evaluaciones estandarizadas implementadas por el Ministerio de Educación de Perú, cuyos

resultados fueron: un rendimiento académico previo al inicio del 15.60 % del total de estudiantes, un 39.70% que están en inicio y un 26.80% en proceso y solo 17.90% de estudiantes que alcanzaron un rendimiento académico previsto (ECE, 2019). Por otro lado, en la institución educativa de estudio, al igual que todas las instituciones educativas secundarias del Perú, cuenta con un Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE), implantado por el Ministerio de Educación de Perú, que tiene como propósito contar con la información oportuna y confiable sobre la trayectoria educativa del estudiante durante su permanencia en la institución educativa (Ministerio de Educación, 2018) en base a la información obtenida de los procesos de matrícula y evaluación.

En la presente investigación se buscó responder identificar los hallazgos relevantes sobre el rendimiento académico considerando los condicionantes socioculturales de los estudiantes a lo largo de su formación secundaria.

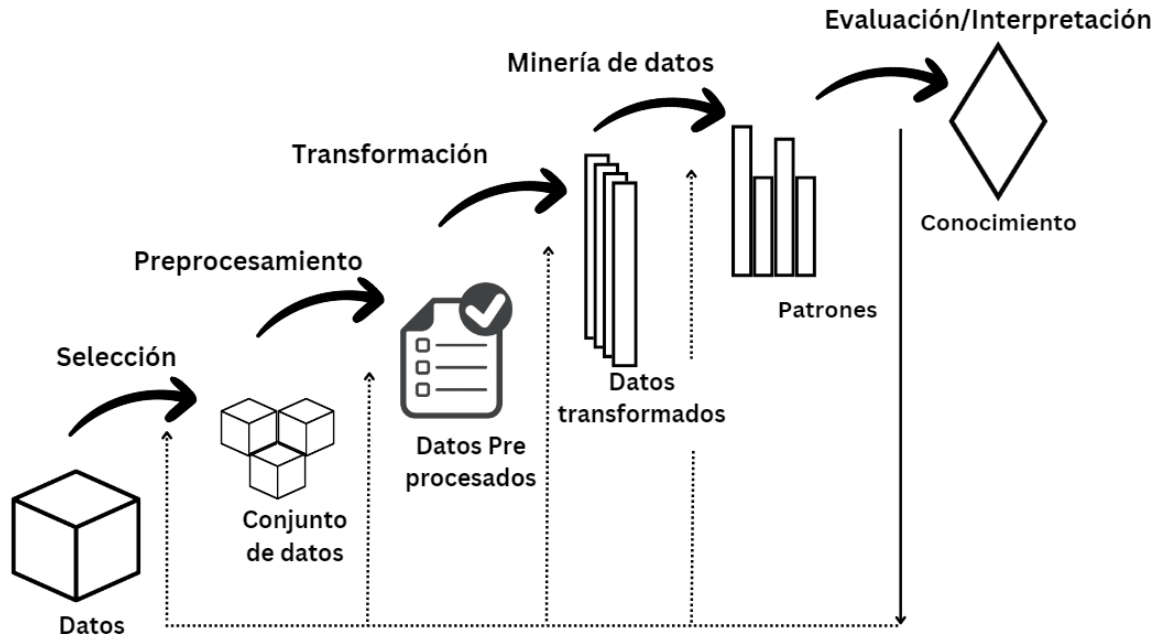
2. MÉTODO.

El presente estudio adoptó un enfoque cuantitativo, ya que se fundamentó en el procesamiento y análisis de datos numéricos mediante algoritmos computacionales (Kamiri & Mariga, 2021). Su diseño es de tipo exploratorio, ya que tuvo como propósito identificar hallazgos relacionados con el rendimiento académico de los estudiantes que permanecieron en la institución educativa durante el periodo escolar 2016-2020, considerando que el objetivo no fue establecer relaciones causales ni verificar hipótesis, sino obtener una comprensión inicial del rendimiento académico de los estudiantes mediante técnicas de análisis automatizado, empleando los datos existentes (Stebbins, 2012).

El proceso empleado para identificar hallazgos relevantes sobre el rendimiento académico de los estudiantes a lo largo de su formación secundaria, fue el proceso de descubrimiento de conocimiento en bases de datos, conocido en inglés como *Knowledge Discovery in DataBases* (KDD), definido como un proceso iterativo que incluye la selección, pre procesamiento, transformación, minería de datos e interpretación de resultados (Frawley et al., 1992) . Este proceso, ha sido aplicado en el ámbito educativo con el propósito de extraer conocimiento útil a partir de los datos disponibles (Romero & Ventura, 2020). Para ello, se siguieron las cinco etapas que lo conforman: selección, pre procesamiento, transformación, minería de datos e interpretación (Buenano-Fernandez et al., 2020). Estas etapas se ilustran en la Figura 1 y se describen a continuación.

Figura 1

Proceso de descubrimiento de conocimiento en bases de datos



Nota. Obtenido de Buenano-Fernandez et al. (2020)

Selección

En esta etapa se empleó un total de 84 registros de estudiantes que permanecieron en la institución educativa durante el periodo escolar 2016-2020. Los datos utilizados fueron extraídos de los reportes generados por el Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE)[2] y proporcionados por la institución educativa con el compromiso de ser utilizados exclusivamente con fines académicos y de investigación. Los datos proporcionados incluyeron información personal, familiar y académica del estudiante, la cual fue debidamente anonimizada, omitiéndose cualquier identificador directo e indirecto; como nombres, número de documento de identidad o direcciones; con el propósito de resguardar la privacidad y confidencialidad de los estudiantes; familiares y académicos del estudiante.

La selección de variables consideradas en este estudio se sustenta en criterios teóricos y a su disponibilidad en la información proporcionada por la institución educativa. En total se seleccionaron 13 variables detallados en la Tabla 1.

En cuanto a las variables académicas, se consideraron las calificaciones promedio del estudiante en las áreas de Matemática, Comunicación y Ciencia y Tecnología (mat_promedio, com_promedio, cta_promedio), debido a que estas son las áreas evaluadas por la Evaluación Censal de Estudiantes (ECE, 2019). Asimismo, se incluyó el promedio de inasistencias (ina_promedio), al ser un indicador relevante del compromiso escolar y un posible predictor del rendimiento académico. En cuanto a las variables personales, se seleccionaron: el grado escolar (grado), que permite situar al estudiante dentro del nivel educativo; la fecha de nacimiento (dia_nac, mes_nac,

[2] Sistema de Información de Apoyo a la Gestión de la Institución Educativa (SIAGIE) es una plataforma informática desarrollada por el Ministerio de Educación del Perú para gestionar y registrar de forma centralizada la trayectoria escolar de los estudiantes desde la educación básica (Ministerio de Educación, 2018).

año_nac), que fue utilizada para calcular la edad relativa del estudiante dentro del grado, un factor que ha demostrado influir significativamente en el desempeño académico (Quílez et al., 2021); y el sexo (sexo), dado que diversos estudios han reportado diferencias de rendimiento por género en ciertas áreas del conocimiento (Marcenaro–Gutierrez et al., 2018) . En lo familiar, se incluyeron variables relacionadas con la situación de los padres (padre_vive, madre_vive), ya que la presencia parental ha sido vinculada al apoyo emocional y académico que incide en el rendimiento estudiantil (Xiong et al., 2021). Asimismo, se consideró la lengua materna del estudiante (lengua_materna), ya que el aprendizaje en una lengua distinta a la propia puede influir negativamente en su desempeño escolar (Mwakira & Mwangi, 2021). Finalmente, se incorporó la escolaridad de la madre (esc_madre), por su reconocida asociación positiva con el rendimiento académico de los hijos (Seden et al., 2020).

Tabla 1

Variables del conjunto de datos seleccionado

Variable	Descripción
grado	Es el grado en el que se ubica el estudiante en el nivel secundaria de la Educación básica Regular de Perú, y este consta de cinco grados (1°,2°,3°,4° y 5°).
dia_nac	Es el día de nacimiento del estudiante.
mes_nac	Es el mes de nacimiento del estudiante.
año_nac	Es el año de nacimiento del estudiante.
sexo	Define el género del estudiante: masculino o femenino.
padre_vive	Si el estudiante tiene a su padre vivo.
madre_vive	Si el estudiante tiene a su madre viva.
Lengua_materna	Es la primera lengua aprendida por el estudiante: quechua o aimara.
esc_madre	Es el grado de instrucción máximo que obtuvo la madre.
ina_promedio	Es el registro promedio de inasistencias del estudiante a la institución educativa durante los años de estudio.
mat_promedio	Es el promedio de la calificación obtenida en el área de Matemática durante los años de estudio.
com_promedio	Es el promedio de la calificación obtenida en el área de Comunicación durante los años de estudio.
cta_promedio	Es el promedio de la calificación obtenida en el área de Ciencia y Tecnología durante los años de estudio.

Preprocesamiento

En el preprocesamiento se crearon cinco variables adicionales (Farnham et al., 2019), como se ilustra en la Tabla 2, que fueron obtenidas a partir de las variables iniciales de la Tabla 1. Como conjunto de datos final se tiene un total de 84 registros de estudiantes con 18 variables.

Tabla 2

Variables creadas

Variable	Descripción
edad	Es la edad obtenida a partir de la diferencia del año cursante y la fecha de nacimiento.
extra_edad	Es cuando se detecta que la edad es fuera del promedio del grado correspondiente, cuyos valores son SÍ o NO.
clasificación_inasistencia	Es el nivel de inasistencia alcanzado por el estudiante, cuyos valores son baja, media y alta
nota_promedio	Es el promedio de las calificaciones obtenidas en las principales áreas que son: Matemática, Comunicación, Ciencia y Tecnología.
resultado_final	Es el logro alcanzado del estudiante de acuerdo a su calificación y se clasifica en: Inicio, Proceso, Logrado y Logro destacado.

En las variables dicotómicas (*padre_vive, extra_edad*) y variables categóricas (*esc_madre, clasificación_inasistencia, resultado_final*) se realizó la codificación a variables numéricas, este proceso facilitó el análisis, la visualización e interpretación de las variables en un contexto determinado. En la Tabla 3, se presenta la codificación de la variable escolaridad de la madre (*esc_madre*) a través de una escala de 1 al 4, donde 1 representa que la madre no tiene formación académica y 4 representa que la madre tiene formación superior. En la Tabla 4, se presenta la codificación de la variable resultado final (*resultado_final*) a través de intervalos de calificaciones obtenidas de una escala vigesimal.

Tabla 3

Codificación de la variable escolaridad de la madre

Identificador	Descripción
esc_madre_id	esc_madre
1	Sin escolaridad
2	Primaria
3	Secundaria
4	Superior

Tabla 4

Codificación de la variable resultado final

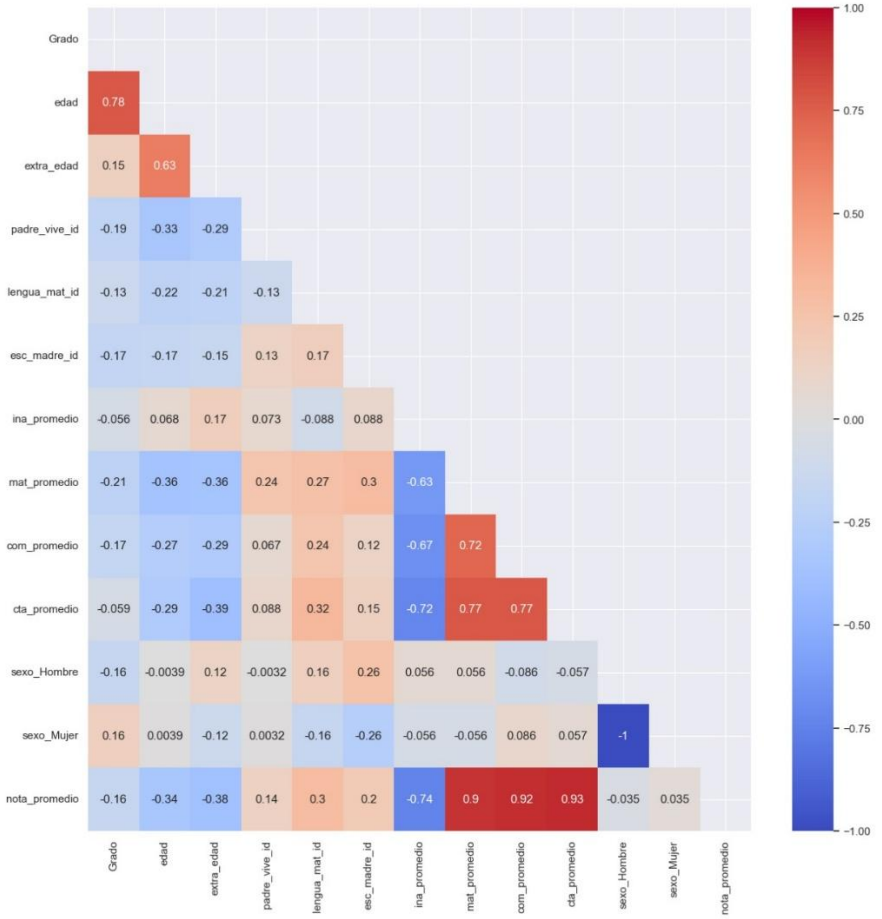
Rango	Descripción
resultado_final	resultado_final
0-10	En inicio
11-13	En proceso
14-17	Logro esperado
18-20	Logro Destacado

Transformación

En esta etapa se empleó la matriz de correlación, que es una tabla que muestra los coeficientes de correlación entre múltiples variables (James et al., 2000), y se identificó la existencia de relaciones lineales entre las variables definidas en la etapa anterior. En la Figura 2, se presenta el mapa de calor de correlación de triángulos entre cada par de variables del conjunto de datos, estas variables muestran las relaciones lineales medidas a través del coeficiente de correlación, donde los cuadrados de color rojo con mayor tonalidad presentan valores cercanos a 1, que identifica una correlación lineal positiva entre variables; y los cuadrados de color azul con mayor tonalidad presentan valores cercanos a -1, que identifica una correlación lineal negativa entre variables (Bruce et al., 2020). En el mapa de calor de correlación de triángulos se muestra que la escolaridad de la madre y la nota promedio tienen un coeficiente de correlación positivo, igual que la lengua materna y la nota promedio, así mismo el padre vive tiene una relación positiva con la nota promedio de matemática; la nota promedio de matemática, la nota promedio de comunicación y la nota promedio de ciencia y tecnología tienen una correlación positiva y alta. Por otro lado, la inasistencia y la nota promedio tienen un coeficiente de correlación negativo y alto, la extra edad del estudiante y la nota promedio tienen una correlación negativa media. El grado y el sexo del estudiante cuentan con bajas relaciones lineales en relación con las otras variables de estudio, en particular con la nota promedio.

Figura 2

Mapa de calor de correlación de triángulos



En esta etapa también se aplicó el análisis de componentes principales (PCA), que funciona mediante el uso de transformaciones ortogonales convirtiendo variables correlacionadas en un conjunto de componentes no correlacionados linealmente (Yang et al., 2018), y se han identificado 3 componentes principales y las variables con mayor importancia en cada componente presentado en la Tabla 5.

Tabla 5

Importancia de las variables en los componentes principales

Variable	Componente Principal 1	Componente Principal 2	Componente Principal 3
extra_edad	**-0.24**	**-0.45**	0.21
sexo_id	0.02	-0.09	**-0.53**
padre_vive_id	0.09	**0.48**	-0.26
lengua_mat_id	0.17	0.14	**0.55**
ina_promedio	**-0.34**	**0.37**	0.07
mat_promedio	**0.42**	-0.01	0.06
com_promedio	**0.41**	-0.17	-0.02
cta_promedio	**0.43**	-0.15	-0.02
nota_promedio	**0.45**	-0.12	0.00

El componente principal 1, es influenciado negativamente por la extra edad e inasistencia del estudiante e influenciado positivamente por las notas alcanzadas en las áreas de matemática, comunicación y ciencia y tecnología. El componente principal 2, es influenciado negativamente por la extra edad del estudiante y positivamente por el padre vivo que tiene el estudiante y la inasistencia de estudiantes. El componente principal 3 es influenciado negativamente por el sexo del estudiante y positivamente por la lengua materna.

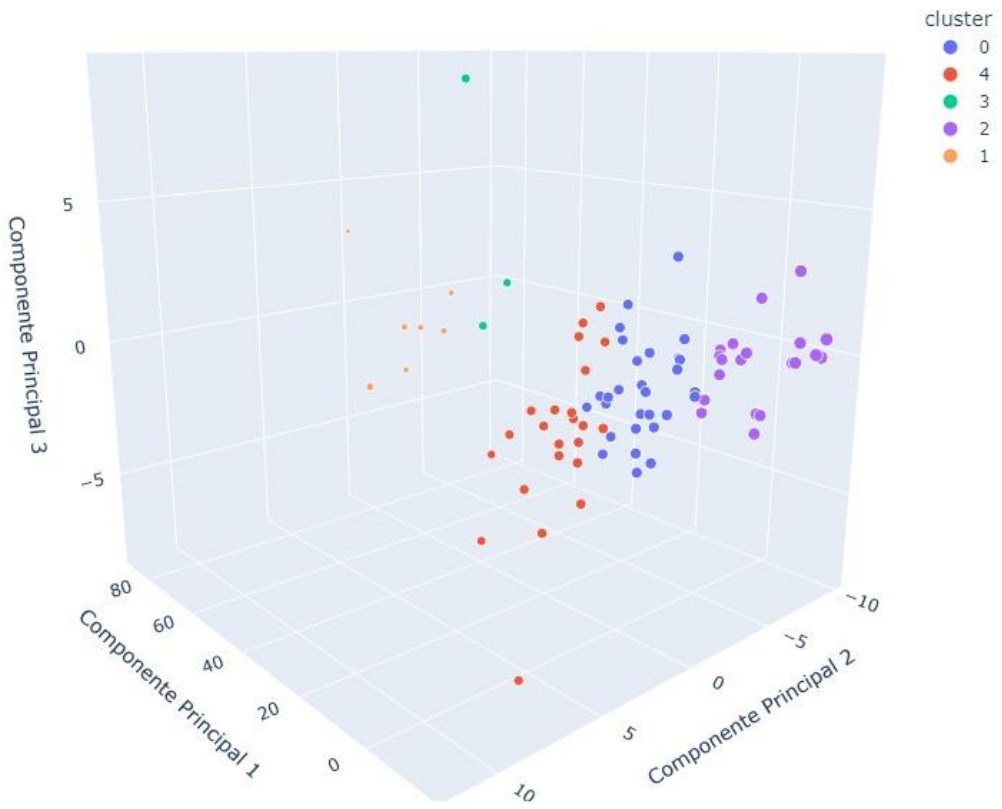
Minería de datos

En la etapa de minería de datos se aplicó el algoritmo de agrupamiento K-means, una técnica representativa del aprendizaje automático no supervisado, ampliamente utilizada para identificar patrones (Liu, 2022). Para determinar el número óptimo de clústeres, se utilizó el método del codo, el cual permitió evaluar que un mayor número de agrupaciones no aportaría una mejora significativa en la cohesión interna de los clústeres (Kelleher et al., 2015). Además, para facilitar la visualización de los resultados en un espacio bidimensional se empleó el método PCA, que permitió proyectar los datos en un plano de dos dimensiones, conservando la mayor variabilidad posible y facilitando la interpretación de los clústeres encontrados.

En la Figura 3 se visualiza cada uno de los cinco clústeres en un espacio de 3 dimensiones donde se aprecia que los clústeres 0, 2 y 4 son los más representativos del total de estudiantes; por el contrario, los clústeres 1 y 3 están compuestos por una cantidad mínima del total de estudiantes.

Figura 3

Visualización de clústeres



Para una mayor interpretación de la visualización se presenta la Tabla 6, donde se muestra el detalle sobre las características de los estudiantes en cada uno de los clústeres generados y el porcentaje de estudiantes asignado a cada uno de los clústeres. Esta caracterización de cada clúster en la siguiente fase permitirá definir perfiles.

Tabla 6

Porcentaje de características de los estudiantes según clúster

Variables (características)		Clúster 0	Clúster 1	Clúster 2	Clúster 3	Clúster 4
extra_edad	Si	0.00%	**8.43%**	**6.02%**	0.00%	4.82%
	No	24.10%	19.28%	32.53%	1.20%	3.61%
padre_vive	Si	24.10%	21.69%	36.14%	1.20%	0.00%
	No	0.00%	**6.02%**	02.41%	0.00%	**8.43%**
lengua_mat	Aimara	0.00%	2.41%	0.00%	0.0%	0.00%
	Quechua	0.00%	**7.23%**	2.41%	0.0%	3.61%
	Castellano	**24.10%**	18.07%	36.14%	1.20%	4.82%
esc_madre	Sin escolaridad	0.00%	3.61%	3.61%	0.00%	0.00%
	Primaria	6.02%	10.84%	**16.87%**	0.00%	2.41%
	Secundaria	**9.64%**	**13.25%**	13.25%	1.20%	6.02%
	Superior	**8.43%**	0.00%	4.82%	0.00%	0.00%
clasificación_inasistencia	Baja	**14.46%**	7.23%	12.05%	0.00%	0.0%
	Media	9.64%	**20.48%**	**26.51%**	1.20%	2.41%
	Alta	0.00%	0.00%	0.00%	0.00%	**6.02%**
resultado_final	En inicio	0.00%	7.23%	0.00%	1.20%	8.43%
	En proceso	0.00%	**20.48%**	21.69%	0.00%	0.00%
	Logrado	**15.66%**	0.00%	16.87%	0.00%	0.00%
	Logro destacado	8.43%	0.00%	0.00%	0.00%	0.00%

3. RESULTADOS Y DISCUSIÓN

Se encontró dos hallazgos principales después del proceso de identificación. La primera revelación es la existencia de dos factores que influyeron directamente en el rendimiento académico de los estudiantes y la segunda es la existencia de cinco perfiles de estudiantes en base a estos factores.

Factores revelados

Los dos factores revelados que influyen en el rendimiento académico de los estudiantes, reflejados en la Tabla 7, están representados por las variables compuestas: *entorno familiar del estudiante* y *nivel de responsabilidad del estudiante*. El primer factor corresponde al entorno familiar del estudiante que está compuesta por las variables: lengua materna del estudiante (Grasso, 2020), que tiene relación directa con la lengua en que se imparte la educación; la escolaridad de la madre, la educación que tiene la madre para transmitir opiniones y consejos a sus hijos; tener el padre vivo, la figura paterna que afecta en el desarrollo de la autonomía y relaciones interpersonales del estudiante. El segundo factor corresponde al nivel de responsabilidad del estudiante, que está compuesta por las variables inasistencia del estudiante a la institución educativa, la elección de asistir a sus sesiones de aprendizaje por los estudiantes (Kassarnig et al., 2018); y la extra edad, de volver a llevar el año

escolar (Abdelmagid & Qahmash, 2023). Los factores revelados aportan a la literatura de factores que influyen en el rendimiento académico de estudiantes en un contexto determinado (Knobbout & Van Der Stappen, 2020).

Tabla 7

Factores que influyeron en el rendimiento académico de los estudiantes del caso de estudio

Factores (cariables compuestas)	Variables	Valores de la variable compuesta
entorno_familiar_estudiante	lengua_mat	Favorable Medianamente favorable Desfavorable
	esc_madre	
	padre_vive	
nivel_responsabilidad_estudiante	clasificación_inasistencia	Alto Medio Bajo
	extra_edad	
	resultado_final	

Perfiles de estudiantes

Los perfiles de estudiantes encontrados en la institución educativa en estudio se detallan en la Tabla 8, estos cinco perfiles fueron revelados con el apoyo del algoritmo de aprendizaje automático K-means. Estos perfiles están basados en el entorno familiar del estudiante y el nivel de responsabilidad del estudiante.

Tabla 8

Perfil de estudiantes según clúster obtenido por K-means

Clúster	Nombre de perfil	Características encontradas del perfil	% de estudiantes
Clúster 0	**Estudiantes destacados** Estudiantes con entorno familiar favorable y alto nivel de responsabilidad.	– No tienen extra edad y tienen a su padre vivo. – Su lengua materna es el castellano. – La escolaridad de la madre es secundaria y superior. – La mayoría tienen inasistencia baja. – Tienen como resultado de aprendizaje Logrado y Logro destacado.	36.14%
Clúster 1	**Estudiantes en riesgo moderado** Estudiantes con entorno familiar medianamente favorable y bajo nivel de responsabilidad.	– Algunos tienen extra edad y no tienen vivo a su padre. – Su lengua materna principalmente es el quechua. – La escolaridad de la madre principalmente es secundaria. – La mayoría tienen inasistencia media. – Tienen como resultado de aprendizaje en inicio y en proceso.	8.43%
Clúster 2	**Estudiantes resilientes** Estudiantes con entorno familiar desfavorable y con nivel de responsabilidad media.	– Tienen extra edad y tiene a su padre vivo. – Su lengua materna principalmente es el castellano. – Principalmente la escolaridad de la madre es primaria. – La mayoría tiene inasistencia media. – Tienen como resultado de aprendizaje en proceso.	24.10%
Clúster 3	**Estudiantes promedio** Estudiantes con entorno familiar favorable y con nivel de responsabilidad media.	– Tienen vivo a su padre. – Su lengua materna es el castellano. – La escolaridad de la madre es secundaria. – Pocos tienen inasistencia media. – Logros de aprendizaje en Proceso y Logrado.	3.61%
Clúster 4	**Estudiantes en riesgo alto** Estudiantes con entorno familiar desfavorable y bajo nivel de responsabilidad.	– No tienen a su padre vivo. – Su lengua principalmente es el castellano. – La escolaridad de la madre es secundaria. – La mayoría tiene inasistencia alta. – Tienen logros de aprendizaje en inicio.	27.71%

El primer perfil corresponde a los estudiantes destacados, que representan el 36,14% del total. Estos estudiantes cuentan con un entorno familiar favorable: su padre está presente, su madre tiene un nivel educativo de secundaria o superior, y su lengua materna es el castellano. Además, se caracterizan por un alto sentido de la responsabilidad, tienen la edad adecuada para el grado que cursan, mantienen una asistencia continua a clases

y tienen un resultado de aprendizaje logrado y logro destacado. El segundo perfil corresponde a los estudiantes en riesgo moderado, que representan el 8,43% del total, estos estudiantes se caracterizan por tener un entorno familiar medianamente favorable y un bajo nivel de responsabilidad. El tercer perfil corresponde a los estudiantes resilientes, quienes representan el 24,10% del total; estos estudiantes se caracterizan por contar con un entorno familiar desfavorable, pero muestran un nivel de responsabilidad medio, lo que les permite superar las adversidades y destacar en su rendimiento académico. El cuarto perfil corresponde a los estudiantes promedio, quienes representan el 3,61% del total; estos estudiantes se caracterizan por contar con un entorno familiar favorable y un nivel de responsabilidad medio, lo que les permite mantener un rendimiento académico equilibrado. Finalmente, el quinto perfil corresponde a los estudiantes en riesgo alto; estos estudiantes se caracterizan por tener un entorno familiar desfavorable y un bajo nivel de responsabilidad, factores que inciden significativamente en su rendimiento académico.

Los hallazgos obtenidos, respaldados por la aplicación de algoritmos de aprendizaje automático, permitieron una comprensión más precisa y detallada de la situación actual de la institución educativa analizada, contribuyendo a lo que plantean Knobbout & Van Der Stappen (2020), ya que los hallazgos obtenidos aportan evidencia empírica sobre el rendimiento académico, lo que demuestra que la aplicación de algoritmos de aprendizaje automático contribuye significativamente a una comprensión más profunda del conocimiento educativo en contextos específicos (Contreras Bravo et al., 2018).

Esta comprensión ha servido como base para proponer sugerencias orientadas a impulsar la mejora continua del rendimiento académico de los estudiantes (Bhatt et al., 2019). En línea con lo planteado por Liu (2022) y Lemay et al. (2021), el uso de técnicas de aprendizaje automático no solo facilita el análisis de grandes volúmenes de datos educativos, sino que también ofrece oportunidades concretas para la toma de decisiones informadas en contextos escolares (Romero & Ventura, 2020). El uso de algoritmos de aprendizaje automático permitió identificar patrones y necesidades específicas en los estudiantes, en ese sentido se alinea con la postura de Contreras Bravo et al. (2018), quienes afirman que la aplicación del aprendizaje automático contribuye a una mejor toma de decisiones, por lo que los resultados obtenidos en el presente estudio sugieren la organización de una escuela para padres de familia, con talleres y charlas dirigidas por especialistas como psicólogos y psicopedagogos, así como la implementación de tutorías grupales e individuales para estudiantes con perfiles específicos. Estas recomendaciones buscan fortalecer el entorno familiar y fomentar la responsabilidad estudiantil, elementos clave para mejorar el rendimiento académico (Knobbout & Van Der Stappen, 2020).

Los hallazgos obtenidos en este estudio se basan en los datos accesibles proporcionados por la institución educativa seleccionada como caso de estudio. Esta disponibilidad de información permitió realizar un análisis contextualizado y pertinente. Asimismo, el enfoque local contribuyó a una comprensión más profunda de las dinámicas académicas propias de la institución, posibilitando la identificación de factores relevantes en el rendimiento estudiantil. Si bien los resultados no pretenden ser generalizables a otras instituciones, sí ofrecen una base para futuras investigaciones comparativas y para la toma de decisiones pedagógicas y administrativas orientadas a la mejora educativa.

4. CONCLUSIÓN

Se identificaron dos hallazgos relevantes: en primer lugar, el entorno familiar y el nivel de responsabilidad del estudiante influyeron directamente en su rendimiento académico; en segundo lugar, se detectaron cinco perfiles de estudiantes basados en estos factores. Estos hallazgos, obtenidos mediante algoritmos de aprendizaje automático, brindan un conocimiento preciso sobre la situación de la institución educativa, lo que facilita la toma de decisiones más asertivas para la mejora continua del rendimiento académico de los estudiantes en el sector rural de Perú.

Se encontraron dos factores que intervinieron en el rendimiento académico de los estudiantes de la institución educativa en estudio; obtenidos con el apoyo del algoritmo de PCA para reducir la dimensionalidad de los datos y destacar las variables más influyentes. El primer factor encontrado es el entorno familiar del estudiante, que es una variable compuesta por las variables lengua materna del estudiante, la escolaridad de la madre y si el padre del estudiante vive; este factor interviene en relación directa con el rendimiento académico del estudiante, es decir un entorno familiar favorable influye positivamente en el logro académico del estudiante, por el contrario, un entorno familiar desfavorable influye negativamente. El segundo factor es el nivel de responsabilidad del estudiante, que es una variable compuesta por las variables inasistencia a la institución educativa, la extra edad del estudiante y los resultados obtenidos; este factor interviene en relación directa con el rendimiento académico del estudiante, es decir un alto nivel de responsabilidad del estudiante influye positivamente en su rendimiento académico, sin embargo, un nivel bajo de responsabilidad del estudiante influye negativamente.

Se identificaron cinco perfiles de estudiantes mediante el uso del algoritmo K-means. El primer perfil agrupa a estudiantes destacados, con un entorno familiar favorable y un alto nivel de responsabilidad. El segundo perfil corresponde a estudiantes en riesgo moderado, con un entorno medianamente favorable y baja responsabilidad. El tercer perfil reúne a estudiantes resilientes, con un entorno desfavorable y responsabilidad media. El cuarto perfil representa a estudiantes promedio, con condiciones familiares favorables y responsabilidad media. Finalmente, el quinto perfil incluye a estudiantes en riesgo alto, afectados por un entorno desfavorable y baja responsabilidad.

Para futuras investigaciones se sugiere aplicar diversos algoritmos de aprendizaje automático a conjuntos de datos con un mayor número de variables, con el fin de analizar el rendimiento académico de estudiantes de instituciones educativas urbanas y de otros países, así como continuar con la validación de los factores identificados en el presente estudio.

Conflicto de intereses / Competing interests:

Los autores declaran que no incurren en conflictos de intereses.

Rol de los autores / Authors Roles:

Donia Ruelas: conceptualización, análisis formal, investigación, metodología, administración del proyecto, supervisión, redacción, revisión y edición.

Elio Ruelas: conceptualización, validación, curación de datos, redacción, revisión y escritura –borrador original.

Marta Carrera: conceptualización, validación, curación de datos, redacción, revisión y escritura –borrador original.

Miguel Aceituno: conceptualización, validación, curación de datos, redacción, revisión y escritura –borrador original.

Fuentes de financiamiento / Funding:

Los autores declaran que no recibieron un fondo específico para esta investigación.

Aspectos éticos / legales; Ethics / legals:

Los autores declaran no haber incurrido en aspectos antiéticos ni haber omitido aspectos legales en la realización de la investigación.

REFERENCIAS

Abdelmagid, A. S., & Qahmash, A. I. (2023). Utilizing the Educational Data Mining Techniques "Orange Technology" for Detecting Patterns and Predicting Academic Performance of University Students. *Information Sciences Letters*, *12*(3), 1415–1431. https://doi.org/10.18576/isl/120330

Baek, C., & Doleck, T. (2020). A Bibliometric Analysis of the Papers Published in the Journal of Artificial Intelligence in Education from 2015-2019. *International Journal of Learning Analytics and Artificial Intelligence for Education (iJAI)*, *2*(1). https://doi.org/10.3991/ijai.v2i1.14481

Bhatt, C., Sajja, P. S., & Liyanage, S. (2019). Utilizing educational data mining techniques for improved learning: Emerging research and opportunities. En *Utilizing educational data mining techniques for improved learning: Emerging research and opportunities*.

Bruce, P., Bruce, A., & Gedeck, P. (2020). Practical statistics for Data scientists. *O'Reilly Media.*, *2nd ed.*

Buenano-Fernandez, D., Gonzalez, M., Gil, D., & Lujan-Mora, S. (2020). Text Mining of Open-Ended Questions in Self-Assessment of University Teachers: An LDA Topic Modeling Approach. *IEEE Access*, *8*(March), 35318–35330. https://doi.org/10.1109/ACCESS.2020.2974983

Contreras Bravo, L. E., Tarazona Bermudez, G. M., & Molano, J. I. R. (2018). Big data: An exploration toward the improve of the academic performance in higher education. *Lecture Notes in Computer Science*, *10943*. https://doi.org/10.1007/978-3-319-93803-5_59

Evaluación Censal de Estudiantes. (2019). Evaluación Censal de Estudiantes/ECE. Ministerio de Educación del Perú. http://umc.minedu.gob.pe/ece2019/

Farnham, B., Tokyo, S., Boston, B., Sebastopol, F., & Beijing, T. (2019). *Hands-on Machine Learning with Scikit-Learn, Keras, and TensorFlow Concepts, Tools, and Techniques to Build Intelligent Systems SECOND EDITION*.

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge Discovery in Databases: An Overview. *AI Magazine*, *13*(3), 57–70.

Grasso, P. (2020). Rendimiento académico : un recorrido conceptual que aproxima a una definición unificada para el ámbito superior. *Revista de Educación*, *11*(20), 87–102.

Hernández, P. J., Contreras, P. J., Inga, F., Ayala, P. B., & Valladares-Garrido, M. J. (2022). Factors associated with academic performance in medical students. *Revista Cubana de Medicina Militar*, *51*(1).

Hilbert, S., Coors, S., Kraus, E., Bischl, B., Lindl, A., Frei, M., Wild, J., Krauss, S., Goretzko, D., & Stachl, C. (2021). Machine learning for the educational sciences. En *Review of Education* (Vol. 9, Número 3). https://doi.org/10.1002/rev3.3310

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2000). An introduction to Statistical Learning. En *Current medicinal chemistry* (Vol. 7, Número 10). https://doi.org/10.1007/978-1-4614-7138-7

Kamiri, J., & Mariga, G. (2021). Research Methods in Machine Learning: A Content Analysis. *International Journal of Computer and Information Technology (2279-0764)*, *10*(2). https://doi.org/10.24203/ijcit.v10i2.79

Kassarnig, V., Mones, E., Bjerre-Nielsen, A., Sapiezynski, P., Lassen, D. D., & Lehmann, S. (2018). Academic performance and behavioral patterns. *EPJ Data Science*, *7*(1). https://doi.org/10.1140/epjds/s13688-018-0138-8

Kelleher, J. D., Namee, B. Mac, & D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics. En *Igarss 2014* (Número 1).

Knobbout, J., & Van Der Stappen, E. (2020). Where is the Learning in Learning Analytics? A Systematic Literature Review on the Operationalization of Learning-Related Constructs in the Evaluation of Learning Analytics Interventions. *IEEE Transactions on Learning Technologies*, *13*(3), 631–645. https://doi.org/10.1109/TLT.2020.2999970

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*. https://doi.org/10.1016/j.caeai.2021.100016

Liu, R. (2022). Data Analysis of Educational Evaluation Using K-Means Clustering Method. *Computational Intelligence and Neuroscience*, *2022*. https://doi.org/10.1155/2022/3762431

Marcenaro–Gutierrez, O., Lopez–Agudo, L. A., & Ropero-García, M. A. (2018). Gender Differences in Adolescents' Academic Achievement. *YOUNG*, *26*(3), 250–270. https://doi.org/10.1177/1103308817715163

Mateus, C., Campis, R., Aguaded, I., Parody, A., & Ruiz, F. (2021). Analysis of personality traits and academic performance in higher education at a Colombian university. *Heliyon*, *7*(5). https://doi.org/10.1016/j.heliyon.2021.e06998

Ministerio de Educación. (2018). Norma que regula el registro de la trayectoria educativa del estudiante de educación básica, a través del Sistema de información de apoyo a la gestión de la institución educativa (SIAGIE). En *Diario oficial El Peruano*. https://cdn.www.gob.pe/uploads/document/file/228118/RM_N__609-2018-MINEDU.PDF

Mwakira, G., & Mwangi, V. (2021).The use of mother tongue language and its effect on academic performance in public schools in Kenya. *International Journal of Linguistics*, *1*(1). https://doi.org/10.47604/ijl.1240

Quílez, A., Moyano, N., & Cortés, A. (2021). Motivational, emotional, and social factors explain academic achievement in children aged 6–12 years: A meta-analysis. En *Education Sciences* (Vol. 11, Número 9). https://doi.org/10.3390/educsci11090513

Romero, C., & Ventura, S. (2020). Educational data mining and learning analytics: An updated survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *10*(3). https://doi.org/10.1002/widm.1355

Seden, K., Willert, S., & Dorji, K. (2020). Impact of Mother's Education on the Academic Achievement of their Children in Three Lower and Secondary Schools of Samtse Dzongkhag: An Enquiry. *Bhutan Journal of Research & Development*, *9*(1).

Stebbins, R. (2012). Exploratory Research in the Social Sciences. En *Exploratory Research in the Social Sciences*. https://doi.org/10.4135/9781412984249

Tacilla, I., Vásquez, S., Verde, E. E., & Colque, E. (2020). Rendimiento académico: universo muy complejo para el quehacer pedagógico. *Revista Muro de la Investigación*, *5*(2), 53–65. https://doi.org/10.17162/rmi.v5i2.1325

Viberg, O., Hatakka, M., Bälter, O., & Mavroudi, A. (2018). The current landscape of learning analytics in higher education. En *Computers in Human Behavior* (Vol. 89). https://doi.org/10.1016/j.chb.2018.07.027

Xiong, Y., Qin, X., Wang, Q., & Ren, P. (2021). Parental Involvement in Adolescents' Learning and Academic Achievement: Cross-lagged Effect and Mediation of Academic Engagement. *Journal of Youth and Adolescence*, *50*(9). https://doi.org/10.1007/s10964-021-01460-w

Yang, S. J. H., Lu, O. H. T., Huang, A. Y. Q., Huang, J. C. H., Ogata, H., & Lin, A. J. Q. (2018). Predicting students' academic performance using multiple linear regression and principal component analysis. *Journal of Information Processing*, *26*. https://doi.org/10.2197/ipsjjip.26.170