Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), julio-agosto 2026,
Volumen 10, Número 4.

https://doi.org/10.37811/cl_rcm.v10i4

PREDICCIÓN DEL RIESGO ACADÉMICO MEDIANTE MINERÍA DE DATOS EN UNA INSTITUCIÓN SECUNDARIA DE QUEVEDO, ECUADOR

PREDICTION OF ACADEMIC RISK USING DATA MINING TECHNIQUES: A CASE STUDY AT A SECONDARY EDUCATION INSTITUTION IN QUEVEDO, ECUADOR

José Steven Cordero Bazurto
Universidad Técnica Estatal de Quevedo, Ecuador

Domenica Mariella Coronado Bustos
Universidad Técnica Estatal de Quevedo, Ecuador

Cristhian Danilo Briones Montalvo
Universidad Técnica Estatal de Quevedo, Ecuador

Sulay Fidela Guaman Caicedo
Universidad Técnica Estatal de Quevedo, Ecuador

Carmen del Rocío Bazurto Zamora
Universidad Técnica Estatal de Quevedo, Ecuador

DOI: https://doi.org/10.37811/cl_rcm.v10i4.25620

Predicción del Riesgo Académico mediante Minería de Datos en una Institución Secundaria de Quevedo, Ecuador

José Steven Cordero Bazurto[1]
jcorderob@uteq.edu.ec
https://orcid.org/0009-0001-2961-6736
Universidad Técnica Estatal de Quevedo
Quevedo, Ecuador

Domenica Mariella Coronado Bustos
dcoronadob@uteq.edu.ec
https://orcid.org/0009-0006-9605-4303
Universidad Técnica Estatal de Quevedo
Quevedo, Ecuador

Cristhian Danilo Briones Montalvo
cbrionesm2@uteq.edu.ec
https://orcid.org/0009-0003-2642-2527
Universidad Técnica Estatal de Quevedo
Quevedo, Ecuador

Sulay Fidela Guaman Caicedo
sulayguaman1992@gmail.com
https://orcid.org/0009-0007-0482-1278
Universidad Técnica Estatal de Quevedo
Quevedo, Ecuador

Carmen del Rocío Bazurto Zamora
bazurtocarmen16@gmail.com
https://orcid.org/0009-0006-2799-1600
Universidad Técnica Estatal de Quevedo
Quevedo, Ecuador

RESUMEN

El rendimiento académico en educación básica se ve afectado por factores como las inasistencias, el apoyo familiar y la salud mental. Este estudio predice el riesgo académico de estudiantes de Noveno Año de una institución secundaria de Quevedo, Ecuador, durante los períodos 2020-2021 a 2025-2026, mediante técnicas de minería de datos. Se siguió la metodología KDD con un enfoque cuantitativo predictivo sobre 695 estudiantes activos de seis cohortes. Se emplearon matrices de correlación, técnicas de agrupamiento (K-Means y DBSCAN) y modelos predictivos (regresión logística, árbol de decisión y redes neuronales), evaluados mediante validación cruzada estratificada con balanceo de clases y exclusión de variables con fuga de información. Los resultados evidenciaron una asociación alta entre comportamiento, inasistencias y riesgo académico. La Regresión Logística fue el modelo más confiable (AUC-ROC = 0.94), superando al Árbol de Decisión (AUC = 0.90) y a las Redes Neuronales (AUC = 0.70). Se concluye que la minería de datos combinada con variables socioemocionales permite anticipar el riesgo académico y diseñar intervenciones oportunas, aunque se recomienda validar el modelo en otras instituciones antes de su uso operativo.

Palabras clave: riesgo académico, minería de datos, predicción, inasistencias, comportamiento estudiantil

[1] Autor principal
Correspondencia: jcorderob@uteq.edu.ec

Prediction of Academic Risk using Data Mining Techniques: A Case Study at a Secondary Education Institution in Quevedo, Ecuador

ABSTRACT

Academic performance in basic education is affected by factors such as absenteeism, family support, and behavioral conduct. This study predicts academic risk among ninth-grade students at a secondary institution in Quevedo, Ecuador, across the 2020-2021 to 2025-2026 school years using data mining techniques. The KDD methodology was applied with a quantitative predictive approach to 695 active students across six cohorts. Correlation matrices, clustering techniques (K-Means and DBSCAN), and predictive models (logistic regression, decision tree, and neural networks) were used, evaluated through stratified cross-validation with class balancing and exclusion of leakage-prone features. Results showed a strong association between behavioral conduct, absenteeism, and academic risk. Logistic Regression was the most reliable model (AUC-ROC = 0.94), outperforming the Decision Tree (AUC = 0.90) and Neural Network (AUC = 0.70). It is concluded that data mining combined with socio-emotional variables enables early detection of academic risk, although validation in other institutions is recommended before operational use.

Keywords: academic risk, data mining, prediction, absenteeism, student behavior

Artículo recibido 20 junio 2026
Aceptado para publicación: 25 julio 2026

INTRODUCCIÓN

En instituciones de educación secundaria del Ecuador, identificar a tiempo a un estudiante en riesgo de bajo rendimiento puede marcar la diferencia entre una intervención oportuna y la deserción escolar; los logros académicos de los estudiantes determinan no solo su desarrollo personal y profesional, sino también la construcción de comunidades más justas. Sin embargo, factores como la inasistencia, el apoyo familiar y la salud mental influyen significativamente en el rendimiento escolar. Este estudio parte de un problema concreto: la institución no contaba con un mecanismo sistemático para anticipar qué estudiantes necesitarían apoyo antes de que sus calificaciones cayeran, durante los períodos lectivos 2020-2021 a 2025-2026.

Diversos estudios respaldan la relevancia de estos factores: la frecuencia de asistencia a clases está relacionada con el éxito académico, y la inasistencia constante puede convertirse en un factor que incrementa la deserción y los comportamientos problemáticos (Pérez Peñaranda et al. 2009). Asimismo, el apoyo familiar es determinante, pues la participación de los padres en la educación de sus hijos mejora la asistencia y las calificaciones (Birioukov 2016). Por su parte, la salud mental (incluyendo ansiedad, depresión u otros problemas psicológicos) genera un impacto negativo en el desempeño académico y puede derivar en el abandono escolar (Finning et al. 2019). A nivel internacional, se ha documentado que hasta un 16.3% de estudiantes de décimo grado reportan ausentismo frecuente, asociado con el nivel de escolaridad de los padres y la falta de supervisión (Henry 2007).

El problema de investigación que orienta este trabajo es: ¿cómo influyen las inasistencias, el apoyo familiar y la salud mental en el rendimiento académico de los estudiantes durante un año escolar? En consecuencia, el objetivo general fue predecir el riesgo académico de los estudiantes de la institución mediante técnicas de minería de datos, con el fin de identificar factores críticos que influyen en el aprendizaje y mejorar las estrategias pedagógicas. Los objetivos específicos incluyeron: (a) revisar la literatura sobre técnicas de minería de datos aplicadas al contexto educativo, (b) identificar la correlación entre inasistencias, apoyo familiar, salud mental y rendimiento académico, (c) agrupar a los estudiantes en clústeres según sus patrones de calificaciones e inasistencias, y (d) implementar un modelo predictivo que identifique a los estudiantes en riesgo de bajo rendimiento.

Esta investigación es pertinente porque proporciona un análisis basado en datos que permite a docentes y autoridades académicas diseñar estrategias de intervención más efectivas y oportunas, contribuyendo además al campo educativo mediante la aplicación de técnicas avanzadas de minería de datos.

Antecedentes y fundamento conceptual

Este trabajo se enmarca dentro de dos campos estrechamente relacionados: la Minería de Datos Educativa (Educational Data Mining, EDM), que aplica técnicas de aprendizaje automático y estadística para descubrir patrones en datos educativos, y la Analítica de Aprendizaje (Learning Analytics), que se centra en la medición, recolección y análisis de datos sobre los estudiantes y su contexto con el fin de comprender y optimizar el aprendizaje y los entornos en los que ocurre. Un producto aplicado de ambos campos son los Sistemas de Alerta Temprana (Early Warning Systems, EWS), diseñados para identificar de forma anticipada a estudiantes en riesgo de bajo rendimiento o deserción, de modo que las instituciones puedan intervenir antes de que el problema se consolide. Akçapınar, Altun y Aşkar (2019) mostraron que un sistema de alerta temprana basado en analítica de aprendizaje permite identificar tempranamente a estudiantes en riesgo con buena precisión, mientras que Chang, Chen y Lee (2025), en un estudio reciente en educación superior taiwanesa, evidenciaron que integrar la predicción algorítmica con intervenciones personalizadas y culturalmente pertinentes incrementa significativamente la tasa de aprobación de estudiantes en riesgo. El presente estudio se inscribe en esta tradición, con la particularidad de aplicarse a educación básica en un contexto latinoamericano, donde este tipo de sistemas ha sido comparativamente menos explorado que en la educación superior.

Las técnicas de minería de datos han sido ampliamente utilizadas en el ámbito educativo. El análisis de agrupamiento (clustering) permite organizar a los estudiantes en conjuntos con características similares (patrones de calificación, inasistencia o problemas de salud mental), siendo K-Means y los algoritmos jerárquicos los más utilizados (Zhang and Paquette 2023). Esta técnica facilita el diseño de intervenciones pedagógicas específicas para cada grupo, como lo demuestran estudios que segmentan a estudiantes de educación superior según patrones de comportamiento académico (Mohd Talib, Abd Majid, and Sahran 2023).

El análisis de regresión, por su parte, modela la relación entre variables dependientes e independientes, permitiendo estudiar la influencia combinada de las inasistencias, el apoyo familiar y la salud mental

sobre las calificaciones (Tsui et al. 2023). Los árboles de decisión, técnica de aprendizaje automático basada en reglas derivadas de los datos, permiten visualizar qué combinación de factores predice mejor el bajo rendimiento académico, apoyando la toma de decisiones oportuna por parte de docentes y directivos (Marcolino et al. 2025).

METODOLOGÍA

El presente estudio adoptó un enfoque cuantitativo de tipo predictivo, con un diseño no experimental y de corte transversal, basado en datos institucionales de seis cohortes correspondientes a los períodos lectivos 2020-2021 a 2025-2026, de estudiantes de Noveno Año de Educación General Básica de la institución.

Para la ejecución del proceso de minería de datos se adoptó la metodología KDD (Knowledge Discovery in Databases), la cual establece un proceso estructurado en fases para la extracción de conocimiento y patrones (Farid and Donyatalab 2023).

Figura 1. Fases de la metodología KDD

Conjunto de datos

Los datos utilizados provienen de dos fuentes institucionales: un conjunto con información académica de los estudiantes y otro con el registro de tutoría, que incluye la asistencia de los padres de familia a citaciones y las intervenciones del área de psicología estudiantil. Ambos conjuntos fueron fusionados para su análisis conjunto, conformando una base de 727 estudiantes de Noveno Año (3635 registros a nivel materia) correspondientes a seis cohortes de los períodos lectivos 2020-2021 a 2025-2026. Para el

análisis predictivo los registros se agregaron a nivel estudiante-período (727 observaciones estudiante-período), promediando las variables académicas entre las cinco materias cursadas, dado que las variables de comportamiento, inasistencias, apoyo familiar y psicología son constantes por estudiante en cada período. El flujo de depuración de la muestra fue el siguiente: 727 estudiantes-período totales: 32 (4.4%) excluidos por tener estado de matrícula "Retirado" (17) o "Trasladado" (15) durante el año lectivo: 695 estudiantes-período utilizados para el entrenamiento y validación de los modelos predictivos y de agrupamiento. Los 32 casos excluidos se trataron de forma diferenciada en el preprocesamiento en lugar de imputarse con la media general. La Tabla 1 resume las principales variables consideradas.

Tabla 1. Variables principales del conjunto de datos.

Variable	Descripción	Tipo	Rango
PROMEDIO GENERAL	Promedio general de calificaciones	Numérico	6.98 – 9.86
PROM_QUIMESTRE_1 / 2	Promedio por quimestre	Numérico	6.85 – 10.00
PORCENTAJE_INASISTENCIAS_Q1 / Q2	Porcentaje de inasistencias por quimestre	Numérico	0 – 16.8 %
COMPORTAMIENTO	Evaluación conductual del estudiante	Numérico	A / B / C / D
REPORTES_PSICOLOGIA	Número de reportes al departamento de psicología	Numérico entero	0 – 5
GENERO / MATERIA	Variables demográficas y académicas	Categórico	-

Procesamiento y transformación de los datos

El preprocesamiento incluyó la unificación de los dos conjuntos de datos en un único archivo y la conversión de columnas numéricas mal tipificadas como texto. Los 32 estudiantes-período con estado de matrícula "Retirado" o "Trasladado" se excluyeron por completo del entrenamiento y validación de los modelos (n final = 695), en lugar de imputarse con el promedio general de la columna, ya que este grupo presenta perfiles de riesgo distintos al resto y su imputación con la media podría atenuar artificialmente la señal de riesgo académico. En la fase de transformación se aplicó codificación ordinal de variables categóricas (género, comportamiento, tipo de apoderado, situación socioeconómica), selección de variables predictoras excluyendo aquellas con fuga de información hacia la variable

objetivo (promedios por quimestre, parciales y exámenes, que componen matemáticamente el promedio general), y estandarización de las variables numéricas (media 0, desviación estándar 1) para evitar que variables de distinta escala distorsionaran el análisis. La variable objetivo "riesgo académico" se define, de forma única y consistente en todo el estudio, como el 20% de estudiantes activos con menor promedio general dentro de la cohorte (umbral relativo, no un corte absoluto), dado que el promedio general osciló entre 6.98 y 9.86 (media = 8.71) y un corte fijo como "promedio menor a 7" habría dejado prácticamente sin casos positivos a la muestra (2 de 695). El procesamiento se realizó en Python 3.12, con las librerías scikit-learn 1.8, pandas 3.0, NumPy y matplotlib/seaborn para las visualizaciones, con semilla aleatoria fija (random_state = 42) en todos los procedimientos estocásticos para garantizar la reproducibilidad de los resultados.

Técnicas aplicadas

Se aplicó una matriz de correlación de Pearson para identificar la relación entre el riesgo académico y el resto de variables, complementada con un mapa de calor. Para el agrupamiento se emplearon los algoritmos K-Means y DBSCAN sobre las variables estandarizadas. El número de conglomerados de K-Means (k=3) se seleccionó comparando el método del codo (inercia) con el coeficiente de silueta para k entre 2 y 6, siendo k=3 el valor con mayor silueta (0.164). Los parámetros de DBSCAN (eps=1.8, min_samples=8) se determinaron mediante una búsqueda en cuadrícula que maximizó el coeficiente de silueta entre las combinaciones evaluadas (eps entre 1.0 y 2.5; min_samples entre 5 y 15). Para la predicción del riesgo académico se implementaron modelos de regresión logística, árbol de decisión y redes neuronales (MLPClassifier), todos con ponderación de clases balanceada (class_weight="balanced") dado el desbalance observado entre estudiantes en riesgo y sin riesgo. La profundidad máxima del árbol de decisión se determinó mediante búsqueda en cuadrícula con validación cruzada sobre el AUC-ROC (valores evaluados: 2, 3, 4, 5, 6, 8 y sin límite), seleccionándose una profundidad de 5 niveles (AUC de validación cruzada = 0.901) como mejor equilibrio entre ajuste y sobreajuste. La red neuronal se configuró con dos capas ocultas de 16 y 8 neuronas, función de activación ReLU, optimizador Adam (tasa de aprendizaje inicial = 0.001) y parada temprana (early stopping) para prevenir sobreajuste, dado el tamaño muestral limitado. La Tabla 2 resume los modelos empleados.

Tabla 2. Modelos y técnicas de minería de datos aplicados.

Modelos	Técnicas	Descripción
Agrupamiento	K-Means	Agrupó a los estudiantes en 3 clústeres según variables relevantes.
Agrupamiento	DBSCAN	Identificó clústeres basados en densidad, útil frente a ruido en los datos.
Predictivo	Regresión logística	Predijo el riesgo académico (20% de estudiantes activos con menor promedio general por cohorte).
Predictivo	Árbol de decisión	Clasificó a los estudiantes según su rendimiento académico.
Predictivo	Redes neuronales	Modelo MLPClassifier aplicado a la predicción de rendimiento.

La evaluación de los modelos predictivos se realizó mediante validación cruzada estratificada de 5 particiones (StratifiedKFold, semilla aleatoria fija = 42 para garantizar reproducibilidad), en lugar de una única partición entrenamiento/prueba, con el fin de obtener estimaciones más estables de precisión, recall, F1-score y área bajo la curva ROC (AUC), complementadas con curvas ROC sobre una partición de reserva del 30%, siguiendo recomendaciones metodológicas previas (Pan 2022). El procesamiento se realizó con Microsoft Excel para la preparación inicial y Python (scikit-learn) para la ejecución de los modelos de minería de datos.

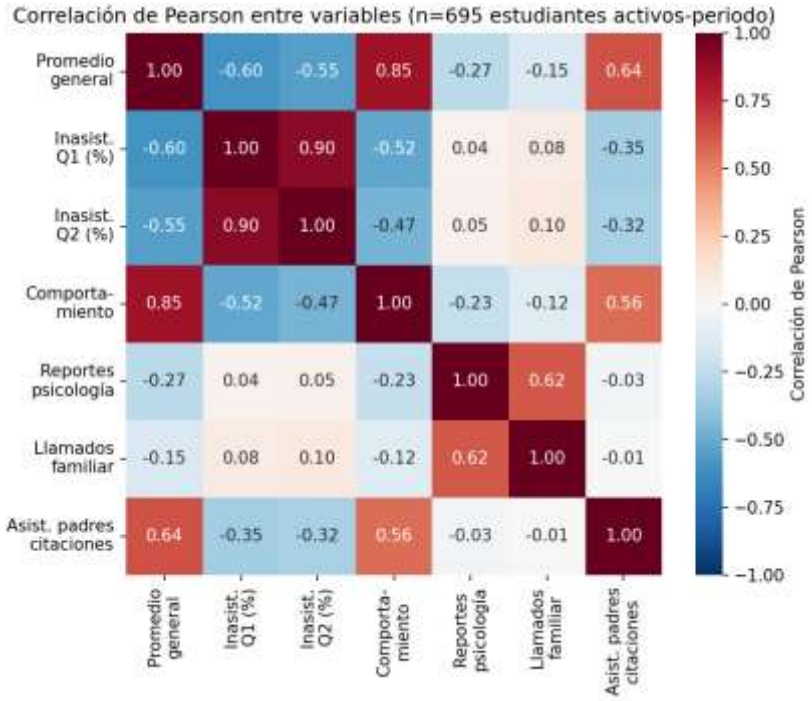
RESULTADOS Y DISCUSIÓN

Correlación entre variables

El análisis de correlación de Pearson, representado en la Figura 2, se calculó sobre los 695 estudiantes activos y únicamente sobre variables sin fuga de información hacia el promedio general (se excluyeron los promedios por quimestre, al ser componentes directos del promedio general). La variable COMPORTAMIENTO mostró la asociación más fuerte con el promedio general (r = 0.85), seguida de la asistencia de los padres a citaciones (r = 0.63). El porcentaje de inasistencias del primer y segundo quimestre presentó una correlación negativa considerable con el promedio general (r = -0.60 y r = -0.55, respectivamente), y los reportes de psicología también se asociaron negativamente (r = -0.27). Estos resultados son consistentes con la literatura: la asociación entre inasistencias, comportamiento y desempeño académico ya había sido documentada, y el presente estudio la confirma en una muestra

ampliada, aunque, al tratarse de un diseño transversal no experimental, estas relaciones deben interpretarse como asociaciones y no como relaciones causales.

Figura 2. Mapa de calor de correlación entre variables



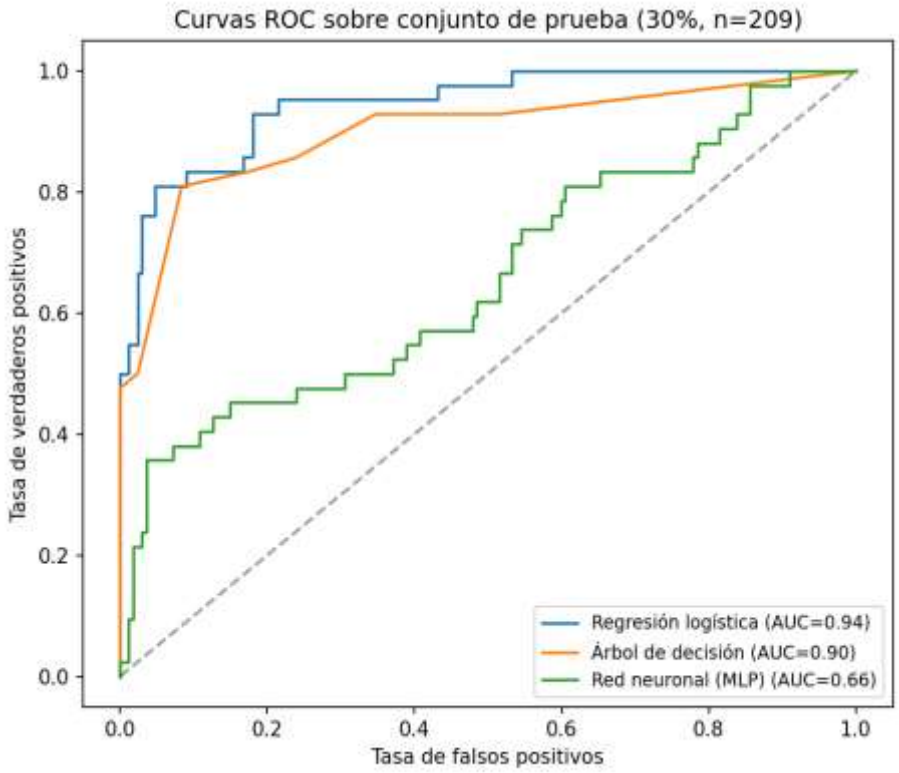
Agrupamiento de estudiantes

La aplicación de K-Means (k=3, seleccionado por el método del codo y confirmado con el coeficiente de silueta) y DBSCAN permitió segmentar a los 695 estudiantes activos en grupos con características similares en comportamiento, inasistencias y reportes psicológicos. K-Means produjo tres clústeres con coeficiente de silueta de 0.164: un grupo pequeño (n=74) que concentró la mayor cantidad de reportes psicológicos (media = 1.66 por estudiante) y el promedio general más bajo, un segundo grupo (n=286) con pocos reportes psicológicos, y un tercer grupo mayoritario (n=335) con el promedio general más alto. DBSCAN, con los parámetros seleccionados mediante búsqueda en cuadrícula que maximizó el coeficiente de silueta (eps=1.8, min_samples=8), identificó 3 clústeres con coeficiente de silueta de 0.03, clasificando como ruido al 23.3% de los estudiantes; su desempeño global fue menos consistente que K-Means, aunque resultó útil para señalar casos atípicos. Los coeficientes de silueta obtenidos (0.03-0.16) indican una separación entre clústeres moderada a baja, por lo que estos grupos deben interpretarse como perfiles exploratorios y no como categorías definitivas.

Predicción del riesgo académico

Con base en la curva ROC obtenida sobre la partición de prueba del 30% (n=209 estudiantes; Figura 3) y en los resultados promediados por validación cruzada de 5 particiones sobre los 695 estudiantes activos, la Regresión Logística resultó el modelo más confiable para predecir el riesgo académico (AUC promedio en validación cruzada = 0.94), seguida del Árbol de Decisión (AUC = 0.90). La Red Neuronal presentó el desempeño más bajo de los tres (AUC promedio en validación cruzada = 0.70; AUC en conjunto de prueba = 0.66, véase Figura 3), probablemente por el tamaño muestral limitado para un modelo de mayor complejidad paramétrica. Los valores de AUC observados en la partición de prueba única (Figura 3) difieren ligeramente de los promedios de validación cruzada por tratarse de una única muestra aleatoria; se reporta la validación cruzada como estimación principal por ser más estable.

Figura 3. Curva ROC de las técnicas de predicción



La Tabla 3 resume las métricas obtenidas para cada modelo.

Tabla 3. Resultados de las métricas de evaluación por modelo (validación cruzada 5-fold; media).

Modelos	Técnica	Métrica principal*	F1-Score	Recall
Agrupamiento	K-Means	0.16*	-	-
Agrupamiento	DBSCAN	0.03*	-	-
Predictivo	Regresión logística	0.60	0.70	0.85
Predictivo	Árbol de decisión	0.53	0.64	0.84
Predictivo	Redes neuronales	0.64	0.40	0.31

*Para K-Means y DBSCAN, la columna "Métrica principal" reporta el coeficiente de silueta (no existe una métrica de precisión en sentido supervisado, al ser técnicas de agrupamiento); para los modelos predictivos, la misma columna reporta la precisión (precision).

Los modelos de agrupamiento presentaron coeficientes de silueta bajos a moderados (0.03 a 0.16), lo que evidencia una separación limitada entre grupos y confirma su carácter exploratorio más que predictivo. Entre los modelos predictivos evaluados mediante validación cruzada sobre los 695 estudiantes activos, la Regresión Logística obtuvo el mejor equilibrio entre precisión (0.60), recall (0.85) y F1-score (0.70), además del AUC-ROC más alto (0.94). El Árbol de Decisión, con la profundidad óptima de 5 niveles determinada por búsqueda en cuadrícula, obtuvo un desempeño cercano (AUC = 0.90; F1 = 0.64). La Red Neuronal mostró el desempeño más bajo (AUC promedio en validación cruzada = 0.70; AUC en conjunto de prueba = 0.66; recall = 0.31), probablemente porque su mayor número de parámetros requiere una muestra más grande que la disponible (n=695) para generalizar adecuadamente. A diferencia del análisis preliminar con una muestra reducida (donde el Árbol de Decisión alcanzaba métricas perfectas de 1.0, señal típica de sobreajuste o fuga de datos), estos resultados, obtenidos sin variables con fuga de información y con balanceo de clases, son más conservadores, pero considerablemente más creíbles y generalizables.

Estos resultados son coherentes con hallazgos previos en minería de datos educativa. Cornell-Farrow y Garrard (2020), al comparar modelos de machine learning frente a la regresión logística para predecir riesgo académico en un millón de estudiantes australianos, encontraron que ningún modelo de machine learning superó de forma consistente a la regresión logística, resultado que el presente estudio reproduce a menor escala: la Regresión Logística (AUC=0.94) superó tanto al Árbol de Decisión (AUC=0.90)

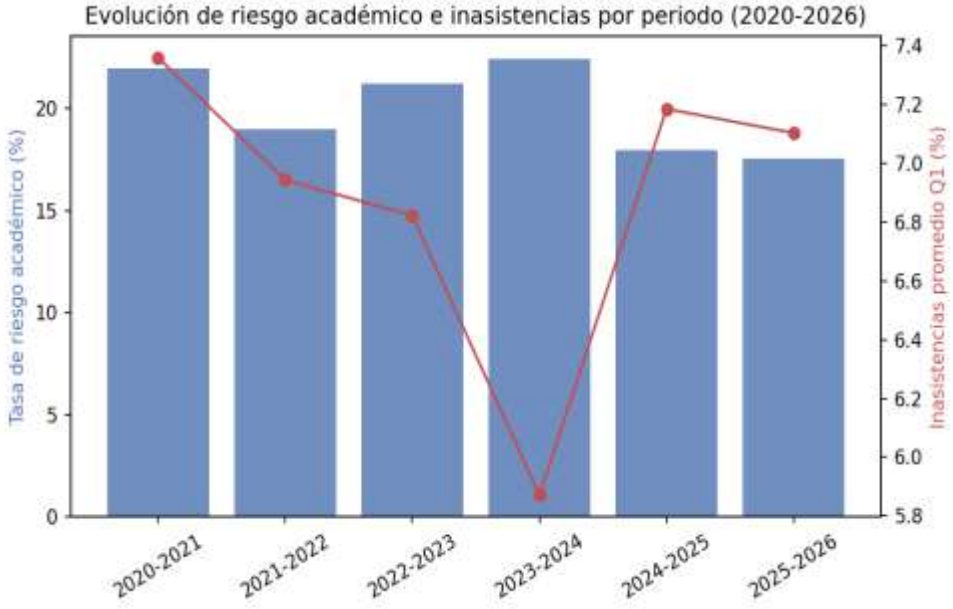
como a la Red Neuronal (AUC=0.70). Yağcı (2022) reportó una exactitud de clasificación de 70-75% al predecir el rendimiento final de estudiantes universitarios con modelos similares, un rango de desempeño comparable al obtenido aquí. Asimismo, Marcolino et al. (2025), en un estudio de predicción de deserción estudiantil basado en registros de plataformas LMS, reportaron valores de AUC-ROC entre 0.70 y 0.95 según la configuración del modelo, rango dentro del cual se ubican los resultados de este estudio. Esta coincidencia respalda la validez externa relativa de los hallazgos, aunque (como se detalla en la sección de limitaciones) la generalización a otras instituciones requeriría validación adicional.

Para hacer explícito el criterio de selección de variables aplicado durante el preprocesamiento y evitar ambigüedad sobre el mecanismo de fuga de datos evitado, se detalla a continuación: Variables predictoras utilizadas: EDAD, GÉNERO, COMPORTAMIENTO, PORCENTAJE_INASISTENCIAS_QUIMESTRE1, PORCENTAJE_INASISTENCIAS_QUIMESTRE2, ASISTENCIA_PADRES_CITACIONES, TIPO_APODERADO, SITUACION_SOCIOECONOMICA, NUMERO_LLAMADOS_DEL_FAMILIAR y REPORTES_PSICOLOGIA. Variables descartadas por fuga de información (data leakage): PROM_QUIMESTRE_1, PROM_QUIMESTRE_2, PARCIAL1_Q1, PARCIAL2_Q1, EXAMEN_Q1, PARCIAL1_Q2, PARCIAL2_Q2 y EXAMEN_Q2, todas ellas descartadas por ser componentes matemáticos directos de PROMEDIO_GENERAL (la variable a partir de la cual se define la etiqueta de riesgo académico). También se descartó MATERIA, al agregarse los registros a nivel estudiante-período, y MOTIVO_REPORTE_PSICOLOGIA y ESTADO_MATRICULA, utilizadas únicamente para el filtrado de la muestra y no como predictores.

Como aporte adicional, aprovechando que el nuevo conjunto de datos cubre seis periodos lectivos (2020-2021 a 2025-2026), se generó un gráfico de evolución temporal (Figura 4) de la tasa de riesgo académico y del porcentaje de inasistencias por cohorte, calculado sobre los 695 estudiantes activos (114 a 118 por periodo). La tasa de riesgo académico se mantuvo relativamente estable entre 17.5% y 22.4% a lo largo de los seis periodos, sin una tendencia sostenida al alza o a la baja, mientras que el porcentaje de inasistencias del primer quimestre mostró mayor variabilidad, con un descenso puntual en el periodo 2023-2024 (5.87%) frente a un promedio general cercano al 6.9%. Esta relativa estabilidad interanual

sugiere que el riesgo académico en la institución no está asociado a un evento coyuntural único, sino a un patrón estructural que amerita intervención sostenida en el tiempo.

Figura 4. Evolución de la tasa de riesgo académico y las inasistencias por periodo lectivo (2020-2026)



Lo anterior no es casual: la literatura ya había advertido que la asociación entre inasistencias, comportamiento y bajo rendimiento académico es un patrón recurrente (Pérez Peñaranda et al. 2009; Finning et al. 2019), y el presente estudio la confirma en el contexto de la institución analizada, con una muestra sustancialmente mayor y una evaluación metodológicamente más rigurosa que la de un estudio de corte único, aportando además un modelo predictivo funcional (Regresión Logística, AUC = 0.94) que puede integrarse como herramienta de apoyo a la gestión académica, sujeto a las limitaciones descritas en la sección siguiente.

Limitaciones del estudio

El estudio presenta limitaciones que deben considerarse al interpretar los resultados. Primero, los datos provienen de una única institución educativa, por lo que los hallazgos no son directamente generalizables a otros contextos sin validación externa. Segundo, el diseño es transversal y no experimental, por lo que las asociaciones reportadas no permiten establecer relaciones de causalidad. Tercero, aunque se amplió considerablemente el tamaño muestral (n=695 estudiantes activos, tras excluir 32 casos de retiro o traslado) respecto de la versión preliminar de este estudio (un solo periodo lectivo), la clase "riesgo académico" sigue siendo minoritaria (20%, 139 de 695), lo que limita el

desempeño alcanzable incluso con balanceo de clases. Cuarto, si bien la profundidad del árbol de decisión y los parámetros de DBSCAN se seleccionaron mediante búsqueda en cuadrícula, la arquitectura de la red neuronal (número de capas y neuronas) se fijó de forma conservadora sin una búsqueda sistemática exhaustiva, por lo que su desempeño podría mejorar con una optimización más amplia. Finalmente, los coeficientes de silueta obtenidos para K-Means y DBSCAN (0.03-0.16) indican una separación entre clústeres moderada a baja, por lo que los perfiles identificados deben tratarse como exploratorios. En síntesis, el estudio utilizó información proveniente de una única institución educativa, por lo que futuras investigaciones deberían incorporar centros educativos de diferentes contextos geográficos y socioeconómicos para fortalecer la generalización del modelo antes de cualquier aplicación institucional más amplia.

Consideraciones éticas y manejo de datos

La base de datos institucional incluye variables sensibles (reportes y motivos de atención psicológica, situación socioeconómica), por lo que los identificadores de estudiante utilizados en el análisis son códigos anonimizados sin datos personales asociados, y los resultados se reportan exclusivamente de forma agregada. Por motivos de confidencialidad, el nombre de la institución se mantiene anonimizado en este artículo; se identifica únicamente como una institución de educación secundaria de la ciudad de Quevedo, Ecuador. El uso de la información contó con una autorización institucional firmada por las autoridades del plantel para fines de investigación académica. Se recomienda que cualquier uso operativo del modelo predictivo sea supervisado por personal del departamento de psicología o consejería estudiantil, y no se utilice como criterio único o automático de decisión sobre un estudiante.

CONCLUSIONES

En respuesta al objetivo de identificar los factores críticos que inciden en el riesgo académico, el estudio confirma que el comportamiento estudiantil y las inasistencias son los factores con mayor asociación al riesgo académico (r=0.85 y r=-0.60 a -0.55, respectivamente), seguidos de la asistencia de los padres a citaciones y los reportes de psicología. Este patrón es consistente a lo largo de las seis cohortes analizadas (2020-2021 a 2025-2026), lo que sugiere que se trata de un factor estructural y no de una situación coyuntural de un año específico.

En cuanto al objetivo de predecir el riesgo académico mediante minería de datos, la Regresión Logística resultó el modelo más adecuado para este conjunto de datos (AUC-ROC = 0.94), un resultado consistente con estudios previos que muestran que modelos más complejos no siempre superan a la regresión logística en la predicción de riesgo académico. Esto, sumado a que evitar la fuga de datos y balancear las clases produjo métricas realistas (en contraste con las métricas perfectas (1.0) del análisis preliminar con una muestra reducida), confirma que la metodología corregida es sustancialmente más confiable para un eventual uso institucional.

Las técnicas de agrupamiento (K-Means y DBSCAN) cumplieron un rol complementario, más exploratorio que predictivo dado el nivel de separación moderado a bajo entre clústeres, permitiendo no obstante identificar un perfil de estudiantes con alta concentración de reportes psicológicos que podría orientar estrategias de intervención diferenciadas.

Finalmente, se recomienda a la institución impulsar programas de acompañamiento conductual e involucramiento familiar como estrategia complementaria a la implementación de un modelo predictivo de minería de datos, validando y actualizando dicho modelo a medida que se disponga de nuevas cohortes de estudiantes, y sujeto a las limitaciones y consideraciones éticas descritas en las secciones correspondientes.

Conflicto de intereses: Los autores declaran no tener ningún conflicto de intereses.

REFERENCIAS BIBLIOGRÁFICAS

Akçapınar, G., Altun, A., & Aşkar, P. (2019). Using learning analytics to develop early-warning system for at-risk students. International Journal of Educational Technology in Higher Education, 16(1), 40. https://doi.org/10.1186/s41239-019-0172-7

Birioukov, A. (2016). Absenteeism. Encyclopedia of Personality and Individual Differences, 1-4. https://doi.org/10.1007/978-3-319-28099-8_725-1

Chang, Y.-H., Chen, F.-C., & Lee, C.-I. (2025). Developing an early warning system with personalized interventions to enhance academic outcomes for at-risk students in Taiwanese higher education. Education Sciences, 15(10), 1321. https://doi.org/10.3390/educsci15101321

Cornell-Farrow, S., & Garrard, R. (2020). Machine learning classifiers do not improve the prediction of academic risk: Evidence from Australia. https://doi.org/10.1080/23737484.2020.1752849

Farid, F., & Donyatalab, Y. (2023). Unlocking insights into mental health and productivity in the tech industry through KDD and data visualization. Lecture Notes in Networks and Systems, 759, 90-98. https://doi.org/10.1007/978-3-031-39777-6_11

Finning, K., Ukoumunne, O. C., Ford, T., Danielson-Waters, E., Shaw, L., Romero De Jager, I., Stentiford, L., & Moore, D. A. (2019). Review: The association between anxiety and poor attendance at school - a systematic review. Child and Adolescent Mental Health, 24(3), 205-216. https://doi.org/10.1111/CAMH.12322

Henry, K. L. (2007). Who's skipping school: Characteristics of truants in 8th and 10th grade. The Journal of School Health, 77(1), 29-35. https://doi.org/10.1111/j.1746-1561.2007.00159.x

Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., & Cechinel, C. (2025). Student dropout prediction through machine learning optimization: Insights from Moodle log data. Scientific Reports, 15, 9840. https://doi.org/10.1038/s41598-025-93918-1

Mohd Talib, N. I., Abd Majid, N. A., & Sahran, S. (2023). Identification of student behavioral patterns in higher education using K-Means clustering and support vector machine. Applied Sciences, 13(5), 3267. https://doi.org/10.3390/app13053267

Pan, H. (2022). Data mining framework of public sports budget performance evaluation index based on smart data transmission algorithm. 4th International Conference on Inventive Research in Computing Applications (ICIRCA 2022), 1035-1038. https://doi.org/10.1109/ICIRCA54612.2022.9985467

Pérez Peñaranda, A., García Ortiz, L., Rodríguez Sánchez, E., Losada Baltar, A., Porras Santos, N., & Gómez Marcos, M. Á. (2009). Función familiar y salud mental del cuidador de familiares con dependencia. Atención Primaria, 41(11), 621-628. https://doi.org/10.1016/j.aprim.2009.03.005

Tsui, K. L., Chen, V., Jiang, W., Yang, F., & Kan, C. (2023). Data mining methods and applications. Springer Handbooks, 797-816. https://doi.org/10.1007/978-1-4471-7503-2_38

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

Zhang, Y., & Paquette, L. (2023). Sequential pattern mining in educational data: The application context, potential, strengths, and limitations. En Mining Educational Data (pp. 219-254). https://doi.org/10.1007/978-981-99-0026-8_6