Asesorías y Tutorías para la Investigación Científica en la Educación Puig-Salabarría S.C.
José María Pino Suárez 400-2 esq a Lerdo de Tejada. Toluca, Estado de México. 7223898475
RFC: ATI120618V12

Revista Dilemas Contemporáneos: Educación, Política y Valores.
http://www.dilemascontemporaneoseducacionpoliticayvalores.com/
Año: XIII Número: 2 Artículo no.:43 Período: 1 de enero del 2026 al 30 de abril del 2026

TÍTULO: Descubriendo patrones educativos en educación superior mediante Técnicas de Minería de Datos.

AUTORES:

1. Dr. Vitervo López Caballero.
2. Dra. Lucia Morales Morales.
3. Dra. Xochitl Morales Morales.

RESUMEN: La Minería de Datos Educativos (EDM) aplica métodos de ciencia de datos para analizar información académica y apoyar la toma de decisiones basada en evidencia. Este artículo de investigación, realizado en una Institución de Educación Superior en México, identifica patrones asociados con éxito o riesgo académico en estudiantes de licenciatura mediante cuatro fases: recopilación, depuración, modelado y análisis. Con técnicas no supervisadas como PCA y K-means se detectaron tres perfiles de estudiantes, uno de ellos con señales de desinterés emocional o académico. En la etapa predictiva, modelos supervisados como MLP, SVM y Regresión Logística alcanzaron precisiones del 95–96%. Los resultados muestran que el aprendizaje automático es clave para identificar estudiantes en riesgo y mejorar la intervención institucional.

PALABRAS CLAVES: minería de datos educativa, K-means, análisis de componentes principales, deserción estudiantil, aprendizaje automático.

TITLE: Discovering educational patterns in higher education through Data Mining Techniques.

AUTHORS:

1. PhD. Vitervo López Caballero.
2. PhD. Lucia Morales Morales.
3. PhD. Xochitl Morales Morales.

ABSTRACT: Educational Data Mining (EDM) applies data science methods to analyze academic information and support evidence-based decision-making. This research article, conducted at a Higher Education Institution in Mexico, identifies patterns associated with academic success or risk among undergraduate students through four phases: data collection, cleaning, modeling, and analysis. Using unsupervised techniques such as PCA and K-means, three student profiles were identified, one of which exhibited signs of emotional or academic disengagement. In the predictive stage, supervised models such as MLP, SVM, and Logistic Regression achieved accuracies of 95–96%. The findings demonstrate that machine learning is a key tool for identifying students at risk and strengthening institutional intervention.

KEY WORDS: educational data mining, K-means, principal component analysis, student dropout, machine learning.

INTRODUCCIÓN.

La Minería de Datos Educativa (Educational Data Mining, EDM) es una disciplina en constante crecimiento que se ha consolidado como una herramienta importante para el análisis de grandes volúmenes de datos generados en contextos académicos. Su principal objetivo es descubrir patrones y tendencias ocultas en los datos, con el fin de generar conocimiento útil que apoye a la toma de decisiones en el ámbito educativo, particularmente en la mejora del rendimiento estudiantil (Trung et al., 2023).

Uno de los principales desafíos en el campo de la EDM radica en tres aspectos críticos que limitan su efectividad:

(1) Limitaciones algorítmicas, ya que muchos modelos utilizados para predecir el rendimiento académico y la deserción escolar presentan deficiencias tanto en precisión como en aplicabilidad contextual.

(2) Dificultades en la integración de datos heterogéneos, dado que una aplicación efectiva de EDM exige combinar datos sociales, conductuales y cognitivos, lo cual se ve obstaculizado por la ausencia de formatos estandarizados y sistemas interoperables.

(3) Baja transferencia a la práctica educativa, ya que persiste una brecha significativa entre los avances teóricos de la EDM y su implementación real en instituciones educativas, especialmente en regiones en vías de desarrollo (Colpo et al., 2024).

En respuesta a estas limitaciones, la presente investigación se orienta al análisis de datos académicos de una Institución de Educación Superior (IES) con el fin de identificar patrones relevantes que permitan predecir el éxito o fracaso académico de estudiantes de nivel licenciatura; para ello, se desarrolla un caso de estudio mediante la aplicación de algoritmos de aprendizaje automático, tanto supervisado como no supervisado, siguiendo una metodología sistemática que contempla las siguientes etapas: recolección y preprocesamiento de datos, selección y validación de modelos, e interpretación de resultados. Este enfoque busca contribuir al cierre de la brecha entre la teoría y la práctica, ofreciendo herramientas analíticas con potencial de aplicación directa en contextos educativos reales.

Los resultados obtenidos mediante los algoritmos de aprendizaje no supervisado permitieron identificar tres grupos distintos de estudiantes; entre ellos, el grupo uno se destacó, ya que mostró indicios de posible desvinculación emocional o académica, lo que sugiere la necesidad de implementar estrategias de intervención temprana tales como programas de orientación psicológica, mentoría personalizada y asesoramiento vocacional. Adicionalmente, se entrenaron varios modelos de aprendizaje supervisado para evaluar su capacidad predictiva. Los algoritmos que demostraron el mejor desempeño fueron el Perceptrón Multicapa (Multilayer Perceptron, MLP), las Máquinas de Soporte Vectorial (Support Vector Machines, SVM) y la Regresión Logística (Logistic Regression, LR). Estos modelos alcanzaron una precisión global (accuracy) del 95%, una precisión (precision) del 96%, una sensibilidad (recall) del 96% y un puntaje F1

(F1-score) también del 96%, lo que evidencia su alta capacidad para clasificar correctamente estudiantes nuevos en función de los datos analizados.

La estructura de este artículo se organiza de la siguiente manera: en la Sección del desarrollo se describen los trabajos relacionados, se describe detalladamente el método utilizado y se presenta el caso de estudio, con el fin de ilustrar paso a paso la aplicación de la metodología. Finalmente, se presentan las conclusiones del estudio.

DESARROLLO.

Revisión de la literatura.

El propósito de esta investigación no es comparar directamente el desempeño de algoritmos con trabajos previos; no obstante, resulta pertinente realizar una revisión de la literatura para identificar los algoritmos más empleados en el campo de la Minería de Datos Educativa. En esta sección se analizan diversos estudios que han aplicado algoritmos de aprendizaje automático para el análisis de datos académicos, proporcionando un marco de referencia contextual que sustenta la presente investigación.

Mengash (2020) aplicó técnicas de minería de datos educativa para evaluar y predecir el rendimiento académico de estudiantes universitarios. En su estudio se entrenaron cuatro algoritmos de aprendizaje automático: las redes neuronales, los árboles de decisión, las máquinas de vectores de soporte (SVM) y Naive Bayes, utilizando un conjunto de datos conformado por 2,039 estudiantes de una universidad pública en Arabia Saudita. Los resultados mostraron que el modelo basado en redes neuronales obtuvo el mejor desempeño, alcanzando una precisión del 79%.

Por su parte, A. Khan & Ghosh (2021) realizaron una revisión sistemática de la literatura centrada en la modelización del rendimiento académico en instituciones de educación superior, considerada una de las problemáticas más relevantes y complejas dentro de la minería de datos educativa. Los autores examinaron 140 artículos publicados entre 2000 y 2018, todos relacionados con la predicción del rendimiento estudiantil, para lo cual emplearon diversas cadenas de búsqueda en Google Scholar. La revisión permitió

identificar los predictores más utilizados, los métodos empleados, el momento en que se realiza la predicción y sus propósitos principales; asimismo, el meta-análisis evidenció futuras líneas de investigación, entre ellas la predicción temprana del rendimiento académico antes del inicio del curso.

De manera complementaria, Xiao et al (2022) reforzaron la importancia de la predicción del rendimiento estudiantil como uno de los principales retos en la minería de datos educativa, especialmente por la necesidad de contar con modelos altamente interpretables. Su estudio consistió en una revisión bibliográfica de trabajos publicados entre los años 2016 y 2021, con el objetivo de identificar los factores que influyen de manera significativa en el proceso de aprendizaje. Los resultados obtenidos ofrecen insumos valiosos para la toma de decisiones orientadas a mejorar el desempeño académico.

Investigaciones recientes también han demostrado, que las técnicas de minería de datos educativa permiten detectar patrones ocultos en el comportamiento académico, los cuales pueden emplearse para anticipar el éxito estudiantil a partir de datos históricos; asimismo, se reconoce la influencia de factores sociales en el rendimiento académico (M. Khan et al., 2023). En este estudio, los autores analizaron diversas técnicas de minería de datos para examinar patrones de conducta y predecir el desempeño estudiantil, encontrando una correlación significativa entre el rendimiento académico y múltiples factores tanto académicos como sociales. Estos hallazgos permiten identificar a estudiantes en riesgo que podrían beneficiarse de intervenciones oportunas por parte del profesorado.

Finalmente, Meneses Claudio (2024) reporta que la deserción estudiantil durante los primeros años universitarios continúa siendo un problema crítico a nivel global. De acuerdo con datos de la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO), la tasa de abandono escolar supera el 30%. Con base en este contexto, los autores presentan un estudio en el que se emplean diversas técnicas de minería de datos educativa, entre ellas el análisis de conglomerados, los árboles de decisión y las redes neuronales. El objetivo fue predecir el rendimiento académico e identificar los factores más influyentes en la deserción de estudiantes de Ingeniería de la Universidad Nacional Autónoma de México

durante 2022. La validación de los modelos mostró una precisión promedio del 91.2%, con una desviación estándar del 2.7%.

Discusión de la revisión de la literatura.

Con base en la revisión de la literatura presentada, se puede observar que existe un consenso claro respecto a la utilidad de la Minería de Datos Educativa como una herramienta poderosa para analizar grandes volúmenes de datos académicos y predecir el rendimiento estudiantil. Esta evidencia respalda la relevancia de nuestra investigación, aunque el propósito no sea comparar directamente los algoritmos, sino proponer un enfoque contextualizado y metodológicamente sólido.

Los estudios analizados destacan el uso recurrente de algoritmos como redes neuronales, árboles de decisión, máquinas de soporte vectorial (SVM) y Naive Bayes, lo cual refleja una tendencia consolidada hacia el empleo de modelos tanto interpretables como robustos. En particular, se observa que las redes neuronales han sido ampliamente adoptadas debido a su capacidad de modelar relaciones complejas en los datos, alcanzando altos niveles de precisión como en el caso de Mengash (2020), donde se logró un 79% de acierto en la predicción del rendimiento académico.

Metodología.

Para llevar a cabo nuestro estudio se utiliza una metodología que consta de cuatro pasos, los cuales son: ***recolección de datos y validación, preprocesamiento, selección de modelos y validación e interpretación de resultados***. A continuación, se describe cada una de las cuatro etapas de la metodología.

a) Recolección de datos.

En esta etapa se elabora un cuestionario que incorpora diversos factores asociados al desempeño estudiantil, entre ellos variables académicas, socioeconómicas, psicológicas, motivacionales e institucionales. Para validar la calidad y pertinencia de los ítems que conforman el instrumento, se emplearon dos métricas ampliamente reconocidas: la V de Aiken, utilizada para evaluar la validez de

contenido (Merino-Soto, 2023), y el alfa de Cronbach, destinado a medir la consistencia interna del cuestionario (Tavakol & Dennick, 2011).

b) Preprocesamiento de la información recolectada.

En esta etapa, se aplica una técnica de preprocesamiento de datos denominada OneHotEncoder. La técnica OneHotEncoder se utiliza en aprendizaje automático para convertir variables categóricas a una representación numérica binaria, esto para que los algoritmos de aprendizaje automático lo puedan entender (Kosaraju et al., 2023).

c) Selección de modelos y validación.

Una vez procesada la información a una representación numérica, de tal modo que los algoritmos lo puedan entender, se aplican técnicas de aprendizaje no supervisado como son algoritmos de reducción de dimensiones y algoritmos de agrupamiento de objetos; esto para encontrar patrones en los datos. Una vez agrupada la información, esta es dividida en 2 conjuntos, el 80% para entrenar los modelos de aprendizaje supervisado y el 20% para probar los modelos. Los modelos de aprendizaje supervisado que se seleccionan son: Perceptrón Multicapa, las Máquinas de Soporte Vectorial, Regresión Logística, y Random Forest. Estos modelos se seleccionan, ya que son de los algoritmos más utilizados para análisis de datos académicos, de acuerdo a la literatura reportada en la sección II de este documento.

Para validar los modelos seleccionados, se genera la matriz de confusión para cada modelo y se generan las métricas de Exactitud (Accuracy), Precisión, Sensibilidad (Recall) y F1-score. La matriz de confusión es una herramienta fundamental para evaluar el rendimiento de los modelos de clasificación, ya que proporcionan una forma estructurada de visualizar los resultados de las predicciones frente a los resultados reales. La métrica de exactitud proporciona la relación entre las instancias predichas correctamente y el total de instancias. La métrica de precisión proporciona la relación entre los verdaderos positivos y la suma de los verdaderos y falsos positivos, que indican la calidad de las predicciones positivas. La métrica de sensibilidad proporciona la relación entre los verdaderos positivos y la suma de los verdaderos positivos

y los falsos negativos, que refleja la capacidad del modelo para identificar instancias relevantes. La métrica de F1-score describe la media armónica de la precisión y la sensibilidad, describe un equilibrio entre ambas métricas (Sowmiya Narayanan & Manimaran, 2024).

d) Interpretación de resultados.

Una vez obtenido los resultados por los modelos, es necesario realizar una interpretación de los resultados, ya que los modelos sólo preprocesan la información, pero no arrojan recomendaciones para tomar decisiones. Es necesario, que el experto en los datos interprete los resultados obtenidos por los modelos para brindar recomendaciones al cliente, y de esta manera, se tomen buenas decisiones.

Caso de Estudio.

Con el propósito de contribuir al cierre de la brecha entre la teoría y la práctica de la EDM, se toma como caso de estudio una Institución de Educación Superior (IES), en este caso el Instituto Tecnológico de Iztapalapa III. Esta institución se selecciona por darnos las facilidades de aplicar EDM en sus estudiantes. A continuación, se describe cada una de las actividades que se siguieron para aplicar las técnicas de minería de datos educativa con el propósito de detectar patrones en los datos.

Recopilación de información y selección de la muestra.

En este apartado se detalla cómo fue que se construye el cuestionario y se selecciona el tamaño de la muestra.

Construcción del cuestionario.

En esta etapa se construye un cuestionario que contiene 31 preguntas agrupadas en cuatro factores los cuales son: factor académico, factor socioeconómico, factor psicológico y motivacional, y finalmente, el factor institucional. Para más información sobre el cuestionario construido, se invita al lector a consultar el siguiente enlace: Hacer clic para consultar cuestionario.

Validación del cuestionario.

Para validar el cuestionario, se utiliza la métrica de V de Aiken, la cual se muestra en la Ecuación 1. La métrica de V de Aiken se evalúa antes de aplicar el cuestionario a los estudiantes y su propósito es validar el contenido de las preguntas que conforman el cuestionario. La métrica de V de Aiken consiste en solicitar a un grupo de expertos que califiquen cada pregunta del cuestionario en una escala ordinaria con respecto a su relevancia, claridad o coherencia. En la presente investigación, la muestra de expertos que se selecciona fue de 8 personas expertas en el área de psicología y la escala que se utiliza fue una escala discreta categórica (No cumple con el criterio, Poco relevante, Neutral, Preciso y Calidad relevante). La fórmula de la V de Aiken es la siguiente:

$$V = \frac{\sum(x_i - l)}{N(k - l)} \qquad (1)$$

donde:

$x_i = Puntuación\ otorgoda\ por\ el\ experto\ i.$

$l = menor\ puntuación\ posible\ en\ la\ escala.$

$k = mayor\ puntuación\ posible\ en\ la\ escala.$

$N = número\ de\ expertos\ que\ calificaron\ el\ item.$

Para más información del cuestionario aplicado a los expertos en psicología, se invita al lector a consultar el cuestionario en el siguiente enlace: hacer clic para consultar el cuestionario. Como se muestra, cada pregunta cuenta con una respuesta en la siguiente escala: No cumple con el criterio (valor 1), poco relevante (valor 2), neutral (valor 3), preciso (valor 4) y finalmente calidad relevante (valor 5).

Selección de la muestra.

Una vez construido y validado el cuestionario por un conjunto de expertos en el área de psicología, se selecciona una muestra de un total de 430 estudiantes. El tamaño de la muestra se selecciona debido a que sólo 430 estudiantes mostraron disponibilidad e interés para responder objetivamente un listado de 31 peguntas seleccionadas. El tamaño de la muestra está conformado como sigue: 177 son estudiantes de

cuarto semestre, 105 de segundo semestre, 50 de octavo semestre, 37 de quinto semestre, 18 estudiantes no egresados, 16 de noveno semestre, 12 de tercer semestre, 11 egresados y titulados. y finalmente. 4 de primer semestre.

Para validar las respuestas de los estudiantes, se utiliza la métrica de Alfa de Cronbach, la cual se muestra en la Ecuación 2. Esta métrica se utiliza después de aplicar el cuestionario y su objetivo es validar la consistencia interna y fiabilidad del conjunto de preguntas del cuestionario. La métrica del alfa de Cronbach indica el mejor valor de alfa cuando $\propto > 0.80$ lo que expresa una *"muy buena consistencia";* por otro lado, cuando $\propto > 0.70$ indica que es *"buena consistencia"*.

$$\propto = \frac{k}{k-1}\left(1 - \frac{\sum \sigma_i^2}{\sigma_T^2}\right) \qquad (2)$$

donde:

$\propto = coeficiente\ de\ fiabilidad\ de\ Alfa\ de\ Cronbach.$

$k = número\ de\ items\ del\ cuestionario.$

$\sigma_i^2 = varianza\ de\ item\ i.$

$\sigma_T^2 = varianza\ total\ de\ test.$

Con el propósito de ejemplificar el cálculo del coeficiente Alfa de Cronbach, en la Tabla 1 se presentan las respuestas de 13 estudiantes a un cuestionario conformado por 6 preguntas. A partir de estos datos, se calcularon las varianzas individuales de cada pregunta y la varianza total del instrumento, obteniendo un valor de Alfa de Cronbach de 0.75. Este resultado indica un nivel adecuado de consistencia interna entre los ítems del cuestionario.

$$\propto = \frac{6}{6-1}\left(1 - \frac{6.1065}{16.5207}\right) = 0.756447$$

Tabla 1. Ejemplo de cálculo de alfa de Cronbach.

No.	Q1	Q2	Q3	Q4	Q5	Q6	Total
1	1	5	4	3	2	4	19
2	4	3	3	4	3	4	21
3	3	4	4	4	4	5	24
4	2	3	3	3	3	5	19
5	4	3	5	3	4	5	24
6	4	5	4	3	2	5	23
7	4	3	3	3	3	3	19
8	3	3	3	2	2	5	18
9	5	4	4	5	4	4	26
10	1	3	3	2	1	1	11
11	3	4	4	2	4	3	20
12	1	3	3	2	2	4	15
13	2	5	4	4	5	5	25
Varianza	1.67	0.67	0.39	0.84	1.23	1.30	16.52

Preprocesamiento de información.

Una vez validado el cuestionario, el siguiente paso de la metodología es limpiar la información; de tal modo, que los algoritmos la puedan entender; para ello, se preprocesa la información verificando valores faltantes y valores infinitos; por otro lado, se codifican las respuestas a una representación numérica utilizando la técnica OneHotEncoder. La técnica OneHotEncoder es empleada en el campo de aprendizaje automático para convertir variables categóricas a una representación numérica binaria; esto para que los algoritmos de aprendizaje automático lo puedan entender.

En la Tabla 2 se muestra un ejemplo de la codificación según las respuestas de los estudiantes. La pregunta de ejemplo es la siguiente: ¿cuál fue tu promedio en el último semestre?, sus posibles respuestas son: 5 o menos-reprobado (x1), 6-suficiente (x2), 7-aprobado (x3), 8-bien (x4), 9-notable (x5) y 10-sobresaliente (x6). Aunque las respuestas tienen una relación ordinal, OneHotEncoder las convierte en variables

dummies binarizadas, generando una nueva columna por cada categoría, con valores de 0 y 1 según la respuesta del estudiante. Con esto se evita que el modelo asuma una relación numérica arbitraria entre las categorías. Para la pregunta de ejemplo mencionada anteriormente, la codificación aplicando OneHotEncoder se representa en la Tabla 2.

Tabla 2. Ejemplo de codificación OneHotEncoder.

Promedio	**x6**	**x1**	**x2**	**x3**	**x4**	**x5**
x1	0	1	0	0	0	0
x2	0	0	1	0	0	0
x3	0	0	0	1	0	0
x4	0	0	0	0	1	0
x5	0	0	0	0	0	1
x6	1	0	0	0	0	0

Selección de modelos y validación.

Una vez preprocesada la información, el siguiente paso consiste en aplicar modelos de aprendizaje no supervisado con el propósito de identificar patrones ocultos en los datos académicos. Tras codificar las variables categóricas mediante la técnica OneHotEncoder, la base de datos educativa alcanza una dimensión de 430 instancias por 148 atributos. El número de atributos incrementa debido a que la técnica OneHotEncoder toma a una pregunta y sus posibles respuestas, y estás las expresa como atributos sin considerar la pregunta original.

Dado el elevado número de atributos, se requiere una visualización efectiva de los datos que permita detectar tendencias y agrupaciones; para ello, se recurre a una técnica de reducción de dimensionalidad conocida como Análisis de Componentes Principales (Principal Component Analysis, PCA). El algoritmo PCA permite transformar el espacio original de atributos en un espacio de menor dimensión, conservando

la mayor cantidad de varianza posible (Alkandari & Aljaber, 2015). En esta investigación, se busca automáticamente el mejor número de componentes usando el criterio de varianza explicada acumulada; en este caso, el 95%. En la Figura 1 se muestra como la varianza explicada acumulada crece conforme se agregan más componentes principales. Se observa, que alrededor de 72 componentes principales explican el 95% de la varianza, lo cual es un umbral comúnmente utilizado para reducción de dimensiones sin perder demasiada información.

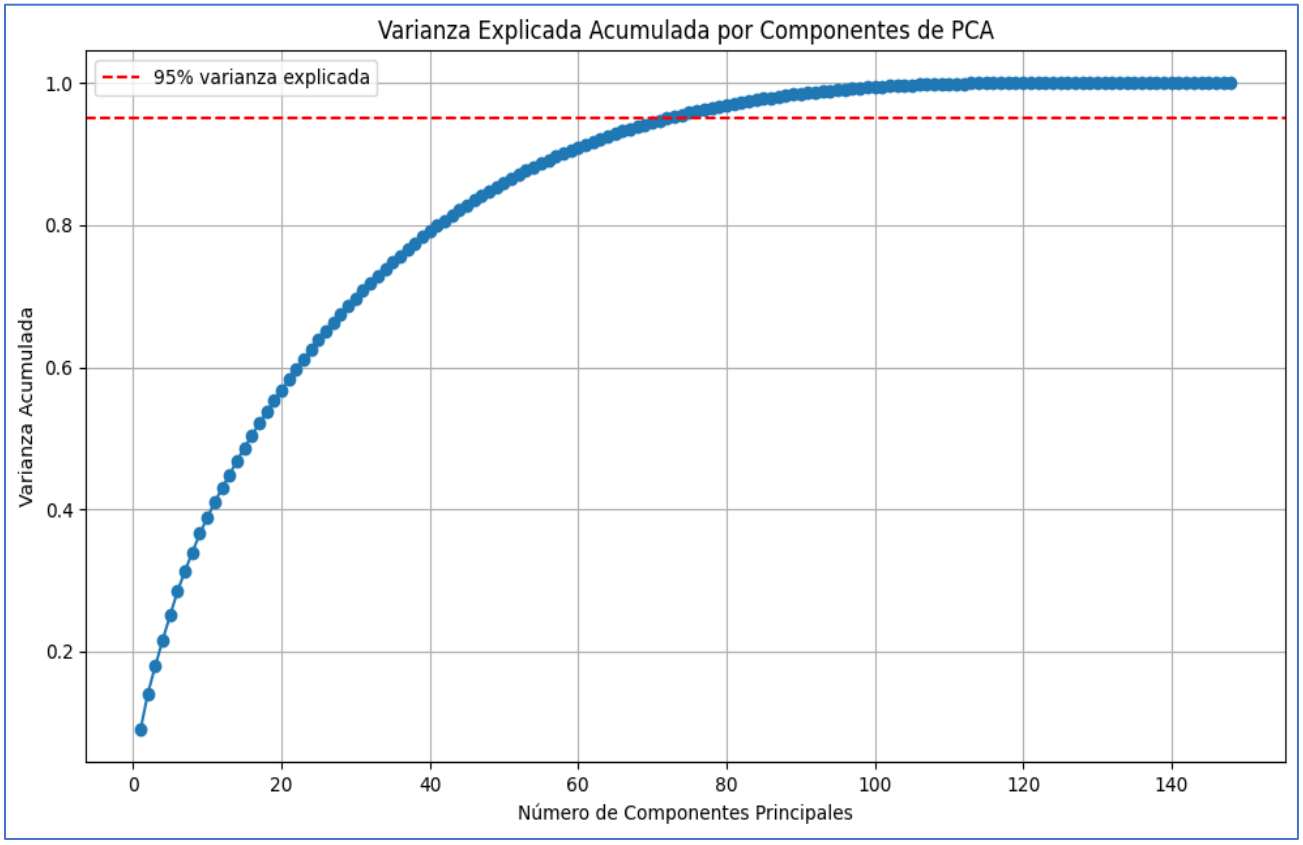

Figura 1. Varianza explicada acumulada.

El siguiente paso consiste en aplicar un algoritmo de aprendizaje no supervisado, específicamente K-means, con el objetivo de identificar patrones subyacentes en los datos académicos. Este algoritmo requiere definir previamente el número de grupos (K) que se desea formar. Para seleccionar un valor adecuado de K, se emplearon dos métodos ampliamente utilizados: el método del codo (elbow method)

(Shi et al., 2021) y el coeficiente de silueta promedio (Mamat et al., 2018). Los resultados obtenidos se presentan en la Figura 2.

En ambas representaciones se observa que la curva de inercia comienza a estabilizarse alrededor de K = 3 o K = 4, lo cual sugiere que estos valores podrían reflejar la estructura natural de los datos, y por tanto, ser los más apropiados para el agrupamiento.

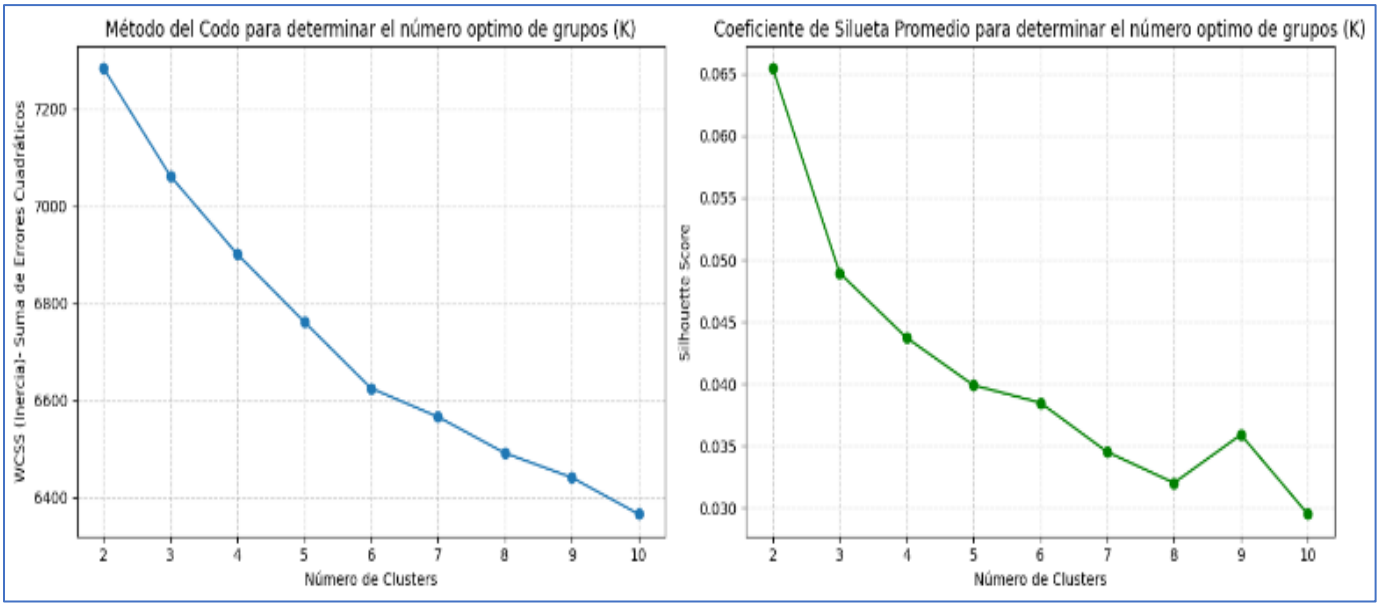

Figura 2. Método del codo y el método de coeficiente de silueta promedio para buscar el número de grupos adecuados.

Una vez identificado el valor adecuado de K, se procede a instanciar el algoritmo de agrupamiento K-means de Scikit-learn, configurándolo con el valor de K=3, una semilla fija (random_state=42) y un número de inicializaciones igual a 10 (n_init=10). El resultado final del agrupamiento se presenta en la Figura 3, y sólo se visualizan los dos primeros componentes principales del algoritmo PCA y en donde se identifican tres grupos principales: grupo 0 (color azul), grupo 1 (color naranja) y grupo 2 (color verde).

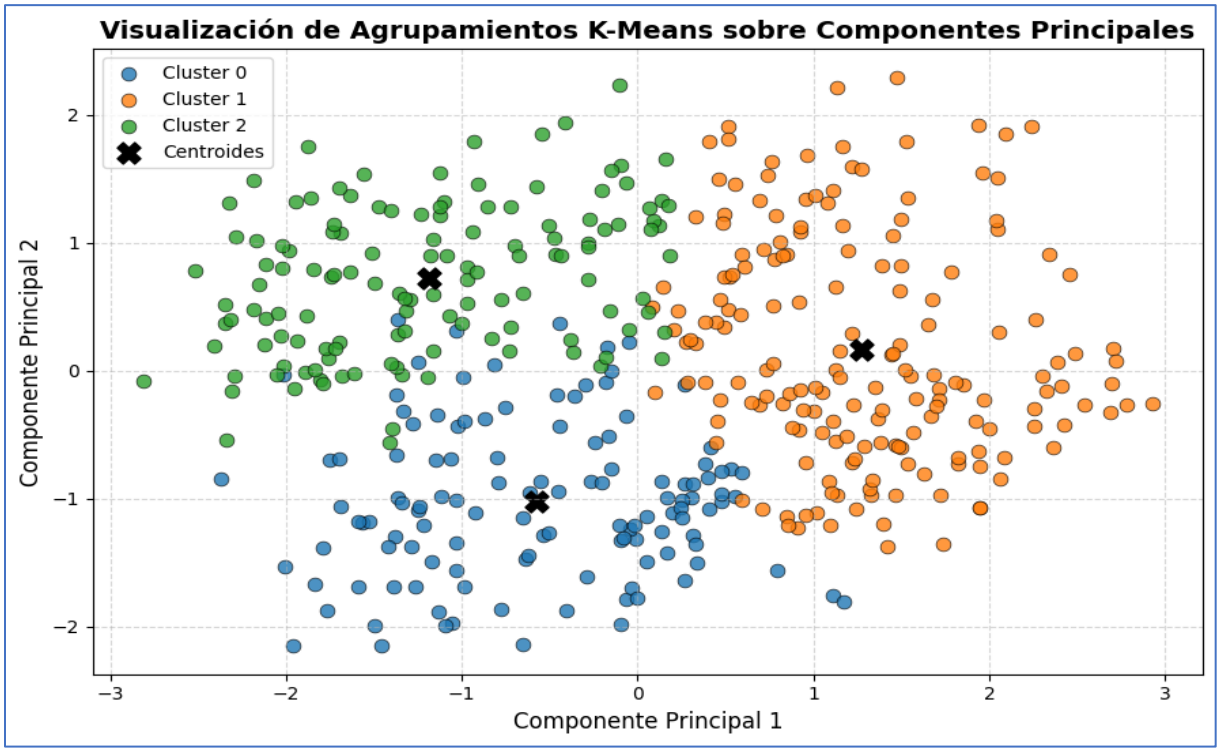

Figura 3. Visualización de los grupos con el algoritmo K-means después de aplicar PCA.

Para dar una interpretación a cada grupo se utilizan dos fuentes clave de información: cálculo de la moda por grupo, y análisis cualitativo de las variables dominantes. A continuación, se explica la interpretación detallada por grupo basada en la moda de las respuestas.

El grupo 0 (Tabla 3), es un grupo con madurez académica, buenos resultados, alto apoyo familiar, y buena percepción institucional. Se sugiere fomentar liderazgo y participación en tutorías o mentorías.

Tabla 3. Descripción del grupo 0 (color azul).

Estudiantes Técnicamente Satisfechos y Apoyados	
Edad:	Más de 21 años
Carrera:	Ingeniería civil
Sexo:	Hombres
Promedio:	Generalmente buenos (“8: bien”)
Satisfacción con enseñanza y métodos:	Alta (“Satisfechos”)

Motivación y apoyo familiar:	Alta (“Muy motivados”, “Siempre”)
Equilibrio vida-escolar:	Bien logrado
Abandono de estudios:	No, nunca.

El grupo 1 (Tabla 4) representa a estudiantes con posibles casos de desvinculación emocional o académica. Se sugiere atención mediante programas de orientación psicológica, mentoría personalizada, y asesoría vocacional.

Tabla 4. Explicación del grupo 1 (color naranja).

Estudiantes Jóvenes y Motivados	
Edad:	Más de 21 años
Carrera:	Ingeniería civil
Sexo:	Hombres
Promedio:	8: bien, es aceptable, pero percepción neutral en muchas respuestas.
Satisfacción con enseñanza y métodos:	Neutral
Motivación y apoyo familiar:	Neutral
Equilibrio vida-escolar:	Neutral
Abandono de estudios:	Posiblemente

El grupo 2 (Tabla 5), representa a estudiantes con alto compromiso y desempeño, probablemente en transición hacia curso avanzados. Se recomienda fortalecer programas de reconocimiento académico y liderazgo estudiantil.

Tabla 5. Explicación del grupo 2 (color verde).

Estudiantes Neutrales y Desconectados	
Edad:	20 años
Carrera:	Ingeniería en informática
Sexo:	Mujer
Promedio:	Generalmente buenos (“8: bien”)
Satisfacción con enseñanza y métodos:	Alta (“Satisfechas”)
Motivación y apoyo familiar:	Alta (“Muy motivadas”, “Siempre”)
Equilibrio vida-escolar:	Bien logrado
Abandono de estudios:	No, nunca.

Para validar los resultados del agrupamiento, se utilizan dos métricas: la primera denominada Índice de Calinski-Harabasz, y la segunda el Coeficiente de Silueta. El Índice de Calinski-Harabasz evalúa qué tan bien separados están los grupos. Cuánto más alto es el valor, mejor la separación inter-grupo frente a la compacidad intra-grupo; por otro lado, el Coeficiente de Silueta evalúa la coherencia interna del agrupamiento. Un valor cercano a 0 indica grupos que se solapan.

Una vez etiquetados nuestros datos con el algoritmo de agrupamiento K-means, el siguiente paso es validar nuestros resultados con algoritmos de aprendizaje supervisados. Para este caso se seleccionan cuatro modelos: el Perceptrón Multicapa (Multilayer Perceptron, MLP), las Máquinas de Soporte Vectorial (Support Vector Machines, SVM), la Regresión Logística (Logistic Regression, LR) y el Random Forest (FR). El desempeño de cada algoritmo se puede observar en la Tabla 6.

Como se observa en la Tabla 6, el algoritmo que obtuvo mejores resultados fue el Perceptrón Multicapa, seguido de la Máquina de Soporte Vectorial, y la Regresión Logística.

Tabla 6. Resultados de algoritmos de aprendizaje supervisado.

Algoritmo	Accuracy	Precision (macro)	Recall (marco)	F1-score (macro)	Precision (weighted)	Recall (weighted)	F1-Score (weighted)
LR	0.9535	0.9587	0.9543	0.9563	0.9543	0.9535	0.9536
FR	0.9302	0.9380	0.9310	0.9290	0.9369	0.9302	0.9280
SVM	0.9535	0.9658	0.9467	0.9542	0.9583	0.9535	0.9538
MLP	0.9535	0.9576	0.9599	0.9582	0.9550	0.9535	0.9536

Interpretación de resultados y discusión.

El análisis de los resultados obtenidos mediante algoritmos de aprendizaje no supervisado, específicamente K-means en conjunto con PCA, el método del codo y el coeficiente de silueta promedio, indicó que la agrupación óptima de los datos académicos corresponde a tres grupos. Dos de estos grupos están conformados por estudiantes con indicadores positivos de madurez académica, rendimiento escolar sobresaliente, alto nivel de apoyo familiar, y percepción institucional favorable; no obstante, el grupo 1 revela patrones asociados a una posible desvinculación emocional o académica, lo que sugiere la necesidad de atención focalizada mediante programas de orientación psicológica, mentoría personalizada y acompañamiento institucional.

Para caracterizar los grupos identificados, se realizó un análisis cualitativo complementado con el cálculo de la moda por variable dentro de cada clúster, lo que permitió interpretar las variables dominantes que describen los perfiles estudiantiles.

Posteriormente, se entrenaron modelos de aprendizaje supervisado con el objetivo de predecir la pertenencia de estudiantes nuevos (no incluidos en los conjuntos de entrenamiento ni de prueba) a alguno de los grupos previamente identificados. Los resultados obtenidos demuestran que es factible realizar predicciones confiables sobre el grupo de pertenencia de nuevos estudiantes con características similares, permitiendo así la detección temprana de posibles casos de riesgo académico o emocional, y facilitando la toma de decisiones para la intervención oportuna por parte de las instituciones educativas.

CONCLUSIONES.

Este estudio presenta una contribución significativa al campo de la Minería de Datos Educativa (EDM) mediante la aplicación de un enfoque metodológico estructurado en cuatro etapas: recolección de datos, preprocesamiento, selección de modelos, e interpretación de resultados.

El caso de estudio presentado se centra en una Institución de Educación Superior perteneciente al Tecnológico Nacional de México, considerando una muestra de 430 estudiantes de distintos niveles académicos. Los datos fueron recolectados a través de un cuestionario validado mediante los coeficientes de V de Aiken y Alfa de Cronbach, garantizando así la fiabilidad y consistencia de las variables consideradas.

Para la reducción de dimensionalidad se empleó el algoritmo PCA (Análisis de Componentes Principales), mientras que el agrupamiento de estudiantes se llevó a cabo utilizando el algoritmo K-means, seleccionando automáticamente el número óptimo de componentes y de grupos mediante el método del codo y el coeficiente de silueta promedio. El análisis permitió identificar un grupo de estudiantes con indicios de desvinculación emocional y/o académica, lo que sugiere la necesidad de implementar estrategias de intervención como programas de orientación psicológica, mentoría personalizada y asesoramiento vocacional. Posteriormente, se entrenaron y compararon cuatro modelos de aprendizaje supervisado: el Perceptrón Multicapa (Multilayer Perceptron, MLP), las Máquinas de Soporte Vectorial (Support Vector Machines, SVM), la Regresión Logística (Logistic Regression, LR) y el Random Forest

(FR), destacando la MLP, SVM y LR por su mayor desempeño. La validación de los modelos se realizó mediante matrices de confusión y métricas estándar como exactitud, precisión, sensibilidad (recall) y F1-score.

Como trabajo futuro, se propone ampliar el tamaño de la muestra, incluir diversas instituciones del mismo subsistema educativo y evaluar el desempeño de modelos emergentes de aprendizaje automático, con el fin de robustecer la generalización de los hallazgos y avanzar hacia sistemas de alerta temprana para la prevención del abandono escolar.

REFERENCIAS BIBLIOGRÁFICAS.

1. Alkandari, A., & Aljaber, S. J. (2015). Principle Component Analysis algorithm (PCA) for image recognition. 2015 Second International Conference on Computing Technology and Information Management (ICCTIM), 76-80. https://doi.org/10.1109/ICCTIM.2015.7224596
2. Colpo, M. P., Primo, T. T., Aguiar, M. S. de, & Cechinel, C. (2024). Educational Data Mining for Dropout Prediction: Trends, Opportunities, and Challenges. Revista Brasileira de Informática Na Educação, 32, 220-256. https://doi.org/10.5753/rbie.2024.3559
3. Khan, A., & Ghosh, S. K. (2021). Student performance analysis and prediction in classroom learning: A review of educational data mining studies. Education and Information Technologies, 26(1), 205-240. https://doi.org/10.1007/s10639-020-10230-3
4. Khan, M., Naz, S., Khan, Y., Zafar, M., Khan, M., & Pau, G. (2023). Utilizing Machine Learning Models to Predict Student Performance From LMS Activity Logs. IEEE Access, 11, 86953-86962. https://doi.org/10.1109/ACCESS.2023.3305276
5. Kosaraju, N., Sankepally, S. R., & Mallikharjuna Rao, K. (2023). Categorical Data: Need, Encoding, Selection of Encoding Method and Its Emergence in Machine Learning Models—A Practical Review Study on Heart Disease Prediction Dataset Using Pearson Correlation. En M. Saraswat, C. Chowdhury, C. Kumar Mandal, & A. H. Gandomi (Eds.), Proceedings of International Conference on

Data Science and Applications (pp. 369-382). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-6631-6_26

6. Mamat, A. R., Mohamed, F. S., Mohamed, M. A., Rawi, N. M., & Awang, M. I. (2018). Silhouette index for determining optimal k-means clustering on images in different color models. International Journal of Engineering and Technology, 7(2.14), Article 2.14. https://doi.org/10.14419/ijet.v7i2.14.11464
7. Meneses Claudio, B. (2024). Application of Data Mining for the Prediction of Academic Performance in University Engineering Students at the National Autonomous University of Mexico, 2022. LatIA, 2, 4.
8. Mengash, H. A. (2020). Using Data Mining Techniques to Predict Student Performance to Support Decision Making in University Admission Systems. IEEE Access, 8, 55462-55470. https://doi.org/10.1109/ACCESS.2020.2981905
9. Merino-Soto, C. (2023). Aiken's V Coefficient: Differences in Content Validity Judgments. MHSalud: Revista en Ciencias del Movimiento Humano y Salud, 20(1), 1-10. https://doi.org/10.15359/mhs.20-1.3
10. Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. EURASIP Journal on Wireless Communications and Networking, 2021(1), 31. https://doi.org/10.1186/s13638-021-01910-w
11. Sowmiya Narayanan, K. J., & Manimaran, A. (2024). Using Decision Risk and Decision Accuracy Metrics for Decision Making for Remote Sensing and GIS Applications. En K. R. Reddy, P. T. Ravichandran, R. Ayothiraman, & A. Joseph (Eds.), Recent Advances in Civil Engineering (pp. 125-136). Springer Nature Singapore. https://doi.org/10.1007/978-981-99-6229-7_11
12. Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. International Journal of Medical Education, 2, 53-55. https://doi.org/10.5116/ijme.4dfb.8dfd

13. Trung, B. D., Son, N. T., Tung, N. D., Son, K. A., Anh, B. N., & Lam, P. T. (2023). Educational Data Mining: A Systematic Review on the Applications of Classical Methods and Deep Learning Until 2022. 2023 IEEE Symposium on Industrial Electronics & Applications (ISIEA), 1-15. https://doi.org/10.1109/ISIEA58478.2023.10212273

14. Xiao, W., Ji, P., & Hu, J. (2022). A survey on educational data mining methods used for predicting students' performance. Engineering Reports, 4(5), e12482. https://doi.org/10.1002/eng2.12482

DATOS DEL AUTOR.

1. **Vitervo López Caballero.** Doctor en Ciencias Computaciones con la especialidad de Ingeniería de Software por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Actualmente se desempeña como profesor investigador de tiempo completo en el CENIDET. México. Correo electrónico: vitervo.lc@cenidet.tecnm.mx ORCID: https://orcid.org/0000-0002-1942-9558

2. **Lucia Morales Morales.** Doctora en Sistemas computacionales por la Universidad del Sur (UNISUR) y Maestra en Ciencias de la Computación por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET), México. Actualmente se desempeña como consultora de desarrollo de software en la empresa Hospisoft, Chihuahua, México. Correo electrónico: cialu5040@gmail.com, ORCID: https://orcid.org/0009-0006-6593-4762

3. **Xochitl Morales Morales.** Doctora en Ciencias en Ingeniería Mecánica con la especialidad en Sistemas Térmicos por el Centro Nacional de Investigación y Desarrollo Tecnológico (CENIDET). Actualmente se desempeña como profesora investigadora del Instituto Tecnológico Superior de Tantoyuca. Correo electrónico: morales.m.x@hotmail.com

RECIBIDO: 25 de octubre del 2025. **APROBADO:** 30 de noviembre del 2025.