https://doi.org/10.69639/arandu.v13i3.2567

Analítica Inteligente de historiales académicos para la toma de decisiones en Educación Superior

Intelligent Analytics of Academic Records for Decision-Making in Higher Education

Renata Aguilar Rodríguez
renata.aguilar.uaem@gmail.com
https://orcid.org/0000-0003-4538-3378
Universidad Autónoma del Estado de México / Tecnológico Nacional de México
México – CDMX

Marco Polo Mendoza Otero
marco.mendoza@tecnm.mx
https://orcid.org/0009-0004-0508-5311
Tecnológico Nacional de México
México – CDMX

Magally Martínez Reyes
mmartinezr@uaemex.mx
https://orcid.org/0000-0002-2643-6748
Universidad Autónoma del Estado de México
México – CDMX

Anabelem Soberanes Martin
asoberanesm@uaemex.mx
https://orcid.org/0000-0002-1101-8279
Universidad Autónoma del Estado de México
México – CDMX

Adriana Bustamante Almaraz
abustamantea@uaemex.mx
https://orcid.org/0000-0002-5476-4140
Universidad Autónoma del Estado de México
México – CDMX

Artículo recibido: 10 julio 2026- Aceptado para publicación:16 agosto 2026
Conflictos de intereses: Ninguno que declarar.

RESUMEN

Las instituciones de educación superior generan grandes volúmenes de datos académicos que permiten analizar factores asociados a la reprobación y el abandono escolar. No obstante, muchas instituciones emplean estos registros para elaborar indicadores descriptivos de situaciones ya ocurridas y aprovechan limitadamente su capacidad para anticipar escenarios de riesgo. Este estudio tiene como objetivo desarrollar un modelo de aprendizaje automático que reconozca patrones recurrentes en historiales académicos y estimar probabilidad de reprobación o abandono. La investigación adopta un enfoque cuantitativo y un alcance predictivo mediante las etapas de comprensión, preparación, modelado y evaluación de los datos. El análisis considera variables como promedio, asignaturas reprobadas, repetición de cursos, carga académica, regularidad, avance curricular y periodos de inactividad. El estudio compara regresión logística, árboles de decisión, bosques aleatorios y máquinas de soporte vectorial. Para valorar su desempeño, utiliza

exactitud, precisión, sensibilidad, especificidad, puntuación F1, matriz de confusión y área bajo la curva característica operativa del receptor. También analiza la importancia de variables para determinar su relación con los resultados académicos desfavorables. El estudio busca identificar combinaciones y secuencias de eventos que el análisis descriptivo tradicional no muestra, además de seleccionar el modelo que ofrezca el mejor equilibrio entre la detección de estudiantes en riesgo y la reducción de clasificaciones incorrectas. Los resultados permitirán transformar los historiales académicos en evidencia interpretable para apoyar la toma de decisiones institucionales mediante enfoques de analítica educativa e inteligencia artificial, favoreciendo intervenciones oportunas y estrategias de acompañamiento académico sin sustituir el juicio profesional de docentes y tutores.

Palabras clave: analítica educativa, aprendizaje automático, historiales académicos, reconocimiento de patrones, toma de decisiones

ABSTRACT

Higher education institutions generate vast amounts of academic data that allow for the analysis of factors associated with course failure and student dropout. However, many institutions use these records primarily to create descriptive indicators of past events, making limited use of their potential to anticipate risk scenarios. This study aims to develop a machine learning model capable of recognizing recurring patterns in academic records and estimating the probability of course failure or dropout. The research adopts a quantitative approach and a predictive scope, proceeding through stages of data understanding, preparation, modeling, and evaluation. The analysis considers variables such as grade point average, failed subjects, repeated courses, academic workload, enrollment status, curricular progress, and periods of inactivity. The study compares logistic regression, decision trees, random forests, and support vector machines. Performance is assessed using accuracy, precision, sensitivity, specificity, F1-score, confusion matrix, and the area under the receiver operating characteristic (ROC) curve. It also analyzes variable importance to determine their relationship with unfavorable academic outcomes. The study seeks to identify combinations and sequences of events that traditional descriptive analysis fails to reveal, while selecting the model that offers the best balance between detecting at-risk students and minimizing misclassifications. The results will enable the transformation of academic records into interpretable evidence to support institutional decision-making through educational analytics and artificial intelligence, facilitating timely interventions and academic support strategies without replacing the professional judgment of instructors and academic advisors.

Keywords: learning analytics, machine learning, academic records, pattern recognition, decision-making

INTRODUCCIÓN

Las instituciones de educación superior reciben una población estudiantil diversa en cuanto a conocimientos previos, condiciones socioeconómicas, ritmos de aprendizaje, responsabilidades personales y experiencias escolares. En este contexto, garantizar el acceso a la educación constituye solo una parte del compromiso institucional, ya que también es necesario favorecer la permanencia, el progreso académico y la conclusión satisfactoria de los estudios. Sin embargo, fenómenos como la reprobación recurrente, el rezago y el abandono continúan afectando el desempeño de los estudiantes y la eficiencia de las instituciones educativas (Papadogiannis et al., 2024; Vaarma & Li, 2024).

Estos fenómenos poseen un carácter multifactorial. El promedio, la cantidad de asignaturas reprobadas, la repetición de cursos, los créditos acumulados, la carga académica, la regularidad, el avance curricular y los periodos de inactividad pueden relacionarse de diversas maneras con los resultados académicos. Su análisis individual permite describir determinadas situaciones, pero no siempre facilita reconocer las combinaciones o secuencias de eventos que preceden a la reprobación o al abandono. En consecuencia, parte de las decisiones institucionales continúa sustentándose en indicadores descriptivos que reflejan situaciones ya ocurridas, lo que limita la posibilidad de implementar intervenciones preventivas (Papadogiannis et al., 2024).

Paralelamente, los sistemas de control escolar, las plataformas educativas y los procesos de evaluación generan grandes volúmenes de datos sobre el comportamiento académico. No obstante, la disponibilidad de registros no garantiza por sí misma la generación de conocimiento útil. Para aprovecharlos, se requieren métodos capaces de integrar múltiples variables, reconocer relaciones no evidentes y convertir los resultados en información comprensible para docentes, tutores y responsables académicos (Papadogiannis et al., 2024).

La minería de datos educativos y la analítica del aprendizaje ofrecen marcos para responder a esta necesidad. La minería de datos educativos se centra en el desarrollo y la aplicación de métodos computacionales para descubrir patrones en conjuntos de datos generados en entornos educativos. Por su parte, la analítica del aprendizaje incorpora una orientación pedagógica e institucional, ya que utiliza los resultados para comprender y optimizar los procesos de enseñanza y aprendizaje. Ambas perspectivas permiten transformar datos académicos en información que puede emplearse para identificar estudiantes en riesgo, estimar posibles resultados y orientar intervenciones (Papadogiannis et al., 2024).

Por otro lado, la analítica del aprendizaje ha sido ampliamente reconocida como un enfoque orientado a comprender y optimizar los procesos educativos mediante la recopilación, análisis e interpretación de datos académicos. Su aplicación permite identificar oportunamente estudiantes en riesgo, diseñar estrategias de intervención basadas en evidencia y apoyar la toma de decisiones en instituciones de educación superior (Ferguson, 2012)

Así mismo la inteligencia artificial, particularmente el aprendizaje automático, amplía las posibilidades de este análisis al permitir que los sistemas aprendan patrones a partir de datos históricos y clasifiquen distintos resultados académicos. Entre los algoritmos utilizados para predecir el rendimiento estudiantil se encuentran la regresión logística, los árboles de decisión, los bosques aleatorios, las máquinas de soporte vectorial, el método de los vecinos más cercanos, Naive Bayes y las redes neuronales. Las revisiones existentes muestran que no existe un algoritmo universalmente superior, ya que su desempeño depende de las características del conjunto de datos, las variables seleccionadas, la distribución de las clases y el objetivo de predicción (Yadav & Deshmukh, 2023; Papadogiannis et al., 2024).

En los últimos años, la inteligencia artificial aplicada a la educación ha evolucionado hacia el desarrollo de sistemas capaces de apoyar procesos de orientación académica, personalización del aprendizaje y toma de decisiones institucionales. Estos enfoques permiten integrar técnicas predictivas con mecanismos de apoyo educativo orientados a mejorar la permanencia y el éxito académico de los estudiantes (Zawacki-Richter et al., 2019).

La evidencia reciente demuestra la utilidad de los historiales académicos para identificar patrones de permanencia y de abandono. Polat & Horzum (2026) analizaron 484,158 registros estudiantiles mediante árboles de decisión (J48), vecinos más cercanos, Naive Bayes y bosques aleatorios. El modelo J48 alcanzó una exactitud de 80.47 % en la clasificación de estudiantes graduados, con graduación tardía y en abandono. Asimismo, los créditos acumulados y el tipo de ingreso presentaron una mayor importancia predictiva que algunas variables demográficas. Estos resultados muestran que los indicadores de progreso académico contienen información relevante para anticipar resultados desfavorables.

No obstante, el mismo estudio encontró dificultades para identificar a los estudiantes con avance lento o graduación tardía, categoría que presentó una sensibilidad del 45.8 %. Esta limitación se atribuyó a la heterogeneidad del grupo y al carácter estático de los datos administrativos. Por ello, los autores recomiendan integrar distintas fuentes de información, examinar subgrupos estudiantiles y utilizar métodos capaces de distinguir niveles de riesgo con mayor precisión. Este hallazgo resulta relevante porque un modelo puede alcanzar una exactitud global aceptable y, al mismo tiempo, presentar un desempeño insuficiente en la categoría que requiere mayor atención.

En consecuencia, la evaluación de modelos predictivos no debe limitarse a la exactitud (*accuracy)*. Cuando las categorías presentan una distribución desigual, un modelo puede favorecer la clase mayoritaria y clasificar incorrectamente a los estudiantes en riesgo. Para obtener una evaluación más completa, es necesario considerar la precisión, la sensibilidad (*recall)*, la especificidad, el valor F1, la exactitud balanceada, la matriz de confusión y el área bajo la curva ROC. Estas métricas permiten examinar el equilibrio entre la detección de casos de riesgo y la generación de falsas alertas (Papadogiannis et al., 2024; Polat & Horzum, 2026).

Además del desempeño predictivo, es necesario considerar la interpretación de los resultados. Los modelos aplicados en contextos educativos deben permitir identificar qué variables contribuyen a cada clasificación (Breiman et al., 1984), ya que una predicción sin explicación presenta una utilidad limitada para la toma de decisiones. Los árboles de decisión ofrecen reglas comprensibles, mientras que los modelos de mayor complejidad pueden complementarse con procedimientos para calcular la importancia de las variables (Breiman, 2001). Esta interpretación permite que docentes y responsables académicos comprendan los factores relacionados con el riesgo y seleccionen acciones de acompañamiento pertinentes (Papadogiannis et al., 2024).

La analítica tampoco debe reducir la experiencia educativa a un conjunto de indicadores. Guzmán-Valenzuela et al. (2021) advierten que una parte de las investigaciones en este campo se ha concentrado más en la capacidad técnica de la analítica que en la comprensión del aprendizaje. De manera semejante, Civit et al. (2024) sostienen que los datos deben utilizarse para ampliar la información disponible y apoyar la toma de decisiones, sin sustituir la valoración de docentes, tutores y responsables institucionales. Por lo tanto, los resultados de un modelo predictivo deben considerarse como elementos de apoyo y no como decisiones automáticas o determinantes sobre la situación de un estudiante.

Desde esta perspectiva, los sistemas inteligentes de apoyo a la decisión constituyen herramientas que no sustituyen el juicio humano, sino que proporcionan evidencia adicional para fortalecer los procesos de acompañamiento académico, la identificación temprana de riesgos y la planeación de intervenciones institucionales oportunas (Ifenthaler & Yau, 2020).

La incorporación institucional de estas tecnologías también depende de la percepción de quienes podrían utilizarlas. Un sistema técnicamente funcional puede encontrar resistencia si sus recomendaciones no se consideran útiles, comprensibles, confiables o justas. Por esta razón, la evaluación preliminar de un sistema inteligente educativo debe contemplar tanto su funcionamiento computacional como la credibilidad percibida por los estudiantes. Aspectos como la utilidad, la facilidad de uso, la explicabilidad, el respeto a las necesidades individuales y la percepción de equidad pueden influir en la disposición a aceptar recomendaciones generadas mediante inteligencia artificial (Davis, 1989; Granić & Marangunić, 2019; Civit et al., 2024).

En este contexto se plantea el Sistema Inteligente para el Desarrollo de Trayectorias Académicas Inclusivas (SIDTAI). La propuesta integra información del historial académico, variables del perfil formativo y restricciones institucionales para apoyar la identificación de situaciones de riesgo y la generación de recomendaciones académicas. Su diseño considera que la inteligencia artificial debe funcionar como una herramienta de acompañamiento y proporcionar información interpretable, sin sustituir la decisión del estudiante ni la valoración de los responsables académicos.

A partir de estos antecedentes, el presente estudio desarrolla una evaluación preliminar organizada en dos componentes complementarios. El primero analiza las respuestas de 54 estudiantes a un instrumento adaptado de 15 reactivos sobre utilidad, facilidad de uso, credibilidad, justicia, explicabilidad y valoración global de SIDTAI. El segundo corresponde a un experimento computacional controlado con perfiles académicos sintéticos, diseñado para comparar el comportamiento de diferentes algoritmos ante escenarios de reprobación y abandono. La separación entre ambos componentes permite distinguir la evidencia empírica obtenida directamente de los participantes de la evaluación técnica realizada en un entorno de simulación.

La contribución del trabajo consiste en articular la valoración preliminar de la credibilidad de un sistema inteligente con el análisis comparativo de modelos de clasificación mediante métricas complementarias. El experimento no pretende sustituir la validación posterior con historiales académicos institucionales, sino comprobar el procedimiento analítico, examinar el efecto del desbalance entre categorías y determinar qué métricas proporcionan la información más adecuada para seleccionar un modelo orientado a la detección de riesgo.

El objetivo de este estudio es evaluar preliminarmente un enfoque de analítica inteligente orientado al reconocimiento de patrones relacionados con la reprobación y el abandono escolar, así como al apoyo de la toma de decisiones en educación superior. Para ello, se analiza la credibilidad percibida de SIDTAI mediante un instrumento aplicado a estudiantes y se desarrolla un experimento computacional controlado para comparar el desempeño de modelos de clasificación supervisada, entre ellos regresión logística, árboles de decisión, bosques aleatorios, máquinas de soporte vectorial y redes neuronales. La evaluación considera exactitud, precisión, sensibilidad, especificidad, valor F1, exactitud balanceada, matriz de confusión y área bajo la curva ROC, además de la importancia de las variables para favorecer la interpretación de los resultados.

Como supuesto de trabajo, se plantea que los algoritmos presentarán diferencias en su capacidad para identificar los casos de riesgo y que el modelo con mayor exactitud global no necesariamente ofrecerá el mejor equilibrio entre precisión y sensibilidad. En consecuencia, la selección de un modelo para apoyar la toma de decisiones educativas deberá considerar su desempeño computacional, la interpretación de sus resultados, las características de los datos y las posibles consecuencias de una clasificación incorrecta.

Como hipótesis se parte de la combinación de variables relacionadas con el rendimiento académico, la repetición de asignaturas, los créditos acumulados, la regularidad y el avance curricular permitirá identificar patrones asociados con resultados académicos desfavorables. Asimismo, se plantea que los algoritmos evaluados presentarán diferencias significativas en su capacidad predictiva y que el modelo con mayor exactitud no necesariamente será el más adecuado para detectar estudiantes en riesgo, por lo que su selección deberá considerar el equilibrio entre precisión y sensibilidad.

MATERIALES Y MÉTODOS

Enfoque y diseño de la investigación

Se desarrolló una investigación aplicada con enfoque cuantitativo y alcance exploratorio y evaluativo. El diseño se organizó en dos componentes complementarios. El primero correspondió a un estudio no experimental y transversal, orientado a analizar la percepción estudiantil sobre la utilidad, facilidad de uso, credibilidad, justicia y explicabilidad del Sistema Inteligente para el Desarrollo de Trayectorias Académicas Inclusivas (SIDTAI). El segundo consistió en un experimento computacional controlado con perfiles académicos sintéticos, en el que se comparó el comportamiento de distintos algoritmos de clasificación ante escenarios de reprobación y abandono.

La organización del proceso analítico se basó en la metodología CRISP-DM, particularmente en sus fases de comprensión del problema, comprensión de los datos, preparación, modelado y evaluación. Este marco permitió documentar las transformaciones aplicadas a cada conjunto de datos y mantener una separación explícita entre la información proporcionada por participantes reales y los datos generados mediante simulación (Martínez et al., 2021; Schröer et al., 2021).

Contexto, población y participantes

La población de referencia estuvo conformada por 710 estudiantes de Ingeniería en Sistemas Computacionales del Instituto Tecnológico de Iztapalapa. La muestra efectiva estuvo integrada por 54 estudiantes, de los cuales 6 cursaban el primer semestre y 48 el segundo.

Se empleó un muestreo no probabilístico por autoselección, ya que participaron los estudiantes que decidieron responder el instrumento tras recibir la invitación electrónica. En consecuencia, los resultados se interpretaron como evidencia exploratoria sobre las expectativas y percepciones de los participantes, sin pretender generalizarlos estadísticamente a la totalidad del programa educativo. Dado a que no se contó con el número exacto de estudiantes que recibieron o visualizaron la invitación, no se calculó una tasa de respuesta.

Técnica e instrumento de recolección de datos

La técnica de recolección de datos fue la encuesta, aplicada mediante un cuestionario electrónico adaptado compuesto por 15 reactivos relacionados con la percepción de un sistema inteligente para el acompañamiento académico. Su construcción retomó elementos del Modelo de Aceptación Tecnológica, principios relacionados con la usabilidad, la credibilidad y la explicabilidad de los sistemas inteligentes, así como componentes de la percepción de justicia y la equidad en las decisiones algorítmicas. (Davis, 1989; Brooke, 1996; Colquitt, 2001; Granić & Marangunić, 2019; Hoffman et al., 2023). Los reactivos se organizaron según las dimensiones presentadas en la Tabla 1. El reactivo 15 se utilizó como valoración global del sistema.

Tabla 1

Dimensiones del instrumento de percepción de SIDTAI

Dimensión	Reactivos	Aspectos evaluados
Utilidad y adopción	1, 3, 4 y 13	Organización académica, disposición de uso, acceso a apoyos y utilidad inclusiva
Facilidad de uso	2, 5, 6, 7 y 14	Facilidad, confianza de uso, experiencia, comprensión y aprendizaje del sistema
Credibilidad	8	Confianza en los resultados o sugerencias
Justicia y explicabilidad	9, 10, 11 y 12	Trato justo, claridad de las decisiones, respeto a las necesidades y participación
Valoración global	15	Percepción conjunta de credibilidad, utilidad y justicia

Las opciones de respuesta fueron: No, No lo sé, Un poco, Sí y Totalmente de acuerdo. Para el análisis principal se asignaron los valores 1, 2, 3, 4 y 5, respectivamente. Debido a que la opción "No lo sé" no representa necesariamente un nivel de acuerdo situado entre "No" y "Un poco", se realizaron análisis de sensibilidad mediante codificaciones alternativas, entre ellas su tratamiento como punto medio y como respuesta ausente. Las respuestas originales se conservaron sin modificaciones.

Procedimiento de aplicación

El instrumento se aplicó el 7 de octubre de 2025 mediante un cuestionario electrónico. El enlace de acceso se distribuyó a través del correo institucional y de un chat grupal de WhatsApp utilizado para la comunicación con los estudiantes.

La invitación presentó a SIDTAI como una herramienta de apoyo docente y académico orientada a facilitar la planeación de trayectorias formativas semestre a semestre. El mensaje explicó que el sistema tendría como función sugerir asignaturas en función de las habilidades previas, las necesidades específicas y el ritmo de aprendizaje de cada estudiante. También señaló como beneficios esperados el fortalecimiento de los conocimientos, la prevención de riesgos académicos, la reducción de la reprobación y del abandono, así como la generación de una experiencia universitaria personalizada, equitativa e inclusiva. Posteriormente, se invitó a los estudiantes a ingresar al enlace y responder el instrumento.

La participación se produjo de manera autoseleccionada y la base utilizada para el análisis no incluyó nombres, números de control, correos electrónicos ni otros identificadores directos. Cada registro estuvo compuesto por el semestre declarado y las respuestas a los 15 reactivos.

Debido a que la invitación describió principalmente los beneficios esperados de SIDTAI, se consideró la posible presencia de un efecto de encuadre favorable. Este efecto se observa cuando la forma de comunicar o presentar una situación influye en las valoraciones o decisiones

de las personas (Tversky & Kahneman, 1981). Por esta razón, las respuestas no se interpretaron como evidencia de que el sistema reduzca efectivamente la reprobación o el abandono, sino como una valoración preliminar de su utilidad, facilidad de uso, credibilidad, justicia y explicabilidad antes de una implementación completa.

Preparación y análisis del instrumento

La base de datos se revisó para identificar registros incompletos, categorías de respuesta no reconocidas, patrones duplicados y respuestas invariantes. Cada fila representó a un participante y cada columna correspondió a uno de los 15 reactivos. No se eliminaron casos del análisis principal por presentar respuestas constantes, ya que el instrumento no incorporó reactivos de control que permitieran determinar si dichos patrones eran inválidos.

Se calcularon frecuencias, medias, desviaciones estándar y puntuaciones promedio por dimensión. La consistencia interna se examinó mediante el coeficiente alfa de Cronbach (Cronbach, 1951) tanto para el instrumento completo como para las dimensiones integradas por más de un reactivo. No se calculó este coeficiente para la dimensión de credibilidad, debido a que estuvo representada por un solo reactivo.

Aunque el alfa continúa siendo uno de los indicadores más utilizados para evaluar la consistencia interna, su interpretación debe considerar el número de reactivos, la dimensionalidad del instrumento y la distribución de las respuestas. Por ello, el coeficiente global se complementó con resultados por dimensión y con análisis de sensibilidad (Xiao & Hau, 2023; Kalkbrenner, 2024).

También se estimó un intervalo de confianza del 95 % para el coeficiente global mediante un procedimiento *bootstrap* de 5,000 remuestreos (Efron, 1979). Como análisis de sensibilidad, se repitieron los cálculos excluyendo temporalmente los patrones de respuesta invariantes. Esta comparación no implicó la eliminación definitiva de participantes, sino que permitió determinar si la consistencia interna dependía principalmente de quienes seleccionaron la misma categoría en todos los reactivos

Para describir la percepción de los estudiantes sobre SIDTAI, se calcularon frecuencias, porcentajes, medias y desviaciones estándar para las dimensiones de utilidad y adopción, facilidad de uso, credibilidad, justicia y explicabilidad, así como para la evaluación global del sistema. Adicionalmente, se establecieron tres niveles descriptivos a partir del promedio de los 15 reactivos: bajo para puntuaciones menores de 3; medio para puntuaciones iguales o superiores a 3 y menores de 4; y alto para puntuaciones iguales o superiores a 4. Estos puntos de corte se utilizaron únicamente para facilitar la descripción de los resultados y no constituyen umbrales diagnósticos validados.

Clasificación exploratoria de la credibilidad percibida

Se realizó un análisis exploratorio para determinar en qué medida las dimensiones del instrumento permitían clasificar la valoración global expresada en el reactivo 15. La variable de

salida se dicotomizó en credibilidad alta, integrada por las respuestas “Sí” y “Totalmente de acuerdo”, y credibilidad no alta, que agrupó las respuestas “No”, “No lo sé” y “Un poco”.

Como variables de entrada se utilizaron las puntuaciones de utilidad y adopción, de facilidad de uso, de credibilidad, de justicia y explicabilidad. Este procedimiento evaluó la coherencia entre las dimensiones y la valoración global del instrumento. Por consiguiente, sus resultados no se interpretaron como una predicción de reprobación, rezago o abandono, ni como evidencia de la eficacia académica de SIDTAI.

Se compararon cinco algoritmos de clasificación supervisada: regresión logística, árbol de decisión CART, bosque aleatorio, máquina de soporte vectorial con kernel de función de base radial y red neuronal de perceptrón multicapa (Breiman et al., 1984), bosque aleatorio (Breiman, 2001), máquina de soporte vectorial con kernel de función de base radial (Cortes & Vapnik, 1995) y red neuronal de perceptrón multicapa (Rumelhart et al., 1986). También se incorporó un clasificador basado en la categoría mayoritaria como referencia mínima de desempeño.

Generación de perfiles académicos sintéticos

Ante la falta de acceso a historiales académicos institucionales anonimizados, se construyó un conjunto controlado de 1,200 observaciones sintéticas de tipo estudiante-periodo. La simulación tuvo como propósito evaluar el procedimiento computacional, comparar métricas y verificar el comportamiento de los algoritmos ante categorías desbalanceadas. Los datos sintéticos no se utilizaron para estimar la prevalencia real de reprobación o de abandono en el Instituto Tecnológico de Iztapalapa.

La selección de variables se basó en investigaciones sobre minería de datos educativos y predicción del abandono, las cuales señalan la utilidad del rendimiento previo, la reprobación acumulada, el progreso curricular y la actividad académica para la identificación de situaciones de riesgo (Papadogiannis et al., 2024; Vaarma & Li, 2024). Las variables incorporadas se presentan en la Tabla 2.

Tabla 2

Variables del experimento computacional controlado

Variable	**Descripción**
Semestre	Periodo académico cursado, con valores entre 1 y 13
Promedio previo	Promedio académico anterior, limitado al intervalo de 50 a 100
Materias reprobadas	Número acumulado de asignaturas no acreditadas
Intentos máximos	Mayor número de intentos registrado en una asignatura
Carga académica	Cantidad de asignaturas cursadas durante el periodo
Dificultad percibida	Nivel de dificultad académica en una escala de 1 a 5
Periodos de inactividad	Número de interrupciones temporales simuladas

Proporción acreditada	Relación entre asignaturas acreditadas y cursadas
Avance curricular	Porcentaje estimado de progreso en el plan de estudios
Reprobación del siguiente periodo	Resultado binario generado probabilísticamente
Abandono del siguiente periodo	Resultado binario generado probabilísticamente

Los periodos académicos se generaron en un intervalo de 1 a 13 semestres, considerando que el plan regular se desarrolla en nueve periodos y que el tiempo máximo de permanencia puede extenderse hasta 13 semestres. Las demás variables se generaron mediante distribuciones acotadas y relaciones probabilísticas acordes con los escenarios académicos examinados.

La probabilidad de reprobación aumentó ante promedios bajos, mayor número de asignaturas reprobadas, repetición de intentos, cargas elevadas y mayor dificultad percibida. La probabilidad de abandono se incrementó ante promedios bajos, reprobación acumulada, periodos de inactividad, intentos repetidos y menor avance curricular.

Las etiquetas se obtuvieron mediante funciones logísticas con ruido aleatorio. De esta manera, las categorías no se determinaron mediante reglas rígidas ni copiadas directamente de las variables predictoras. No obstante, las relaciones y los coeficientes utilizados constituyen supuestos del experimento y no parámetros estimados a partir de la población institucional. Se empleó una semilla aleatoria fija para garantizar la reproducibilidad del conjunto de datos.

Configuración de los modelos

Los cinco algoritmos se aplicaron tanto al escenario de reprobación como al de abandono. La regresión logística utilizó una ponderación balanceada entre las categorías. El árbol CART empleó el criterio de impureza de Gini, una profundidad máxima de 3 niveles y un mínimo de 15 observaciones por hoja. El bosque aleatorio se configuró con 300 árboles, profundidad máxima de 6 niveles y ponderación balanceada (Breiman et al., 1984; Breiman, 2001).

La máquina de soporte vectorial utilizó un kernel de función de base radial, un parámetro de regularización igual a 1 y ponderación balanceada (Cortes & Vapnik, 1995). La red neuronal multicapa se integró con dos capas ocultas de ocho y cuatro neuronas, un optimizador de memoria limitada de Broyden-Fletcher-Goldfarb-Shanno y regularización para reducir el sobreajuste (Rumelhart et al., 1986). Las variables numéricas se estandarizaron antes del entrenamiento de la regresión logística, de la máquina de soporte vectorial y de la red neuronal.

La configuración de los modelos buscó contener su complejidad y reducir el riesgo de sobreajuste. No se realizó una búsqueda exhaustiva de hiperparámetros debido a que la finalidad del experimento consistió en comparar el comportamiento general de distintas familias de clasificadores y no en obtener un modelo definitivo para su despliegue institucional.

Evaluación del desempeño

La evaluación se realizó mediante validación cruzada estratificada en cinco particiones. Este procedimiento permitió conservar la proporción de las categorías durante la división de los

datos y estimar el desempeño esperado de los algoritmos sobre observaciones distintas de aquellas empleadas durante su entrenamiento. La interpretación de los resultados consideró la variabilidad producida por las particiones y al tamaño de la muestra (Bates et al., 2024).

Para la clasificación exploratoria de la credibilidad, se repitió el procedimiento 20 veces, lo que dio lugar a 100 evaluaciones por algoritmo. Para los escenarios sintéticos de reprobación y abandono, se realizaron cinco repeticiones, lo que equivale a 25 evaluaciones por algoritmo. Se calcularon la media y la desviación estándar de las métricas obtenidas.

Las métricas consideradas fueron exactitud, exactitud balanceada, precisión, sensibilidad, especificidad, valor F1, valor F1 macro y área bajo la curva ROC. La exactitud representó la proporción total de clasificaciones correctas; la exactitud balanceada otorgó el mismo peso al desempeño en ambas categorías; la precisión indicó la proporción de alertas positivas que correspondieron a casos positivos; la sensibilidad midió la capacidad para detectar casos de riesgo; la especificidad evaluó la capacidad de identificar casos sin riesgo; y el valor F1 expresó la media armónica entre precisión y sensibilidad.

La utilización de métricas complementarias resultó necesaria porque la exactitud y el valor F1 pueden ofrecer interpretaciones incompletas cuando las categorías se encuentran desbalanceadas. En estas condiciones, un modelo puede obtener una exactitud elevada al favorecer la categoría mayoritaria sin detectar adecuadamente los casos positivos (Chicco & Jurman, 2020).

También se generaron matrices de confusión mediante predicciones fuera de muestra obtenidas con validación cruzada estratificada. La comparación con el clasificador mayoritario permitió identificar escenarios en los que una exactitud elevada no correspondía con una detección adecuada de los casos de riesgo.

Finalmente, se calculó la importancia de las variables mediante permutación. Para este procedimiento se reservó el 25 % de las observaciones sintéticas como conjunto de prueba y se midió la disminución de la exactitud balanceada después de modificar aleatoriamente cada variable durante 30 repeticiones. Una disminución mayor indicó que la variable aportaba más información al proceso de clasificación.

El procesamiento se efectuó con Python 3.12.13 y las bibliotecas pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0, SciPy 1.17.0, Matplotlib 3.10.8 y Seaborn 0.13.2. La implementación principal de los algoritmos se realizó con scikit-learn (Pedregosa et al., 2011). Se utilizó la misma semilla aleatoria en las etapas de simulación, partición y entrenamiento para favorecer la reproducibilidad de los resultados.

Consideraciones éticas

La base del instrumento no contenía nombres, números de control, correos electrónicos ni otros identificadores directos. La información se utilizó exclusivamente con fines académicos y los resultados se analizaron de forma agregada. Asimismo, se mantuvo una separación explícita entre las respuestas auténticas de los 54 participantes y los perfiles académicos sintéticos.

Los resultados de la simulación no se atribuyeron a estudiantes reales ni se utilizaron para tomar decisiones académicas individuales. El enfoque propuesto concibe los modelos como herramientas de apoyo para estudiantes, docentes, tutores y responsables institucionales, y no como mecanismos autónomos para determinar la situación académica de una persona.

La interpretación de los resultados consideró que los participantes respondieron tras recibir una descripción centrada en los beneficios esperados de SIDTAI. Este encuadre pudo generar expectativas favorables, aquiescencia o deseabilidad social. Por ello, los resultados representan percepciones preliminares del concepto del sistema y no una evaluación de su funcionamiento tras una experiencia prolongada de uso.

RESULTADOS

Características de las respuestas y percepción de SIDTAI

Se analizaron 54 respuestas válidas de estudiantes de Ingeniería en Sistemas Computacionales. Seis participantes cursaban el primer semestre (11.1%) y 48 el segundo semestre (88.9%). Debido a que la participación fue voluntaria y se utilizó un muestreo no probabilístico, los resultados describen exclusivamente las percepciones del grupo participante y no deben generalizarse a la totalidad de la población estudiantil.

La Tabla 3 presenta los resultados descriptivos del instrumento adaptado. En una escala de 1 a 5, el promedio global fue de 3.663 (DE = 0.883). La dimensión con mayor puntuación fue la utilidad y adopción (M = 3.722), seguida de la facilidad de uso (M = 3.700). La evaluación global del sistema, medida mediante el reactivo 15, obtuvo una media de 3.778. La justicia y la explicabilidad alcanzaron una media de 3.560, mientras que el reactivo individual de credibilidad presentó una media de 3.537.

Tabla 3

Resultados descriptivos y consistencia interna del instrumento adaptado

Dimensión	Reactivos	Media	Desviación Estándar	Alfa de Cronbach
Utilidad y adopción	4	3.722	0.874	0.883
Facilidad de Uso	5	3.700	0.872	0.913
Credibilidad	1	3.537	1.094	-
Justicia y explicabilidad	4	3.560	1.013	0.921
Evaluación global	1	3.778	1.003	-
Instrumento completo	15	3.663	0.883	0.971

Nota. Escala de respuesta de 1 = No a 5 = Totalmente de acuerdo. No se calculó alfa para las dimensiones compuestas por un solo reactivo. Fuente: elaboración propia.

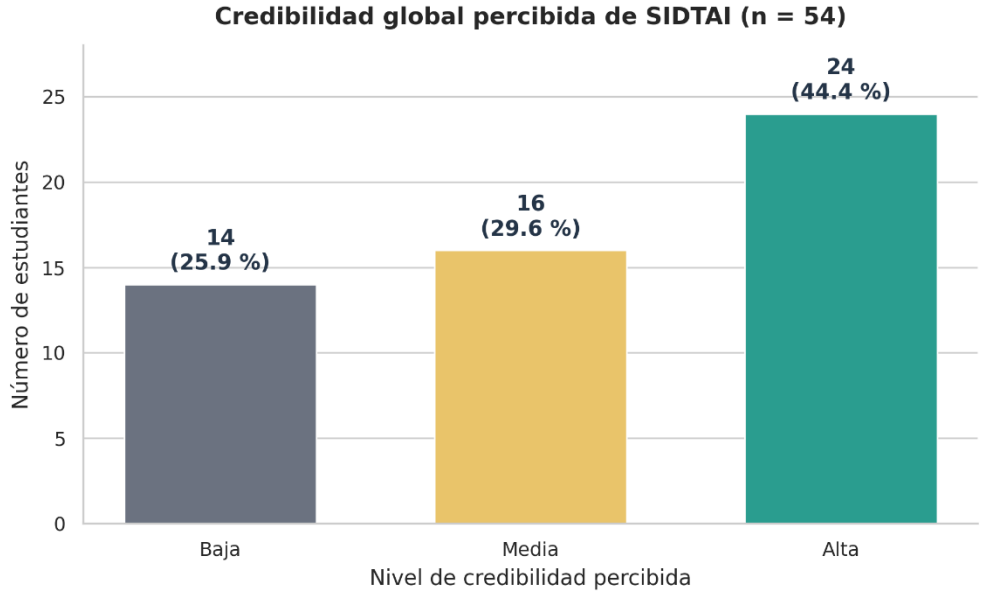
El instrumento completo obtuvo un alfa de Cronbach de 0.971, con un intervalo de confianza bootstrap del 95 % entre 0.953 y 0.982. Al excluir el reactivo de evaluación global, el coeficiente permaneció elevado (α = 0.967). No obstante, se identificaron 12 patrones constantes de respuesta (22.2%), 32 participantes que utilizaron únicamente una o dos categorías de respuesta (59.3%) y 10 vectores duplicados (18.5%).

Para analizar la estabilidad del coeficiente, se realizaron pruebas de sensibilidad. Al excluir los 12 patrones constantes, el alfa global fue de 0.953. Cuando la respuesta "No lo sé" se codificó como punto medio, el coeficiente fue de 0.962; al tratarla como valor ausente e imputarla con la mediana, se obtuvo un alfa de 0.957. Por consiguiente, la elevada consistencia interna no dependió exclusivamente de la codificación utilizada, aunque deben examinarse la posible redundancia entre reactivos y la tendencia de algunos participantes a responder de forma uniforme.

La clasificación del promedio total mostró 24 estudiantes con credibilidad alta (44.4 %), 16 con credibilidad media (29.6 %) y 14 con credibilidad baja (25.9 %), como se muestra en la Figura 1. En conjunto, los resultados indican una percepción predominantemente favorable hacia SIDTAI, aunque más de la mitad de los participantes se ubicó en los niveles medio o bajo.

Figura 1

Distribución del nivel de credibilidad global percibida de SIDTAI



Nota. Resultados correspondientes a las 54 respuestas válidas. Fuente: elaboración propia.

Clasificación exploratoria de la credibilidad

El reactivo 15 se dicotomizó para distinguir entre credibilidad global alta —respuestas "Sí" y "Totalmente de acuerdo"— y credibilidad no alta. Bajo este criterio, se identificaron 34 respuestas de credibilidad alta y 20 de credibilidad no alta. Se compararon cinco algoritmos de clasificación con una base mayoritaria mediante validación cruzada estratificada de cinco particiones repetida 20 veces.

Tabla 4

Desempeño de los modelos para la clasificación exploratoria de credibilidad

Modelo	Exactitud	Exactitud Balanceada	F1 macro	Área bajo la curva ROC
Base mayoritaria	0.629	0.500	0.386	0.500
Regresión logística	0.903	0.908	0.897	0.964
Árbol de clasificación y regresión	0.885	0.900	0.880	0.929
Bosque aleatorio	0.884	0.889	0.878	0.954
Máquina de soporte vectorial	0.907	0.918	0.904	0.956
Red neuronal multicapa	0.881	0.877	0.870	0.956

Nota. Los valores corresponden al promedio de la validación cruzada estratificada de cinco particiones repetida 20 veces. Fuente: elaboración propia.

La máquina de soporte vectorial obtuvo la mayor exactitud (0.907), exactitud balanceada (0.918) y F1 macro (0.904). La regresión logística presentó el área bajo la curva ROC más elevada (0.964). Al excluir los patrones constantes, la exactitud de la regresión logística disminuyó a 0.867, mientras que el árbol de clasificación alcanzó 0.823 y la red neuronal multicapa 0.847. Esta disminución confirma la influencia de los patrones uniformes, aunque los resultados permanecieron por encima de la base mayoritaria.

Estos resultados representan la coherencia interna entre las dimensiones del instrumento y la evaluación global expresada por los mismos participantes. Por tanto, no constituyen evidencia de que SIDTAI pueda predecir reprobación o abandono, ni demuestran todavía su efectividad posterior a la implementación.

Experimento controlado con perfiles académicos sintéticos

El segundo componente consistió en un experimento computacional con 1,200 observaciones sintéticas de tipo estudiante-periodo. Las relaciones entre las variables y las clases se generaron probabilísticamente, incluyendo incertidumbre aleatoria para evitar reglas de clasificación deterministas. La prevalencia de reprobación en el siguiente periodo fue de 31.5 %, mientras que la prevalencia de abandono fue de 9.3 %.

Tabla 5

Desempeño de los modelos en el experimento controlado con datos sintéticos

Resultado	Modelo	Exactitud	Exactitud balanceada	Sensibilidad	F1 de riesgo	Área bajo la curva ROC
Reprobación	Base mayoritaria	0.685	0.500	0.000	0.000	0.500
Reprobación	Regresión logística	0.719	0.697	0.636	0.587	0.752
Reprobación	Árbol de clasificación y regresión	0.687	0.655	0.567	0.532	0.707
Reprobación	Bosque aleatorio	0.716	0.683	0.593	0.568	0.734
Reprobación	Máquina de soporte vectorial	0.712	0.682	0.601	0.567	0.728
Reprobación	Red neuronal multicapa	0.719	0.633	0.403	0.473	0.694
Abandono	Base mayoritaria	0.907	0.500	0.000	0.000	0.500
Abandono	Regresión logística	0.645	0.592	0.528	0.217	0.636
Abandono	Árbol de clasificación y regresión	0.612	0.559	0.495	0.188	0.578
Abandono	Bosque aleatorio	0.836	0.550	0.198	0.181	0.615
Abandono	Máquina de soporte vectorial	0.669	0.521	0.338	0.160	0.555
Abandono	Red neuronal multicapa	0.888	0.508	0.041	0.065	0539

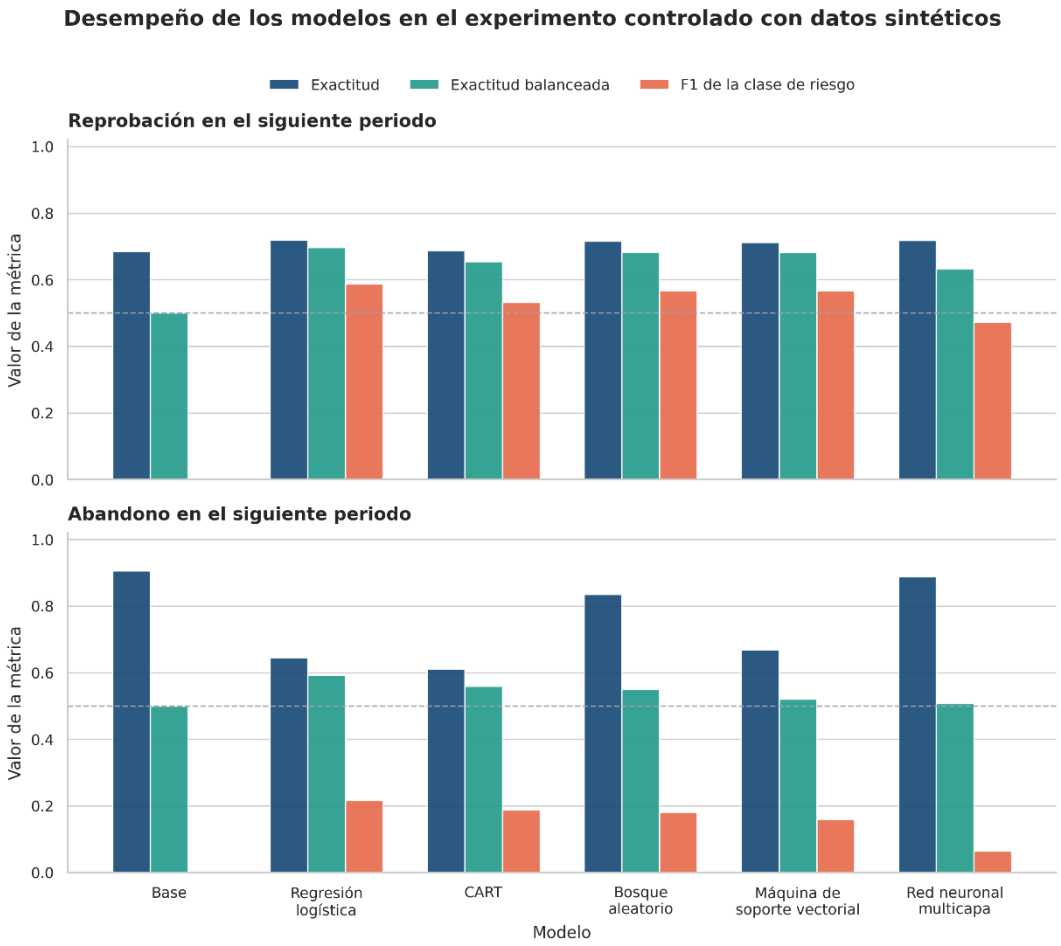
Nota. Resultados promedio de la validación cruzada estratificada de cinco particiones repetida cinco veces. La sensibilidad y la F1 corresponden a la clase positiva de riesgo. Estos resultados provienen de datos sintéticos y no representan historiales institucionales reales. Fuente: elaboración propia.

En la clasificación de reprobación, la regresión logística presentó el mejor equilibrio general, con una exactitud balanceada de 0.697, una sensibilidad de 0.636, un F1 de 0.587 y un área bajo la curva ROC de 0.752. El bosque aleatorio y la máquina de soporte vectorial mostraron resultados cercanos. Aunque la red neuronal multicapa alcanzó una exactitud de 0.719, su sensibilidad fue de 0.403, por lo que omitió una proporción mayor de los casos de riesgo.

En el abandono se observó con mayor claridad el efecto del desbalance de clases. La base mayoritaria alcanzó una exactitud de 0.907, pero obtuvo sensibilidad y F1 iguales a cero, ya a que clasificó todas las observaciones como permanencia. Por el contrario, la regresión logística redujo su exactitud a 0.645, pero alcanzó la mayor sensibilidad (0.528), F1 de riesgo (0.217), exactitud balanceada (0.592) y el área bajo la curva ROC (0.636) entre los modelos evaluados.

Figura 2

Comparación del desempeño de los modelos en el experimento controlado con datos sintéticos

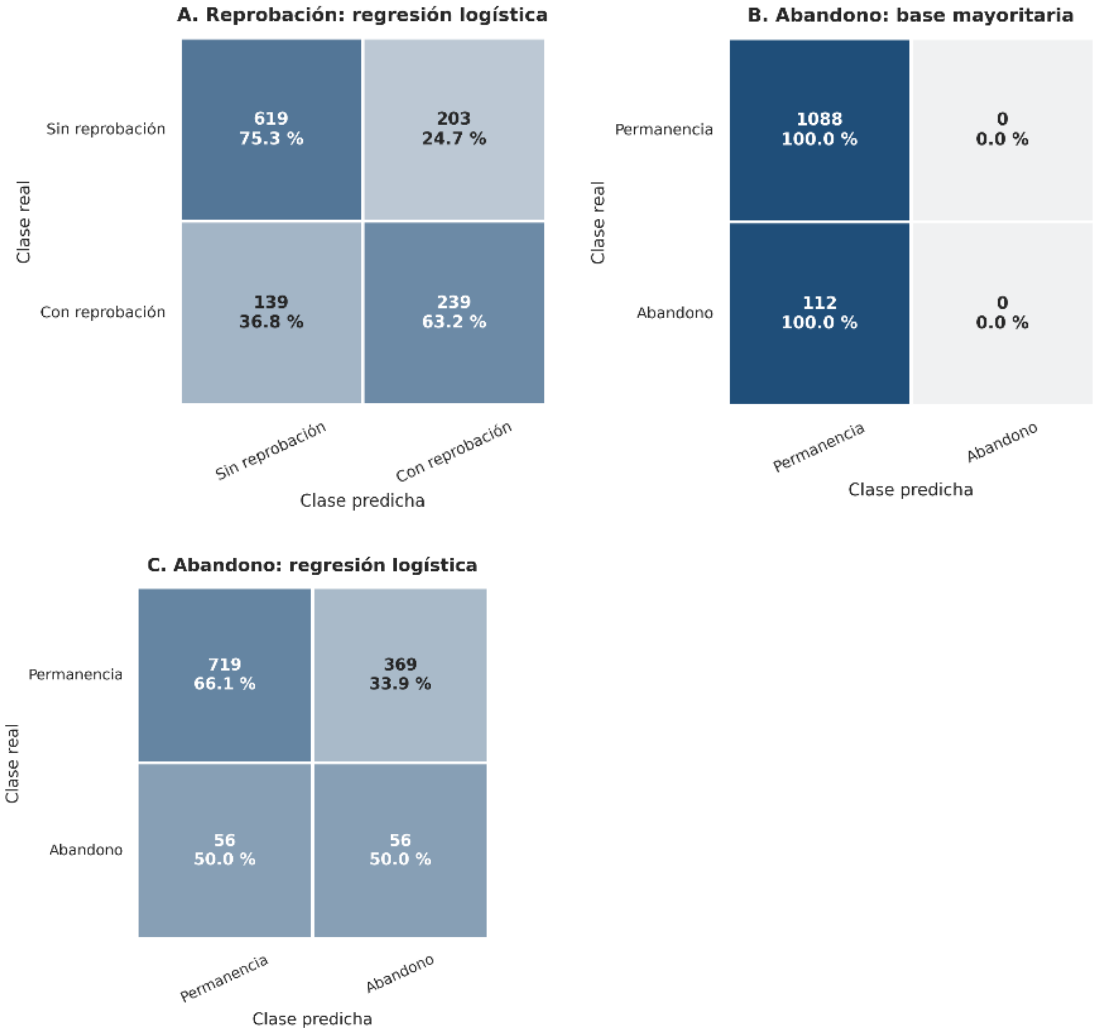


Nota. F1 corresponde a la clase positiva de riesgo. Los resultados proceden de 1,200 observaciones sintéticas y no representan historiales institucionales reales. Fuente: elaboración propia.

Las matrices de confusión de la Figura 3 permiten observar este comportamiento. Para la reprobación, la regresión logística identificó 239 de 378 casos positivos y clasificó correctamente 619 de 822 casos sin reprobación. En el abandono, la base mayoritaria clasificó incorrectamente los 112 casos positivos como permanencia. La regresión logística detectó 56 de esos 112 casos, pero produjo 369 alertas falsas. En consecuencia, la mejora de la sensibilidad implicó un aumento en el número de estudiantes que requerirían una revisión posterior por parte de tutores o responsables académicos.

Figura 3

Matrices de confusión seleccionadas para reprobación y abandono en el siguiente periodo

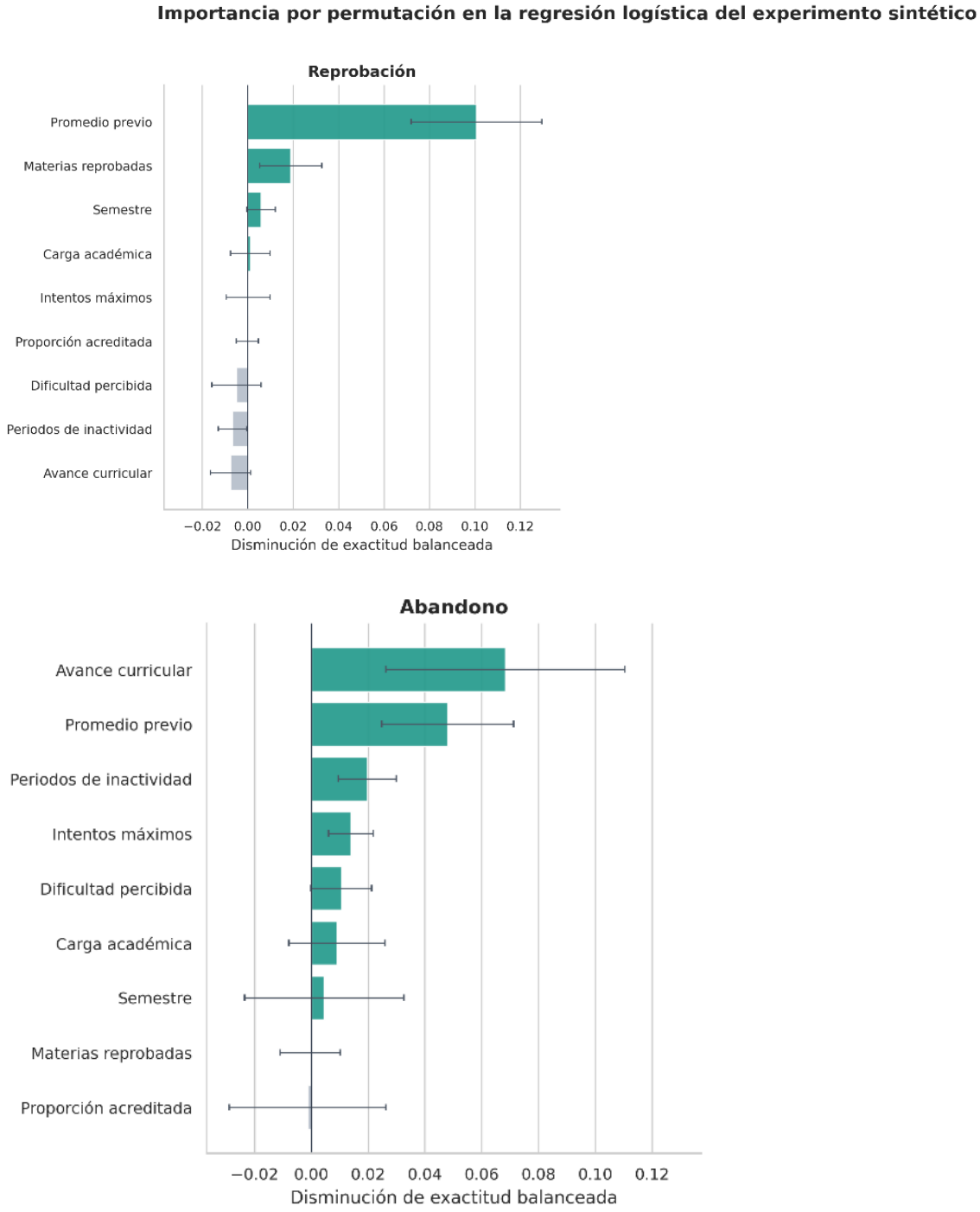


Nota. Cada celda presenta el número de observaciones y el porcentaje respecto de su clase real. Datos correspondientes al experimento controlado con perfiles sintéticos. Fuente: elaboración propia.

La importancia de la permutación mostró que, dentro del escenario simulado, el promedio previo presentó la mayor contribución en la clasificación de reprobación. En el abandono destacaron el avance curricular, el promedio previo y los periodos de inactividad. Las restantes variables mostraron contribuciones menores o intervalos que abarcaron valores cercanos a cero. Debido a las correlaciones existentes entre los indicadores académicos, estos resultados deben interpretarse como indicadores de importancia predictiva dentro de la simulación y no como relaciones causales.

Figura 4

Importancia por permutación de las variables en la regresión logística del experimento sintético



Nota. Las barras de error representan una desviación estándar en 30 permutaciones. Fuente: elaboración propia.

DISCUSIÓN

Los resultados del instrumento muestran una disposición favorable hacia un sistema que emita recomendaciones académicas personalizadas. Las medias obtenidas en utilidad, facilidad de uso y evaluación global sugieren que los participantes reconocen un valor potencial en SIDTAI. Sin embargo, las respuestas se recopilaron antes de una implementación completa y tras presentar a los estudiantes los beneficios esperados del sistema. En consecuencia, los resultados expresan expectativas y aceptación anticipada, no satisfacción derivada de una experiencia real de uso.

El alfa de Cronbach de 0.971 indica una alta asociación entre los reactivos. No obstante, un coeficiente cercano a uno no demuestra por sí solo la validez del instrumento. También puede reflejar reactivos parcialmente redundantes, una escasa diferenciación entre dimensiones o una tendencia general de aceptación. La interpretación del alfa debe considerar el número de reactivos, la correlación entre ellos y la estructura teórica de la escala (Cronbach, 1951; Xiao & Hau, 2023; Kalkbrenner, 2024). Por esta razón, en aplicaciones posteriores será necesario revisar la redacción de los reactivos, incorporar reactivos inversos o controles de atención y realizar análisis factoriales con una muestra más amplia.

La capacidad de los modelos para clasificar la evaluación global a partir de las dimensiones del mismo instrumento también debe interpretarse con prudencia. El desempeño elevado indica coherencia entre las respuestas de utilidad, facilidad, credibilidad, justicia y la evaluación general. No obstante, existe proximidad conceptual entre las variables predictoras y el resultado, por lo que este ejercicio constituye una clasificación exploratoria y no una validación externa del sistema. La validación cruzada permite estimar el comportamiento interno de los modelos, pero no sustituye la evaluación en una población independiente ni la validación temporal con nuevos participantes (Bates et al., 2024).

El experimento controlado confirmó que la exactitud puede producir conclusiones equivocadas cuando las clases están desbalanceadas. En abandono, un modelo que no detectó ningún caso de riesgo alcanzó un 90.7% de exactitud porque la permanencia fue la clase mayoritaria. Este resultado respalda la necesidad de utilizar exactitud balanceada, sensibilidad, especificidad, F1, área bajo la curva ROC y matrices de confusión para seleccionar modelos educativos (Chicco & Jurman, G., 2020; Papadogiannis et al., 2024).

La dificultad para detectar categorías minoritarias también se observa en estudios realizados con registros educativos reales. Polat & Horzum, (2026), por ejemplo, obtuvieron una exactitud global de 80.47 %, pero reportaron una sensibilidad de 45.8 % para estudiantes con graduación tardía. Este contraste muestra que una clasificación global aceptable puede coexistir con un desempeño limitado precisamente en los grupos que requieren mayor atención institucional.

En el escenario de abandono, la regresión logística ofreció el mejor equilibrio para identificar la clase minoritaria, aunque generó un número considerable de falsas alertas. Esto implica que el modelo no debería utilizarse para tomar decisiones automatizadas sobre los estudiantes. Su aplicación práctica más pertinente sería priorizar casos para una revisión humana posterior, en la que tutores y responsables académicos incorporen información contextual no incluida en los registros. La analítica debe ampliar la capacidad institucional de interpretación y de acompañamiento, no sustituir el juicio pedagógico (Guzmán-Valenzuela et al., 2021; Civit et al., 2024).

La novedad de la propuesta radica en integrar dos dimensiones que normalmente se estudian por separado: la percepción de credibilidad, la utilidad, la facilidad y la justicia del sistema, y la evaluación técnica de modelos predictivos mediante métricas sensibles al desbalance de clases. Esta integración permite plantear SIDTAI no solo como un clasificador, sino también como una herramienta de apoyo a decisiones que deberá ser técnicamente consistente, comprensible para los usuarios y percibida como justa.

Entre las aplicaciones prácticas se encuentra la incorporación de un módulo de alertas tempranas que muestre la probabilidad estimada, los factores relacionados con la alerta y el nivel de incertidumbre. La decisión final sobre el acompañamiento deberá quedar a cargo de la responsabilidad institucional. También será necesario ajustar los umbrales de clasificación de acuerdo con el costo relativo de no detectar a un estudiante en riesgo frente al de generar una alerta falsa.

Las principales limitaciones son el tamaño reducido de la muestra del instrumento, el predominio de estudiantes de segundo semestre, el muestreo voluntario, la posible influencia del mensaje de invitación y la ausencia de una prueba posterior al uso real de SIDTAI. En el componente computacional, la limitación central radica en el empleo de perfiles sintéticos. Aunque la simulación permitió comprobar el procedimiento analítico y comparar métricas, no permite afirmar que los mismos resultados se producirán con historiales del Tecnológico Nacional de México.

Como trabajo futuro, se propone recopilar historiales académicos anonimizados, establecer definiciones institucionales verificables de reprobación, rezago y abandono, efectuar una división temporal entre entrenamiento y prueba, validar el modelo en cohortes independientes, calibrar las probabilidades y examinar su comportamiento en diferentes grupos estudiantiles. Asimismo, será necesario evaluar la utilidad, la usabilidad, la credibilidad y la justicia percibida después de que los participantes interactúen con el sistema.

CONCLUSIONES

El estudio permitió establecer una evaluación preliminar de SIDTAI a partir de dos fuentes claramente diferenciadas: 54 respuestas genuinas sobre la percepción anticipada del sistema y un

experimento controlado con 1,200 observaciones académicas sintéticas. Los estudiantes participantes mostraron una percepción predominantemente favorable, especialmente en las dimensiones de utilidad, facilidad de uso y evaluación global.

El instrumento presentó una consistencia interna elevada; sin embargo, este resultado no debe interpretarse como evidencia suficiente de validez. La concentración de respuestas, los patrones constantes y la cercanía conceptual entre algunos reactivos justifican una revisión posterior del instrumento y su aplicación a una muestra más amplia tras el uso real del sistema.

El experimento computacional mostró que el modelo de mayor exactitud no necesariamente es el más adecuado para detectar estudiantes en riesgo. En abandono, la base mayoritaria alcanzó un 90.7% de exactitud, pero no identificó ningún caso positivo. La regresión logística obtuvo una exactitud menor, aunque presentó el mejor equilibrio entre sensibilidad, F1 y discriminación de la clase minoritaria. Por tanto, la selección de modelos para SIDTAI deberá priorizar métricas complementarias y considerar el costo educativo de los falsos negativos y de las falsas alertas.

SIDTAI demuestra su viabilidad como arquitectura de apoyo para convertir la información académica en alertas y recomendaciones interpretables. No obstante, su eficacia para anticipar reprobación, rezago o abandono deberá demostrarse mediante historiales institucionales anonimizados y su validación en cohortes reales. Hasta contar con esa evidencia, los resultados deben considerarse preliminares y el sistema deberá emplearse como apoyo al acompañamiento humano, nunca como sustituto de la decisión académica.

REFERENCIAS

Bates, S., Hastie, T., & Tibshirani, R. (2024). Cross-validation: What does it estimate and how well does it do it? *Journal of the American Statistical Association*, *119*(546), 1434-1445. https://doi.org/https://doi.org/10.1080/01621459.2023.2197686

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5-32. https://doi.org/https://doi.org/10.1023/A:1010933404324

Brooke, J. (1996). SUS: A «quick and dirty» usability scale. En P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.). *Usability evaluation in industry*, 189-194.

Chicco, D., & Jurman, G., G. (2020). The advantages of the Matthews correlation coefficient over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(6). https://doi.org/https://doi.org/10.1186/s12864-019-6413-7

Civit, M., Escalona, M. J., Cuadrado, F., & Reyes-de-Cozar, S. (2024). Class integration of ChatGPT and learning analytics for higher education. *Expert Systems*, *41*(12). https://doi.org/https://doi.org/10.1111/exsy.13703

Colquitt, J. A. (2001). On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386-400. https://doi.org/10.1037/0021-9010.86.3.386

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334. https://doi.org/https://doi.org/10.1007/BF02310555

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, *13*(3), 319-340. https://doi.org/10.2307/249008

Granić, A., & Marangunić, N. (2019). Technology acceptance model in educational context: A systematic literature review. *British Journal of Educational Technology*, *50*(5), 2572-2593. https://doi.org/10.1111/bjet.12864

Guzmán-Valenzuela, C., Gómez-González, C., Rojas-Murphy Tagle, A., & Lorca-Vyhmeister, A. (2021). Learning analytics in higher education: A preponderance of analytics but very little learning? *Int J Educ Technol High Educ*. https://doi.org/https://doi.org/10.1186/s41239-021-00258-x

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2023). Metrics for explainable AI: Challenges and prospects. *AI Magazine*, *44*(2), 165-183. https://doi.org/10.1002/aaai.12073

Kalkbrenner, M. T. (2024). Choosing between Cronbach's coefficient alpha, McDonald's coefficient omega, and coefficient H: Confidence intervals and the advantages and drawbacks of interpretive guidelines. *Measurement and Evaluation in Counseling and Development*, *57*(2), 93-105. https://doi.org/https://doi.org/10.1080/07481756.2023.2283637

Martínez Plumed, F., Ferri, C., Nieves, D., & Hernández Orallo, J. (2021). *Missing the missing values: The ugly duckling of fairness in machine learning*. https://doi.org/10.1002/int.22415

Papadogiannis, I., Wallace, M., & Karountzou, G. (2024). Educational Data Mining: A Foundational Overview. *Encyclopedia*, *4*, 1644-1664. https://doi.org/https://doi.org/10.3390/encyclopedia4040108

Polat, A., & Horzum, M. B. (2026). Predicting Student Outcomes in Open High Schools Using Educational Data Mining. *International Review of Research in Open and Distributed Learning*, *27*(3).

Schröer, C., Kruse, F., & Gómez, J. M. (2021). A systematic literature review on applying CRISP-DM process model. *Procedia Computer Science*, *181*, 526-534. https://doi.org/10.1016/j.procs.2021.01.199

Vaarma, M., & Li, H. (2024). Predicting student dropouts with machine learning: An empirical study in Finnish higher education. *Technology in Society*, *76*(102474). https://doi.org/https://doi.org/10.1016/j.techsoc.2024.102474

Xiao, L., & Hau, K. T. (2023). Performance of coefficient alpha and its alternatives: Effects of different types of non-normality. *Educational and Psychological Measurement*, *83*(1), 5-27. https://doi.org/https://doi.org/10.1177/00131644221088240

Yadav, N. R., & Deshmukh, S. S. (2023). Prediction of Student Performance Using Machine Learning Techniques: A Review. *ICAMIDA 2022*, (05), 735-741. https://doi.org/https://doi.org/10.2991/978-94-6463-136-4_63