ISSN-e: 2737-6419 Revista Athenea
Período: enero-marzo de 2026 Vol.7, Número 23, (pp. 66–76)

Artículo de investigación https://doi.org/10.47460/athenea.v7i23.134

Modelos explicables para identificar dificultades en matemáticas universitarias: análisis del dataset MathE

Eduardo Pozo Valdiviezo*
https://orcid.org/0000-0003-0480-5669
eduardo.pozo@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Sandra Elizabeth Tenelanda Cudco
https://orcid.org/0000-0001-6215-9517
stenelanda@unach.edu.ec
Universidad Nacional de Chimborazo
Riobamba, Ecuador

Martha Ximena Dávalos Villegas
https://orcid.org/0000-0001-7865-6307
martha.davalos@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

Gustavo Javier Ávila Gaibor
https://orcid.org/0009-0005-6873-7927
gustavo.avila@espoch.edu.ec
Escuela Superior Politécnica de Chimborazo
Riobamba, Ecuador

*Autor de correspondencia: eduardo.pozo@espoch.edu.ec

Recibido: (28/10/2025), Aceptado: (11/01/2026)

Resumen. Este estudio aplica un enfoque de analítica del aprendizaje y modelos explicables para identificar contenidos de mayor dificultad en matemáticas universitarias a partir de registros de interacción de una plataforma de práctica y evaluación. Se realizó un análisis cuantitativo secundario del conjunto de datos MathE, considerando variables de contenido (tema, subtema y palabras clave) y contexto (país y nivel). Primero se estimaron tasas de error por tema y subtema y se sintetizaron patrones mediante visualizaciones comparativas. Luego se entrenaron modelos complementarios, una regresión logística por su interpretabilidad y un modelo no lineal de árboles de gradiente para capturar interacciones, validando la generalización con partición por estudiante. La explicabilidad se abordó mediante atribución de contribuciones para interpretar factores asociados al error. Los hallazgos señalan mayores dificultades en contenidos de Diferenciación, Interpretación Funcional y Probabilidad, junto con debilidades transversales de manipulación algebraica, con apoyos adicionales en Métodos Numéricos e Integración.

Palabras clave: analítica del aprendizaje, educación matemática, inteligencia artificial explicable, orientación didáctica.

Explainable Models for Identifying Difficulties in University Mathematics: Analysis of the MathE Dataset

Abstract. This study applies a learning analytics approach and explainable models to identify the most difficult content areas in university mathematics based on interaction records from a practice and assessment platform. A secondary quantitative analysis of the MathE dataset was conducted, considering content variables (topic, subtopic, and keywords) and contextual variables (country and level). First, error rates were estimated by topic and subtopic, and patterns were synthesized through comparative visualizations. Then, complementary models were trained: a logistic regression model for its interpretability and a nonlinear gradient-boosted tree model to capture interactions, validating generalization through student-level partitioning. Explainability was addressed through contribution attribution to interpret factors associated with errors. The findings indicate greater difficulties in Differentiation, Functional Interpretation, and Probability, together with cross-cutting weaknesses in algebraic manipulation, with additional support needs in Numerical Methods and Integration.

Keywords: learning analytics, mathematics education, explainable artificial intelligence, instructional guidance.

I. INTRODUCCIÓN

Las dificultades persistentes en asignaturas de matemáticas en la educación superior se reflejan no solo en bajas calificaciones o repetición de curso, sino también en trayectorias académicas irregulares y desmotivación temprana. En este contexto, los docentes y la administración académica suelen tomar decisiones con información incompleta; puesto que, saben que existen temas difíciles, pero no siempre se dispone de evidencia sistemática sobre qué temas generan más errores y en qué condiciones ocurren, especialmente cuando la evaluación se apoya en recursos digitales.

El crecimiento de las plataformas de aprendizaje y evaluación ha ampliado la posibilidad de observar el aprendizaje mediante analítica del aprendizaje, es decir, utilizando datos generados por la interacción de los estudiantes con actividades y recursos. Este enfoque permite describir patrones de rendimiento y orientar mejoras pedagógicas con base empírica, en lugar de basarse únicamente en percepciones o evidencia anecdótica [1]. Sin embargo, una limitación común en la investigación regional es que muchos estudios se basan en datos institucionales cerrados o difíciles de reutilizar, lo que reduce la replicabilidad y la comparabilidad entre contextos.

Paralelamente, el interés en aplicar inteligencia artificial (IA) en educación matemática ha aumentado en los últimos años, especialmente para predecir el rendimiento y personalizar el apoyo. Las revisiones y mapeos bibliométricos muestran que este campo se ha expandido y diversificado, consolidando el uso de métodos basados en datos en la educación matemática [2]. Sin embargo, un problema persistente es que los modelos de alta precisión a menudo funcionan como cajas negras, lo que dificulta traducir sus resultados en acciones didácticas concretas. Por lo tanto, la inteligencia artificial explicable (XAI por sus siglas en inglés) ha cobrado impulso en la educación, con el objetivo de generar explicaciones comprensibles y útiles para docentes y estudiantes, y fortalecer la confianza y la toma de decisiones informada [3].

Con el objetivo de proporcionar evidencia accesible y replicable, este estudio utiliza MathE, un conjunto de datos abierto asociado a una plataforma de práctica y evaluación de matemáticas para la educación superior. El conjunto de datos contiene 9546 respuestas de 372 estudiantes a 833 preguntas e incluye variables como país, nivel de pregunta, tema, subtema y palabras clave [4]. En particular, contar con estos atributos permite a los investigadores no solo estimar la dificultad de un tema, sino también construir modelos que expliquen qué características se asocian más con los errores, lo que contribuye a una interpretación pedagógica más directa.

En este marco, el objetivo del trabajo fue analizar, a través de un enfoque de analítica del aprendizaje con énfasis en la explicabilidad, los patrones de acierto y error en matemáticas universitarias utilizando el conjunto de datos abiertos MathE, con el fin de identificar contenidos de mayor dificultad y variables asociadas que puedan orientar las decisiones didácticas y de evaluación. Debido a que el dataset está compuesto por trazas de interacción y etiquetas de contenido, el estudio se plantea con un enfoque exploratorio y diagnóstico, priorizando la interpretabilidad y la utilidad pedagógica; por ello, el modelado se interpreta principalmente como un soporte explicable para identificar contenidos críticos más que como un sistema de predicción individual de alto desempeño.

El artículo se organiza de la siguiente manera: luego de esta introducción, se presentan los fundamentos teóricos sobre analítica de aprendizaje, inteligencia artificial en educación matemática y explicabilidad; luego se describe la metodología y el tratamiento de los datos; posteriormente, se presentan los resultados y su discusión; y finalmente, se resumen las conclusiones y principales implicaciones del estudio.

II. MARCO TEÓRICO

La enseñanza y el aprendizaje de las matemáticas en la educación superior se desarrollan en un contexto de dos desafíos persistentes; por un lado, la heterogeneidad en la preparación previa de los estudiantes y, por otro, la necesidad de evidencia objetiva que oriente las decisiones en materia de docencia, evaluación y apoyo académico. En este escenario, el auge de los entornos digitales de práctica y evaluación ha permitido registrar rastros de interacción (respuestas correctas o incorrectas, selección de contenidos, niveles de dificultad), lo que facilita una comprensión más precisa de qué contenidos presentan dificultades y en qué circunstancias surgen.

A. Analítica del aprendizaje e inteligencia artificial en educación matemática

La analítica del aprendizaje se ha consolidado como un enfoque centrado en la recopilación, medición, análisis y generación de informes de datos sobre los estudiantes y sus contextos para comprender y optimizar el aprendizaje y los entornos en los que se desarrolla [5]. En la práctica, este enfoque está estrechamente vinculado a la minería de datos educativos, ya que ambas comunidades convergen en métodos para extraer patrones procesables de los datos educativos como la predicción del rendimiento, la identificación de estructuras de aprendizaje y la visualización para revisión humana [6].

En la educación matemática, la integración de técnicas de inteligencia artificial y analítica ha crecido rápidamente, impulsada por la disponibilidad de datos y la necesidad de personalizar el apoyo en cursos con alta complejidad conceptual. Un mapeo bibliométrico y una revisión sistemática ampliamente citados identifican temas recurrentes como los sistemas de tutoría inteligente, el aprendizaje adaptativo, la predicción del rendimiento y el apoyo a la toma de decisiones pedagógicas [2]. Sin embargo, este crecimiento conlleva un requisito metodológico adicional: los resultados deben ser interpretables y relacionables con decisiones educativas concretas, evitando que el modelado se limite a mejorar las métricas predictivas sin traducirse en orientación didáctica.

En este marco se encuentra el uso de repositorios y plataformas para la práctica matemática en la educación superior. Un ejemplo es MathE, una plataforma que almacena las respuestas de los estudiantes a bancos de preguntas organizados por tema, subtema y nivel. Su conjunto de datos se ha descrito como un recurso para el estudio del aprendizaje y la evaluación en matemáticas universitarias, registrando las respuestas (correcta o incorrectas) y los metadatos de los ítems junto con información contextual (por ejemplo, país del estudiante) [4]. Este tipo de datos permite análisis que abarcan desde la caracterización descriptiva de las dificultades hasta modelos estadísticos que explican la probabilidad de error basándose en los atributos y el contexto de los ítems.

B. Medición de la dificultad y modelización del error en los ítems

En las evaluaciones basadas en ítems, una primera aproximación a la dificultad es la proporción de respuestas correctas o incorrectas por ítem o por área de contenido (tasa de error). Existen enfoques de medición más formales, como la teoría de respuesta al ítem (TRI) y el modelo de Rasch, que permiten calibrar ítems y separar la habilidad del estudiante y la dificultad del ítem [7]; no obstante, su implementación requiere supuestos y procedimientos que superan el alcance de este estudio. Por lo tanto, la tasa de error puede utilizarse como indicador empírico de dificultad y complementarse con modelos predictivos explicables para la interpretación de los factores asociados al error.

En lugar de ajustar un modelo IRT en este estudio, el componente explicativo se apoya en un modelo lineal generalizado y un modelo no lineal de árboles de gradiente, incorporando covariables de contenido (tema, subtema y palabras clave) y de contexto (país y nivel). Esta estrategia permite estimar asociaciones interpretables y detectar interacciones entre contenidos, manteniendo trazabilidad para las decisiones didácticas.

De forma complementaria, y especialmente apropiada al trabajar con variables de contenido categóricas, la regresión logística ofrece un marco flexible para modelar la probabilidad de error basándose en predictores observables, y puede extenderse a estructuras jerárquicas cuando hay respuestas repetidas por estudiante y por pregunta. En su formulación estándar, modela el logit de la probabilidad de un evento como una combinación lineal de variables explicativas; su uso está ampliamente establecido en la investigación aplicada debido a su interpretabilidad mediante razones de momios y su capacidad para incorporar múltiples factores [8]. En el contexto de plataformas como MathE, esto permite evaluar, por ejemplo, si ciertos temas o niveles están asociados con mayores probabilidades de error, controlando otros atributos del ítem.

C. Modelos explicables y su valor didáctico

El auge de los modelos productivos en educación ha puesto de manifiesto una tensión; los modelos más precisos tienden a ser menos interpretables, lo que limita su utilidad para docentes, coordinadores y estudiantes. Esto subraya la relevancia de la inteligencia artificial explicable (XAI) en educación, entendida como un conjunto de métodos que buscan transparentar las razones de una predicción o clasificación, atendiendo a necesidades específicas del contexto educativo (confianza, responsabilidad) [3]. En la práctica, la explicabilidad permite que un modelo pase de predecir quién fallará a explicar

qué factores y contenidos se asocian con el fracaso, que es precisamente el tipo de evidencia útil para planificar el refuerzo, ajustar las secuencias de contenido o revisar el diseño de los ítems.

Entre los modelos de inteligencia artificial explicable más utilizados se encuentra Explicaciones Aditivas de Shapley (SHAP), que propone una familia de explicaciones aditivas basadas en valores de Shapley para atribuir las contribuciones de cada variable a una predicción específica [9]. Su adopción en educación es frecuente porque permite tanto explicaciones globales como locales, lo que facilita una comprensión clara para la toma de decisiones pedagógicas, sin abandonar los modelos no lineales cuando estos ofrecen capacidad predictiva.

En resumen, el marco teórico del estudio articula: la analítica del aprendizaje como fundamento para la explotación de datos de interacción educativa, el modelado del desempeño en ítems desde aproximaciones de medición y regresión, y la explicabilidad como condición para convertir resultados predictivos en insumos interpretables y procesables para la enseñanza.

III. METODOLOGÍA

El estudio empleó un enfoque cuantitativo, basado en el análisis secundario del archivo MathE dataset, que recopila registros de las respuestas de estudiantes universitarios a preguntas de matemáticas en una plataforma de práctica y evaluación. La unidad de análisis fue cada intento de respuesta. Las variables consideradas para este estudio fueron el identificador del estudiante, el país, el identificador de la pregunta, el tipo de respuesta (correcta/incorrecta), el nivel de la pregunta (básico/avanzado), el tema, el subtema y las palabras clave. Se analizaron un total de 9546 respuestas de 372 estudiantes y 833 preguntas.

Puesto que cada estudiante responde a múltiples ítems y cada ítem es respondido por varios estudiantes, los datos presentan una estructura de clasificación cruzada. En este estudio el objetivo principal es el diagnóstico basado en el contenido y la explicación de los patrones de error; por lo tanto, se priorizó una evaluación predictiva robusta mediante partición por estudiante y validación cruzada por grupos, evitando que el modelo "aprenda" a estudiantes específicos. Se reconoce que un enfoque inferencial integral podría incorporar un modelo logístico mixto con interceptos aleatorios por estudiante y por ítem; esta extensión se plantea como una vía inmediata para la robustez en futuras versiones del estudio, especialmente si se busca inferencia formal sobre los efectos con supuestos de independencia más estrictos.

A. Preparación y depuración de los datos

Inicialmente, se preparó el conjunto de datos, se importó el archivo csv de MathE y se verificó la consistencia de la estructura de columnas y la codificación de categorías. Se revisaron los valores faltantes y los posibles duplicados; se eliminaron los registros exactamente iguales para evitar la doble contabilización, mientras que se conservaron los registros correspondientes a diferentes intentos, ya que representaban interacciones reales dentro de la plataforma. Para garantizar la consistencia de las comparaciones, se estandarizaron las etiquetas de país, nivel, tema y subtema. La variable dependiente del estudio, definida como error (1=respuesta incorrecta y 0=respuesta correcta), se construyó a partir del campo "Tipo de respuesta". El campo "Palabras clave" se procesó como un conjunto de etiquetas separadas por comas.

Para el control de la dimensionalidad, se definió un umbral mínimo de frecuencia de 100 ocurrencias, y las 40 palabras clave más frecuentes ($K = 40$) se mantuvieron dentro de este conjunto, incorporándolas como variables binarias. Con la codificación *one-hot* (usando una categoría de referencia para evitar multicolinealidad perfecta) se obtuvo un total de 84 predictores: país (7), nivel (1), tema (13), subtema (23) y palabras clave (40), además del intercepto. Para mitigar el riesgo de multicolinealidad, se evitó la trampa de variables ficticias empleando una categoría de referencia, y se empleó regularización en la regresión logística; en el modelo de árboles de gradiente, la colinealidad no afecta el ajuste de la misma forma.

B. Análisis descriptivo de dificultad por contenido

Posteriormente, se realizó un análisis descriptivo para caracterizar la dificultad por área de contenido. Se estimaron las frecuencias y los porcentajes de participación por país, nivel, tema y subtema para contextualizar el peso relativo de cada grupo dentro del conjunto de datos. Luego, se calculó la tasa de

error para cada tema y subtema, y se comparó el comportamiento entre niveles (básico y avanzado). Para priorizar el contenido crítico, se elaboró una clasificación de los subtemas con las tasas de error más altas, considerando también el volumen de respuestas por categoría para evitar conclusiones basadas en muestras muy pequeñas. La incertidumbre de las estimaciones se abordó mediante intervalos de confianza calculados con una aproximación binomial o un remuestreo, según correspondiera al tamaño de la muestra.

C. Modelado predictivo e interpretable

La fase de modelado tuvo como objetivo estimar la probabilidad de error y, en particular, identificar factores asociados de forma interpretable. Se incorporaron como predictores del nivel, el país, el tema, el subtema y las palabras clave seleccionadas. Las variables categóricas se transformaron mediante codificación *one-hot*, mientras que, las palabras clave se incluyeron como indicadores binarios. Se entrenaron dos enfoques complementarios: un modelo de regresión logística de referencia, debido a su interpretabilidad directa mediante *odds ratios*, y un modelo no lineal basado en árboles de gradiente, diseñado para capturar relaciones e interacciones complejas entre el contenido y las etiquetas. Se evaluó el equilibrio entre clases (correcto e incorrecto) y, de ser necesario, se aplicaron ponderaciones de clase para reducir los sesgos derivados de una posible desproporción. Para el modelo de árboles de gradiente, se realizó un ajuste de hiperparámetros mediante validación cruzada por grupos (estudiantes) en el conjunto de entrenamiento, explorando combinaciones de profundidad del árbol, tasa de aprendizaje, número de estimadores y submuestreo. La configuración final se seleccionó maximizando el AUC promedio en la validación.

D. Evaluación del rendimiento

Para evaluar el rendimiento y reducir el riesgo de sobrestimación debido a la dependencia entre registros, se realizó una validación separando los datos por estudiante, asegurando que el conjunto de prueba no compartiera participantes con el conjunto de entrenamiento. Se aplicó una validación cruzada entre los grupos durante el entrenamiento para ajustar el modelo no lineal y verificar la estabilidad. Las métricas reportadas incluyeron el área bajo la curva característica operativa del receptor (AUC-ROC), exactitud, precisión, completitud, puntuación $F1$ y una medida de calibración, así como matrices de confusión para describir los errores de clasificación. Se reportaron intervalos de confianza al 95% para las métricas principales utilizando un *Bootstrap* estratificado por estudiante ($B = 2000$) sobre el conjunto de prueba, con el fin de reflejar la variabilidad bajo la dependencia por participante.

E. Explicabilidad y herramientas de análisis

Finalmente, la explicabilidad se abordó de dos maneras. En la regresión logística, los coeficientes se interpretaron como cambios relativos en la razón de probabilidades de error, manteniendo constantes todas las demás variables. En el método no lineal, se emplearon explicaciones globales y locales mediante un método de atribución de contribuciones de variables (SHAP) para identificar qué factores aumentan o disminuyen la probabilidad de error y cómo estos efectos varían según el contenido. El informe priorizó las visualizaciones y los resúmenes concisos. El procesamiento, análisis y visualización se realizaron en Python, utilizando pandas y NumPy para el tratamiento de datos, se usó scikit-learn para la regresión logística y validación, para el modelo de árboles de gradiente se empleó XGBoost y Excel para la elaboración de figuras.

IV. RESULTADOS

Se analizaron 9546 intentos de respuestas de 372 estudiantes en 833 preguntas, con datos de 8 países, 14 temas y 24 subtemas. La variable dependiente se definió como error ($1 =$ respuesta incorrecta; $0 =$ respuesta correcta). La tasa de error general fue del $53,2\%$, lo que confirma un nivel de dificultad suficiente para identificar patrones por contenido y respaldar un análisis explicable [4].

En términos de contexto, la participación fue desigual entre países, mostrando que Portugal representó la mayoría de respuestas (5495), seguido de Lituania (1443) e Italia (1358), mientras que, otros países contribuyeron con volúmenes menores. Este desequilibrio sugiere interpretar el país como una variable contextual y centrar las conclusiones en el contenido matemático (tema, subtema y palabras clave) [10]. Por nivel, se observó una tasa de error ligeramente mayor en el Básico que en el Avanzado, lo que sugiere que la dificultad empírica está más estrechamente relacionada con el contenido específico que con la etiqueta del nivel [11].

Debido al predominio de Portugal en el registro, el "país" se interpretó como un factor contextual, y la sensibilidad del diagnóstico basado en el contenido se verificó mediante un análisis de sensibilidad. Las clasificaciones de dificultad se replicaron en el subconjunto de países con mayor volumen, y se comparó con el subconjunto "Portugal vs no-Portugal". El patrón general se mantuvo; las mayores fuentes de error se concentraron en el contenido relacionado con la Derivación, Interpretación funcional y probabilidad. Por ello, las conclusiones se formulan principalmente a nivel de contenido y el sesgo de composición se reconoce como limitación.

Para identificar las áreas de mayor dificultad, se estimó la tasa de error de cada tema. La Fig. 1 resume la clasificación de dificultad por tema y muestra que los errores más frecuentes se concentran en Derivación, Funciones Reales de una Variable, Probabilidad, Optimización y Métodos Numéricos. Este patrón sugiere que las dificultades más frecuentes se relacionan con contenidos donde coexisten procedimientos y comprensión conceptual (reglas de diferenciación y lectura funcional), lo que requiere intervenciones didácticas más específicas que la práctica habitual [12].

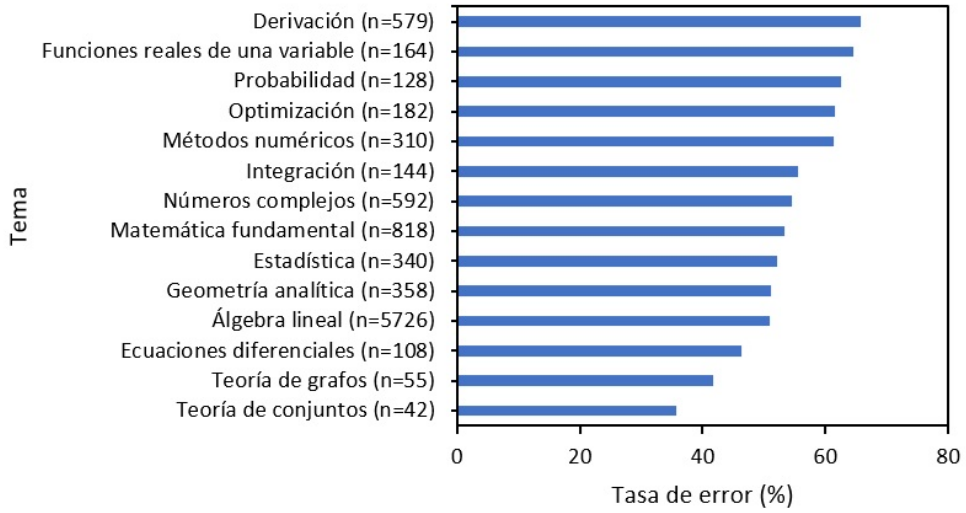

Fig. 1. Tasa de error por tema (ordenando de menor a mayor).

Para traducir el diagnóstico en decisiones pedagógicas concretas, se estimó la tasa de error para cada subtema, priorizando aquellos con evidencia suficiente ($n \geq 100$) para evitar conclusiones basadas en muestras pequeñas. La Fig. 2 muestra que los subtemas con las tasas de error más altas incluyen Diferenciación Parcial, Dominio, Imagen y Gráficos, y Derivadas, además de áreas de enfoque significativas Probabilidad, Métodos Numéricos, Técnica de Integración y habilidades fundamentales como Expresiones Algebraicas, Ecuaciones y Desigualdades. Esta combinación sugiere que una parte significativa del error no se atribuye únicamente al cálculo, sino también a dificultades de interpretación y habilidades algebraicas transversales, coherente con hallazgos recientes que advierten que las trazas suelen capturar predominantemente comportamiento y requieren complementar su lectura con marcos pedagógicos [10].

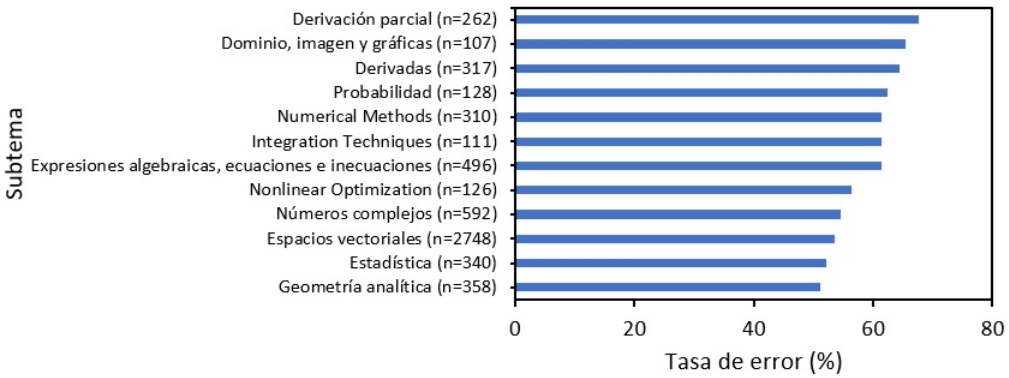

Fig. 2. Subtemas con mayor tasa de error ($n \geq 100$).

Para integrar simultáneamente el tema y el subtema, se generó un mapa de calor de la tasa de error para cada combinación. La Fig. 3 confirma visualmente que los picos de error no se distribuyen uniformemente, ya que se concentran en combinaciones específicas, especialmente en áreas relacionadas con la diferenciación y la interpretación funcional, con énfasis adicional en la Probabilidad y los Métodos Numéricos. Este resultado es útil para diseñar actividades de refuerzo secuencial, manteniendo la coherencia curricular, y coincide con tendencias recientes que recomiendan el uso de visualizaciones orientadas al aprendizaje para apoyar decisiones pedagógicas procesables [12].

Para sintetizar los hallazgos descriptivos y facilitar su lectura, la Tabla 1 presenta una selección de los temas y subtemas con mayor dificultad empírica, indicando el tamaño de la muestra (n) y la tasa de error (%). Esta síntesis permite priorizar el contenido crítico con base en la evidencia, complementando el patrón conjunto mostrado en la Fig. 3. lo que es consistente con la literatura que enfatiza reportes concisos y orientados a intervención cuando se usa analítica en educación superior [11].

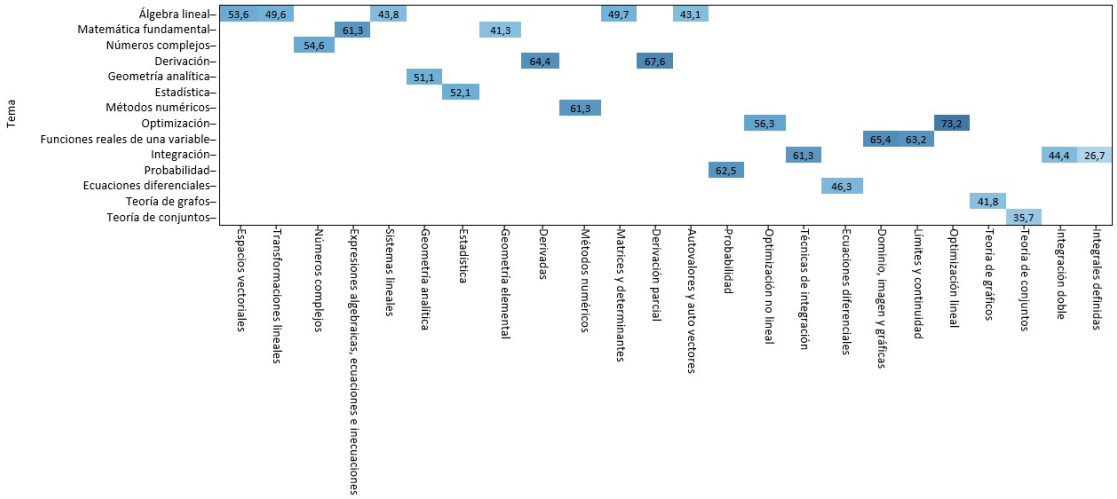

Fig. 3. Mapa de calor de la tasa de error por tema y subtema. Los valores y colores representan la tasa de error (%), calculada como respuestas incorrectas/total; valores más altos indican mayor dificultad empírica. Las celdas vacías corresponden a combinaciones sin registros.

Tabla 1. Síntesis de dificultad por contenido (temas y subtemas críticos).

Nivel	**Categoría**	n	**Tasa de error (%)**
Tema	Diferenciación	579	65,8
	Funciones reales de una sola variable	164	64,6
	Probabilidad	128	62,5
	Optimización	182	61,5
	Métodos numéricos	310	61,3
Subtema ($n \geq 100$)	Diferenciación parcial	262	67,6
	Dominio, imagen y gráficos	107	65,4
	Derivadas	317	64,4
	Probabilidad	128	62,5
	Métodos numéricos	310	61,3
	Expresiones algebraicas, ecuaciones e inecuaciones	496	61,3
	Técnicas de integración	111	61,3

Luego del diagnóstico basado en el contenido, se entrenaron dos enfoques complementarios; una regresión logística para su interpretabilidad (*odds ratios*) y un modelo de árbol de gradiente no lineal para capturar las relaciones e interacciones entre el tema, el subtema y las palabras clave, usando un enfoque de *boosting* ampliamente adoptado en problemas de clasificación [13]. La evaluación se realizó con partición por estudiante.

Para contextualizar el rendimiento de los modelos, se incluyeron dos *baselines* explícitos; un modelo nulo (solo intercepto), equivalente a predecir una probabilidad constante igual a la prevalencia de error; y un clasificador trivial que siempre predice la clase mayoritaria (error), sin utilizar información de contenido. En los dos casos, el AUC-ROC es de aproximadamente $0,5$, lo que representa un nivel de discriminación cercano al azar.

Los AUC-ROC obtenidos ($\approx 0,54$) indican que, con las variables disponibles, la capacidad del modelo para discriminar entre respuestas correctas e incorrectas es limitada. En consecuencia, el objetivo del modelado se interpreta como explicativo-diagnóstico en lugar de como un sistema de predicción individual de alto rendimiento. Este resultado sugiere que las mejoras en el rendimiento requerirían variables no incluidas en el dataset (por ejemplo, historial de intentos, tiempo de respuesta, entre otros).

Para comprobar el rendimiento predictivo sin sobreestimar por dependencia entre intentos, la evaluación se realizó con partición por estudiante. En este escenario, ambos modelos alcanzaron una capacidad de discriminación modesta pero consistente; la regresión logística obtuvo AUC-ROC $= 0,537$ y $F1 = 0,666$, mientras que el modelo de árboles de gradiente alcanzó AUC-ROC $= 0,544$ y $F1 = 0,678$. En los dos casos, el patrón de errores en la matriz de confusión evidenció una mayor sensibilidad para detectar respuestas incorrectas (completitud cercana a $0,70$) que, para reconocer respuestas correctas, lo que es coherente con un problema dominado por señales de contenido. La calibración fue estable en términos generales (Brier $0,25$), por lo que las probabilidades estimadas son útiles para análisis comparativos por tema/subtema más que para decisiones individuales de alto riesgo. La Tabla 2 resume las métricas principales.

Tabla 2. Desempeño de los modelos y *baselines* (validación por estudiante).

Modelo	AUC-ROC	Exactitud	Precisión	Completitud	F1	Brier
Baseline 1: modelo nulo	0,50	-	-	-	-	0,247
Baseline 2: trivial	0,50	0,590	0,590	1,00	0,742	0,410
Regresión logística	0,537	0,578	0,625	0,713	0,666	0,248
Árboles de gradiente	0,544	0,589	0,631	0,734	0,678	0,248

Nota: Baseline 1 se reporta como modelo probabilístico constante. F1 del trivial puede salir alto porque predice todo como error (recall=1), por eso el AUC y la calibración son más informativos para comparar modelos en este caso.

En la partición por estudiante, la regresión logística alcanzó AUC-ROC$= 0,537$ (IC95%: $0,471$–$0,603$) y F1$= 0,666$ (IC95%: $0,619$–$0,709$). El modelo de árboles de gradiente obtuvo AUC-ROC$= 0,544$ (IC95%: $0,483$–$0,607$) y F1$= 0,678$ (IC95%: $0,627$–$0,725$). Los intervalos de confianza se estimaron mediante *Bootstrap* por estudiante ($B = 2000$) sobre el conjunto de prueba. El modelo no lineal se entrenó con hiperparámetros seleccionados por validación (profundidad baja y tasa de aprendizaje moderada), priorizando la estabilidad y evitando el sobreajuste dada la estructura dependiente de los datos.

Además de la puntuación Brier (Tabla 2), la calibración se evaluó mediante un diagrama de fiabilidad (*bins* de probabilidad) y con pendiente e intercepto de calibración. En general, las probabilidades predichas mostraron un rango relativamente estrecho y una calidad de calibración coherente con el rendimiento discriminativo moderado; por lo tanto, las salidas probabilísticas se interpretan como útiles para la comparación y el diagnóstico por contenido, más que para decisiones individuales de alto riesgo.

La explicabilidad del modelo no lineal se resume en la Fig. 4, que tiene significancia global según las contribuciones de tipo SHAP, un marco ampliamente utilizado para explicar modelos basados en árboles y consolidar interpretaciones globales a partir de explicaciones locales [14]. El patrón respalda la interpretación descriptiva, mostrando que, la señal dominante proviene de las variables de contenido (tema, subtema) y de etiquetas (palabras clave) asociadas con procedimientos críticos, lo que sugiere que la probabilidad de error aumenta cuando los intentos se vinculan a reglas y transformaciones específicas (por ejemplo, derivación y simplificación), en lugar de depender del nivel general de la pregunta. En consecuencia, el valor del modelado en este estudio no reside en predecir con alta precisión, sino en explicar de forma práctica qué componentes de contenido se asocian con mayores tasas de error para orientar refuerzos didácticos y decisiones de evaluación, alineados con principios de XAI orientados a interpretabilidad y uso responsable [14], [15]. Dado el modesto desempeño discriminativo (Tabla 2), las explicaciones SHAP se interpretan como exploratorias y apuntan a identificar señales de contenido consistentes con el diagnóstico descriptivo, más que como evidencias causales o como una explicación fiable a nivel individual.

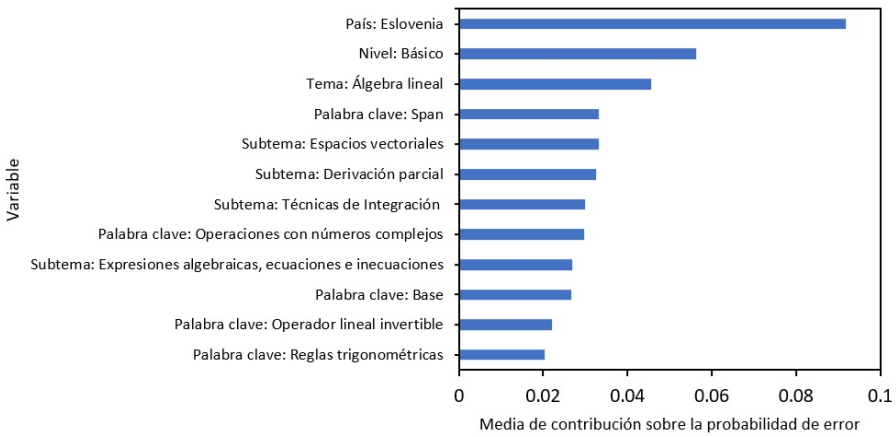

Fig. 4. Importancia global (SHAP) del modelo de árboles de gradiente (12 primeros). La barra representa la media de la contribución absoluta de cada predictor sobre la probabilidad de error.

En conjunto, las evidencias presentadas (clasificación por tema, priorización por subtema, patrón tema-subtema y explicabilidad general del modelo) convergen en una implicación clara. Los contenidos con mayor necesidad de refuerzo se concentran en la Diferenciación, la Interpretación Funcional, la Probabilidad y las habilidades transversales como la Simplificación y la Manipulación Algebraica, con apoyo adicional en Métodos Numéricos y Técnicas de Integración. Esta convergencia permite pasar de una percepción general de temas difíciles a un mapa de contenidos específicos que puede orientar las decisiones didácticas y de evaluación con base empírica.

CONCLUSIONES

Los resultados obtenidos del conjunto de datos MathE muestran una tasa de error global del $53,2\%$, lo que confirma un nivel de dificultad suficiente para caracterizar patrones de rendimiento estables por contenido. En cuanto a la dificultad curricular, el diagnóstico por tema y subtema indica que la mayor necesidad de refuerzo se concentra en Diferenciación, Interpretación Funcional y Probabilidad, acompañada de dificultades transversales en la manipulación y simplificación algebraica, con apoyo adicional en Métodos Numéricos y Técnicas de Integración.

El análisis también confirma que la dificultad observada depende más del contenido específico (tema, subtema y etiquetas) que de la etiqueta global de nivel (básico, avanzado), lo que es relevante para el diseño de secuencias de refuerzo. En lugar de incrementar la dificultad general, es aconsejable intervenir en subprocesos específicos que se repiten en diferentes temas.

El componente de modelado explicable aporta un valor práctico adicional. Tanto el enfoque interpretable (regresión logística) como el modelo no lineal (árboles de gradiente) muestran que los predictores que más contribuyen se asocian principalmente con los atributos de contenido (tema, subtema y palabras clave). En consecuencia, la principal contribución del modelado no es solo la predicción, sino también la identificación de factores accionables para orientar las decisiones didácticas y de evaluación.

Como limitaciones, los hallazgos deben interpretarse considerando que; la participación por país es desequilibrada, la unidad de análisis corresponde a los intentos en la plataforma, y no se dispone de variables adicionales como tiempo de respuesta, historial completo, condiciones de evaluación o variables académicas, que podrían enriquecer la explicación. Como futuras líneas de investigación, se recomienda validar el patrón en otras instituciones, incorporar modelos jerárquicos cuando sea posible y, sobre todo, realizar una fase aplicada donde los subtemas críticos identificados se traduzcan en una intervención breve y se evalúe su efecto sobre el error y la progresión del aprendizaje.

REFERENCIAS

[1] O. Viberg, M. Hatakka, O. Bälter, and A. Mavroudi, "The current landscape of learning analytics in higher education," *Comput. Human Behav.*, vol. 89, pp. 98–110, dec 2018, doi: 10.1016/J.CHB.2018.07.027.

[2] G. J. Hwang and Y. F. Tu, "Roles and research trends of artificial intelligence in mathematics education: A bibliometric mapping analysis and systematic review," *Mathematics*, vol. 9, no. 6, mar 2021, doi: 10.3390/MATH9060584.

[3] H. Khosravi *et al.*, "Explainable artificial intelligence in education," *Computers and Education: Artificial Intelligence*, vol. 3, p. 100074, jan 2022, doi: 10.1016/J.CAEAI.2022.100074.

[4] B. F. Azevedo, M. F. Pacheco, F. P. Fernandes, and A. I. Pereira, "Dataset of mathematics learning and assessment of higher education students using the mathe platform," *Data Brief*, vol. 53, p. 110236, apr 2024, doi: 10.1016/J.DIB.2024.110236.

[5] P. D. Long and G. Siemens, "Penetrare la nebbia: tecniche di analisi per l'apprendimento," *Revista Italiana de Tecnología Educativa*, vol. 22, no. 3, pp. 132–137, dec 2014, doi: 10.17471/2499-4324/195.

[6] D. J. Lemay, C. Baek, and T. Doleck, "Comparison of learning analytics and educational data mining: A topic modeling approach," *Computers and Education: Artificial Intelligence*, vol. 2, p. 100016, jan 2021, doi: 10.1016/J.CAEAI.2021.100016.

[7] A. Zeileis, "Examining exams using rasch models and assessment of measurement invariance," *Austrian Journal of Statistics*, vol. 54, no. 3, pp. 9–26, sep 2024, doi: 10.17713/ajs.v54i3.2055.

[8] Z. Zhang, "Model building strategy for logistic regression: Purposeful selection," *Ann. Transl. Med.*, vol. 4, no. 6, p. 111, mar 2016, doi: 10.21037/ATM.2016.02.15.

[9] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Adv. Neural Inf. Process. Syst.*, vol. 2017-December, pp. 4766–4775, may 2017, disponible en: https://arxiv.org/pdf/1705.07874. Accedido: Jan. 20, 2026.

[10] N. Bergdahl, M. Bond, J. Sjöberg, M. Dougherty, and E. Oxley, "Unpacking student engagement in higher education learning analytics: A systematic review," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, p. 63, dec 2024, doi: 10.1186/S41239-024-00493-Y/TABLES/6.

[11] D. Ifenthaler, J. Yau, and Y.-K. Yau, "Utilising learning analytics to support study success in higher education: A systematic review," *Educational Technology Research and Development*, vol. 68, no. 4, pp. 1961–1990, jun 2020, doi: 10.1007/S11423-020-09788-Z.

[12] L. Paulsen and E. Lindsay, "Learning analytics dashboards are increasingly becoming about learning and not just analytics – a systematic review," *Education and Information Technologies*, vol. 29, no. 11, pp. 14 279–14 308, jan 2024, doi: 10.1007/S10639-023-12401-4.

[13] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, vol. 13-17-August-2016, aug 2016, pp. 785–794, doi: 10.1145/2939672.2939785.

[14] S. M. Lundberg *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 56–67, jan 2020, doi: 10.1038/s42256-019-0138-9.

[15] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, jun 2020, doi: 10.1016/J.INFFUS.2019.12.012.