# ESCUELA POLITÉCNICA NACIONAL

## FACULTAD DE INGENIERÍA DE SISTEMAS

## DESARROLLO DE UN SIMULADOR DE CLASES PERSONALIZADAS CON IA GENERATIVA PARA EL APRENDIZAJE UNIVERSITARIO

## IMPLEMENTACIÓN DEL SISTEMA DE EVALUACIÓN ADAPTATIVA

**TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO DE SOFTWARE o EN CIENCIAS DE LA COMPUTACIÓN**

**CARLOS ANDRÉS CÓRDOVA ACARO**
**carlos.cordova02@epn.edu.ec**

**DIRECTOR: ENRIQUE ANDRÉS LARCO AMPUDIA, PhD.**
**andres.larco@epn.edu.ec**

**DMQ, enero 2026**

# CERTIFICACIONES

Yo, **Carlos Andrés Córdova Acaro**, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.
Asimismo, declaro que he utilizado herramientas de inteligencia artificial (ChatGPT, Gemini y Claude) únicamente como apoyo en la generación de tablas y como guía para la investigación de librerías en Python, sin atribuirles autoría, y que todo el contenido derivado ha sido revisado, validado y es de mi exclusiva responsabilidad.

---

**CARLOS ANDRÉS CÓRDOVA ACARO**

Certifico que el presente trabajo de integración curricular fue desarrollado por **Carlos Andrés Córdova Acaro**, bajo mi supervisión.

---

**ENRIQUE ANDRÉS LARCO AMPUDIA, PhD.**
Director

# DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmo que el trabajo de integración curricular aquí descrito, así como el producto resultante del mismo, es de carácter público y estará a disposición de la comunidad académica a través del repositorio institucional de la Escuela Politécnica Nacional.
No obstante, la titularidad de los derechos patrimoniales corresponde al autor del presente trabajo, de conformidad con las disposiciones establecidas por el órgano competente en materia de propiedad intelectual, la normativa interna institucional y la legislación vigente.

**Carlos Andrés Córdova Acaro**

**Enrique Andrés Larco Ampudia, PhD.**

# DEDICATORIA

Dedico este trabajo a mi familia, cuyo apoyo constante, esfuerzo y confianza han sido fundamentales a lo largo de mi formación académica y personal. Su acompañamiento incondicional ha sido un pilar esencial para alcanzar este logro.

# AGRADECIMIENTOS

Agradezco a la Escuela Politécnica Nacional por la formación académica brindada a lo largo de mi carrera universitaria. De manera especial, expreso mi gratitud al director de este trabajo, PhD. Enrique Andrés Larco Ampudia, por su orientación, observaciones y acompañamiento durante el desarrollo del presente proyecto.
Asimismo, agradezco a los docentes, compañeros y familiares que, de distintas maneras, contribuyeron con su apoyo y motivación a la culminación de esta etapa académica.

# ÍNDICE DE CONTENIDO

# ÍNDICE GENERAL

# RESUMEN

El presente trabajo desarrolla un sistema de evaluación adaptativa orientado al contexto universitario, cuyo objetivo es estimar de forma precisa el nivel de conocimiento del estudiante y personalizar la selección de actividades de aprendizaje. El sistema se fundamenta en un modelo híbrido que integra la Teoría de Respuesta al Ítem (IRT) con técnicas de Knowledge Tracing bayesiano, permitiendo un seguimiento continuo del estado de dominio del estudiante.

La arquitectura propuesta se implementa como un motor adaptativo desacoplado, capaz de interactuar con sistemas externos de generación de contenidos educativos. La validación del sistema se realiza mediante simulación computacional con estudiantes virtuales, analizando métricas de precisión, estabilidad y rendimiento. Los resultados obtenidos evidencian la viabilidad técnica del enfoque propuesto y su potencial aplicación en entornos de aprendizaje personalizados.

# ABSTRACT

This work presents the development of an adaptive assessment system designed for higher education environments. The main objective of the system is to accurately estimate students'knowledge levels and to personalize learning activities accordingly. The proposed approach is based on a hybrid model that combines Item Response Theory (IRT) with Bayesian Knowledge Tracing techniques.

The system is implemented as a decoupled adaptive engine capable of interacting with external content generation components. Validation is carried out through computational simulation using virtual students, evaluating accuracy, stability, and performance metrics. The results demonstrate the technical feasibility of the proposed model and its potential applicability in personalized learning systems.

# 1 DESCRIPCIÓN DEL COMPONENTE

## 1.1 Descripción general del componente

Este proyecto se compone de cuatro componentes interrelacionados. El Componente B constituye el motor que traduce la evidencia de interacción del estudiante en decisiones pedagógicas accionables. Su función principal es calcular el nivel de habilidad del estudiante, identificar qué conocimientos domina con suficiente confianza y determinar qué actividad o ítem presentar a continuación, incluyendo la dificultad apropiada y el tipo de soporte necesario. Los objetivos del componente son duales: por un lado, medir con precisión el desempeño del estudiante; por otro, favorecer el aprendizaje mediante trayectorias personalizadas, haciendo uso de la retroalimentación constante que mantiene con el Componente A, responsable de generar las actividades y las clases. Este enfoque se alinea con la lógica del aprendizaje adaptativo que predomina en la actualidad, basada en realizar ajustes del proceso de aprendizaje a partir de datos individuales en lugar de seguir rutas fijas predeterminadas [1], [2], [3].

## 1.2 Objetivo general

Desarrollar un motor adaptativo operativo que traduzca la evidencia de interacción del estudiante en decisiones pedagógicas fundamentadas, mediante la combinación de modelos psicométricos extensivamente validados (IRT y BKT/KT) para medir con precisión el desempeño del estudiante y construir trayectorias de aprendizaje genuinamente personalizadas que se ajusten dinámicamente a partir de datos individuales.

## 1.3 Objetivos específicos

### 1.3.1 Implementar un sistema de medición dual del perfil del estudiante mediante señales complementarias

Desarrollar un sistema que recoja eventos de interacción y calcule dos señales complementarias: un nivel continuo de habilidad $\theta$ basado en IRT para ordenar ítems informativos y reducir el error estándar de medición, y una probabilidad de dominio por habilidad fundamentada en BKT/KT para guiar la práctica espaciada y el refuerzo cuando el objetivo sea consolidar conocimientos de forma sostenida.

### 1.3.2 Desarrollar una política de selección adaptativa basada en IRT para contextos de diagnóstico y certificación

Implementar una política de selección que escoja el ítem que maximiza la información en torno al $\theta$ estimado, reduciendo rápidamente el error estándar y permitiendo alcanzar precisión localizada donde más importa, incorporando reglas de detención, restricciones de contenido curricular y limitaciones de exposición siguiendo las buenas prácticas establecidas en IRT y CAT.

### 1.3.3 Implementar un sistema de rastreo de conocimiento basado en BKT/KT para orientar la práctica guiada

Desarrollar mediante BKT/KT un sistema que mantenga una probabilidad de dominio por habilidad y la actualice tras cada interacción, modelando fenómenos como la adivinación y el desliz, de modo que la selección de la actividad subsiguiente se decida según el beneficio esperado: confirmar dominio incipiente, reducir incertidumbre o fomentar el aprendizaje en la zona de desarrollo más productiva.

### 1.3.4 Diseñar una arquitectura fundamentada en el patrón MAPE-K para sistemas auto-adaptativos

Implementar el Componente B como un bucle MAPE-K que monitorice respuestas y patrones de desempeño, analice el perfil del estudiante, planifique la siguiente actividad especificando ítem, dificultad y tipo de apoyo, y ejecute enviando recomendaciones explícitas al Componente A, garantizando trazabilidad, explicabilidad y la posibilidad de integrar organización semántica del contenido mediante grafos o rutas de aprendizaje.

### 1.3.5 Establecer un contrato explícito de integración entre componentes que garantice trazabilidad

Diseñar la integración $A \leftrightarrow B$ como un contrato de datos simple y explícito donde el Componente A envíe datos del usuario, tema activo e historial de interacciones, y el Componente B responda con el ítem propuesto, dificultad objetivo, tipo de ayuda recomendada y justificación concisa de la decisión, haciendo las decisiones auditables y proporcionando al docente evidencia clara del proceso.

### 1.3.6 Implementar salvaguardas de validez, equidad y mecanismos de monitoreo continuo del sistema

Desarrollar salvaguardas que resguarden la validez y equidad mediante el reporte de precisión alcanzada, validación del ajuste del modelo, monitorización de exposición equilibrada de temas e ítems, reporte de ganancia pre-post, control de métricas predictivas como AUC o log-loss, y realización de pruebas DIF y análisis de brechas entre perfiles para detectar posibles sesgos.

## 1.4 Alcance del componente

El alcance del Componente B comprende el desarrollo de un motor adaptativo operativo con las siguientes características funcionales y técnicas:

I. **Integración de modelos complementarios:** Combinar IRT para diagnóstico o evaluación sumativa con BKT/KT para guiar la práctica continua, aprovechando las fortalezas de ambos enfoques para ofrecer una evaluación integral que mida y fomente el aprendizaje.

II. **Orquestación de la selección de ítems:** Orquestar la selección de ítems mediante reglas claras de parada, cobertura curricular y exposición equilibrada que permitan determinar cuándo la evaluación ha alcanzado suficiente precisión, garantizando que todos los temas relevantes sean cubiertos y evitando la sobreutilización de ítems específicos.

III. **Panel de métricas para validación docente:** Publicar un panel de métricas (precisión diagnóstica, eficiencia, progreso del estudiante, calidad predictiva y equidad) accesible para que los docentes validen el funcionamiento del sistema, facilitando la transparencia y permitiendo intervenciones informadas cuando resulte necesario.

IV. **Gobernanza de datos y privacidad:** Documentar la gobernanza de datos y privacidad, especificando qué se registra, para qué propósito y cómo se protege la información, asegurando el cumplimiento de estándares éticos y regulatorios en el manejo de datos educativos sensibles.

Todo ello integrado con el Componente A de manera que cada estudiante reciba un reto apropiado, en el momento oportuno, con las explicaciones y los apoyos adecuados a su nivel y necesidades específicas, logrando una experiencia de aprendizaje genuinamente personalizada y fundamentada en evidencia [1], [3], [4], [5], [6].

## 1.5 Marco teórico

### 1.5.1 Aprendizaje adaptativo

El aprendizaje adaptativo es una forma de enseñanza que se ajusta a cada estudiante en tiempo real. En vez de proponer la misma ruta para todos, el sistema observa evidencias (aciertos, errores, tiempo de respuesta, interacciones) y decide qué contenido, qué nivel de dificultad y qué apoyo conviene a continuación. Así, la progresión deja de ser lineal y se vuelve personalizada, manteniendo el foco en el dominio gradual de objetivos. Esta idea se formaliza y se sostiene en la literatura reciente sobre evaluación, personalización y uso responsable de IA en educación [1], [3].

Conviene distinguir lo adaptativo de lo simplemente ”personalizado”. La personalización puede implicar variedad de actividades o estilos, pero no siempre supone que el sistema mida y ajuste continuamente con base en datos. Lo adaptativo, en cambio, depende de un ciclo continuo de diagnóstico-retroalimentación-reajuste, y se apoya en un buen diseño instruccional: objetivos claros, progresiones definidas y evidencias útiles para decidir los siguientes pasos [3].

En la práctica, el ciclo luce así: (1) un breve diagnóstico, (2) selección de recursos y tareas ajustadas al nivel detectado, (3) retroalimentación oportuna, (4) una nueva medición que confirma avances o sugiere refuerzos y (5) ajustes de la ruta. La clave no es aumentar la cantidad de ejercicios, sino ofrecer los adecuados en el momento preciso. Este principio didáctico se alinea con marcos como los de Reigeluth y los primeros principios de Merrill, que recomiendan activar saberes previos, demostrar, aplicar e integrar lo aprendido [3].

La IA generativa ha aportado un motor útil para redactar explicaciones, proponer ejemplos y crear ejercicios alineados con la ruta de cada estudiante. Sin embargo, la evidencia disponible también advierte que estas herramientas no reemplazan la pedagogía ni la evaluación rigurosa; su valor aumenta cuando operan bajo criterios claros de calidad, ética y supervisión docente [1].

Un ejemplo concreto es PathRAG, que organiza el conocimiento como un grafo (conceptos y relaciones) para trazar caminos pertinentes según el perfil del estudiante. Estudios recientes en contextos universitarios híbridos reportan mejoras en participación, logro de competencias y percepción de inclusión cuando se integran rutas personalizadas con apoyo de IA generativa; aun así, subrayan límites metodológicos y la necesidad de diseños más robustos [2].

### 1.5.2 Teoría de Respuesta al Ítem (IRT)

La Teoría de Respuesta al Ítem (TRI o IRT) es una forma moderna de entender las pruebas: en lugar de mirar solo el puntaje total, analiza cómo responde una persona a cada ítem y, a partir de ello, estima su nivel en el rasgo que se quiere medir $\theta$. Con esa estimación, es posible seleccionar mejores preguntas, ubicar la dificultad donde más hace falta y conocer cuán precisa es la medición en cada tramo del continuo. Frente a la Teoría Clásica de Tests,

su aporte central es la ínvariancia': medir con la misma escala, aunque cambien los sujetos o los ítems (dentro de ciertos supuestos) [1], [2].

#### 1.5.2.1 Ideas clave

- **Curva característica del ítem (CCI):** es un gráfico que muestra, para cada nivel de $\theta$, la probabilidad de elegir la opción "clave" del ítem (por ejemplo, responder correctamente o indicar mayor rasgo). Su forma creciente refleja que, a mayor nivel del rasgo, mayor probabilidad de dar la respuesta asociada al rasgo [1], [2].
- **Parámetros $a$, $b$ y $c$:** $a$ indica cuánto discrimina el ítem (qué tan bien separa a personas con niveles cercanos de $\theta$), $b$ ubica la dificultad del ítem (el punto de la escala donde el ítem decide), y $c$ modela el azar o pseudo-adivinación en ítems de opción correcta/incorrecta. No todos los modelos usan los tres: Rasch (1PL) usa solo b, 2PL usa a y b, y 3PL usa a, b, c [1], [2].
- **Información del ítem/test:** indica dónde el ítem o el conjunto de ítems mide con mayor precisión. En IRT la precisión no es plana: puede ser excelente en un rango de $\theta$ y más baja en otros. Esto permite construir bancos de ítems que cubran la escala con precisión donde más importa [2].

#### 1.5.2.2 Supuestos relevantes

- **Unidimensionalidad:** los ítems de una escala deben reflejar esencialmente un solo rasgo dominante; si influyen varios rasgos a la vez, conviene usar modelos multidimensionales o depurar la escala [2].
- **Independencia local:** si ya sabemos el nivel de $\theta$, las respuestas a ítems distintos no deben depender entre sí. Cuando hay pistas entre ítems o se agrupan demasiado, este supuesto se rompe y la medición pierde calidad [1], [2].

#### 1.5.2.3 Modelos más usados (visión práctica)

- **Ítems dicotómicos (correcto/incorrecto):** 1PL (Rasch), 2PL y 3PL. Rasch asume igual discriminación y sin azar; 2PL permite que la discriminación varíe; 3PL incluye el parámetro de pseudo-adivinación. Elegir el modelo depende del contexto y los datos [1], [2].
- **Ítems politómicos (escalas Likert):** Modelos como Respuesta Graduada (Samejima) o Crédito Parcial. En el Modelo de Respuesta Graduada, cada salto entre categorías tiene un umbral de dificultad, y un único parámetro $a$ de discriminación para el ítem. Esto es muy útil para cuestionarios con varias opciones de respuesta [7].

#### 1.5.2.4 Bancos de ítems y pruebas adaptativas (CAT)

Al estimar $\theta$ en tiempo real y conocer la información de cada ítem, es posible elegir la siguiente pregunta que aporte máxima precisión justo alrededor del nivel estimado del estudiante. Así nacen los tests adaptativos: cada persona responde un conjunto distinto de preguntas, pero todos son evaluados en la misma escala. En tu proyecto, esto es clave para que el Componente B seleccione o recomiende ítems con mayor ”ganancia informativa” [7], [8].

#### 1.5.2.5 Integración con los Componentes A y B

1. (B) A partir de las respuestas del estudiante, se estima $\theta$ con un modelo IRT apropiado (2PL/3PL para ítems dicotómicos; Respuesta Graduada para Likert).
2. (B) se consulta el banco de ítems para identificar cuáles ofrecen más información alrededor del $\theta$ actual (o del umbral de dominio).
3. (B$\rightarrow$A) se envía al Componente A la dificultad objetivo y, si aplica, los ítems recomendados o las pautas de complejidad.
4. (A) el Componente A genera la siguiente actividad con esa dificultad y pistas adecuadas.
5. (B) tras la actividad, se actualiza $\theta$ y se repite el ciclo. Este bucle mantiene rutas personalizadas y medibles [7], [9].

### 1.5.3 Sistemas de Tutoría Inteligente (ITS)

Un Sistema de Tutoría Inteligente (ITS) es un software que intenta parecerse a una tutoría humana: observa cómo aprende el estudiante, le ofrece explicaciones y actividades a la medida, y retroalimenta en los momentos clave. La idea no es reemplazar al docente, sino multiplicar su apoyo para que cada persona avance a su propio ritmo y con la ayuda justa. Las revisiones recientes muestran que, bien implementados, los ITS mejoran el rendimiento y la participación, personalizan contenidos y apoyan la autorregulación del aprendizaje [4].

#### 1.5.3.1 Componentes típicos

- **Modelo del estudiante:** mantiene un ”perfil vivo” con aciertos, errores, tiempos y progreso.
- **Modelo del tutor:** decide qué explicar, qué actividad proponer y qué pista dar.
- **Modelo de dominio:** representa el conocimiento de la materia (conceptos, habilidades, reglas).
- **Interfaz:** es la cara del sistema (pantallas, ejercicios, feedback).

Con estos componentes, el ITS puede ajustar dificultad, secuencias y apoyos en tiempo real [5].

#### 1.5.3.2 Integración con los Componentes A y B

1. (B) observa respuestas, estima el nivel del estudiante en los temas clave y detecta dificultades.
2. (B→A) Envía al Componente A la dificultad recomendada, objetivos prioritarios y el tipo de intervención.
3. (A) El Componente A genera la actividad/clase con esa dificultad y apoyo.
4. (B) tras la actividad, el ITS vuelve a medir y ajusta la ruta. Este bucle mantiene trayectorias personalizadas y medibles a lo largo del curso [4], [5].

### 1.5.4 Algoritmos de selección adaptativa

Seleccionar ”lo siguiente” no es al azar: es decidir, con evidencia, cuál actividad o ítem conviene presentar para medir mejor o para ayudar a aprender mejor. En este proyecto, esa decisión vive en el Componente B (evaluación) y retroalimenta al Componente A (generación de clases). A grandes rasgos, hay dos familias bien establecidas: selección guiada por IRT (cuando buscamos medir con precisión) y selección guiada por BKT/KT (cuando buscamos acompañar la adquisición de habilidades en el tiempo).

#### 1.5.4.1 Selección guiada por IRT (medición precisa)

La Teoría de Respuesta al Ítem (IRT) modela la probabilidad de respuesta correcta según el nivel del rasgo latente $\theta$ y los parámetros del ítem. Con esa base, la selección adaptativa típica elige el siguiente ítem con más información alrededor del $\theta$ estimado del estudiante. Esto reduce el error estándar de medición con menos preguntas y mantiene la dificultad ”justo donde más informa”. En la práctica, se inicia con un $\theta$ neutro o con un breve arranque *warm-up*; y tras cada respuesta se reestima $\theta$ y se elige el ítem que maximiza la información (o criterios cercanos como la información de Fisher o la divergencia KL). Para mantener validez y equidad, se aplican restricciones de contenido (temas/objetivos), control de exposición (evitar sobreuso de ciertos ítems) y límites de longitud o precisión objetivo. En escalas politómicas (tipo Likert), la lógica es análoga: cada categoría aporta información en zonas distintas de la escala, y la selección prioriza donde la precisión es más útil [7], [8].
Qué aporta al Componente A↔B cuando el objetivo es certificar dominio o ubicar con exactitud el nivel, IRT permite pedir menos y medir mejor. El Componente B devuelve a A el rango de dificultad recomendada (y la cobertura temática pendiente), de modo que A genere actividades acordes a ese nivel y no ”sobre-o-subestime” la exigencia [7], [8].

#### 1.5.4.2 Selección guiada por BKT/KT (apoyo al aprendizaje)

Bayesian Knowledge Tracing (BKT) sigue, para cada habilidad, la probabilidad de dominio del estudiante a lo largo del tiempo: considera un estado ”domina / no domina” y cuatro parámetros intuitivos (conocimiento inicial, probabilidad de aprender tras una práctica, adivinación y desliz). Con ese perfil, el sistema decide el siguiente ejercicio según el mayor beneficio esperado: reducir la incertidumbre, confirmar dominio o provocar aprendizaje en la zona adecuada. En contextos reales se combinan además prerequisitos, espaciado para combatir el olvido y señales de compromiso (tiempos, rachas). La evidencia reciente muestra que BKT es eficaz para personalizar secuencias y mejorar resultados cuando la meta es progresar en habilidades específicas y no sólo ”medir una vez con precisión” [10], [11].
El aporte dentro del flujo entre los Componentes A↔B, BKT entrega al Componente A no sólo una dificultad recomendada, sino también la habilidad prioritaria, el tipo de apoyo (pista, ejemplo guiado, práctica adicional) y el momento oportuno para espaciado o refuerzo. Tras la actividad de A, B actualiza las probabilidades de dominio y repite el ciclo [6], [11].

#### 1.5.4.3 Estrategia híbrida

En etapas tempranas o con bancos pequeños, BKT tiende a funcionar mejor porque necesita menos calibración de ítems y entrega señales útiles para enseñar. Cuando el banco crece y se busca una estimación fina del nivel, IRT gana relevancia: permite fijar una precisión objetivo y optimizar la ruta de ítems. Una política práctica es: usar BKT para guiar la práctica diaria (progreso por habilidades) y activar selección IRT en cortes de evaluación (diagnósticos o certificaciones). En la arquitectura del proyecto, esto se implementa como un bucle MAPEK: Monitorizar (respuestas), Analizar (IRT/BKT), Planificar (siguiente ítem o actividad) y Ejecutar (enviar a A), con conocimiento compartido del perfil del estudiante y del banco de ítems [6].

### 1.5.5 Métricas de desempeño en sistemas adaptativos

Las métricas no son un listado de números: son la forma en que demostramos que el sistema realmente ayuda a aprender y que lo hace de manera eficiente y justa. En un entorno adaptativo, medir implica dos planos que se retroalimentan: la calidad de la medición (£qué tan bien estimamos el nivel del estudiante?) y la calidad de la enseñanza (£qué tanto aprende y con qué esfuerzo?). A continuación se presentan métricas nucleares, escritas en lenguaje claro, conectando la literatura de pruebas adaptativas y trazado del conocimiento [8], [12].

#### 1.5.5.1 Precisión y validez de medición (IRT/CAT)

En pruebas adaptativas orientadas a medir con exactitud, la precisión se observa en el error estándar del estimador de habilidad, $SE(\hat{\theta})$, y en la información del test alrededor

del nivel estimado. Un buen algoritmo reduce $SE(\hat{\theta})$ con menos ítems: ese equilibrio entre precisión y longitud del test es central. Para comparar métodos, es útil fijar la eficiencia (mismo número medio de ítems) y contrastar la precisión resultante, o al revés. En estudios recientes se equiparan ambos métodos (por ejemplo, IRTCAT vs. enfoques alternativos) y se evalúa la diferencia media y la correlación de los puntajes con respecto a un "test completo" considerado referencia [12].
Otro indicador clave es la calibración/ajuste del modelo: el sesgo (diferencia promedio entre el estimado y el valor de referencia), la RMSE (raíz del error cuadrático medio) y las curvas de calibración por tramos de la escala. Para motores más flexibles, se recurre a medidas de divergencia como $KL(\pi||\pi)$, que cuantifican la pérdida de información entre la densidad verdadera de puntajes y la estimada; valores pequeños señalan mejor ajuste. En simulaciones de calibración y selección de modelo, la elección del criterio de información (p. ej., BIC) afecta de forma tangible la precisión final de las estimaciones [12].

#### 1.5.5.2 Eficiencia y carga de respuesta

La eficiencia refleja cuántos ítems o cuánto tiempo necesita el sistema para alcanzar una precisión aceptable. Métricas prácticas incluyen: número promedio de ítems administrados, desviación típica de ese número (variabilidad de carga entre estudiantes), tiempo por objetivo alcanzado y porcentaje de ítems no administrados (ahorro respecto del banco total). Un buen sistema es más corto sin sacrificar precisión. Además, conviene analizar cómo varía la carga según el nivel verdadero: algunos algoritmos exigen más ítems en los extremos de la escala, otros en el centro; reconocer ese patrón ayuda a planificar bancos y reglas de parada [12].

#### 1.5.5.3 Aprendizaje y progreso

Cuando el objetivo es que el estudiante aprenda (no solo medir), importan indicadores de progreso: la ganancia entre pre- y post-prueba normalizada por la dificultad, la tasa de dominio por unidad, el tiempo/ítems hasta alcanzar un umbral de dominio, y la retención tras un intervalo (espaciado). En sistemas con trazado del conocimiento (KT/BKT), puede reportarse la probabilidad de dominio por habilidad y su evolución, verificando que las decisiones (refuerzo, explicación, práctica guiada) incrementen esa probabilidad de forma sostenida [8], [10].

#### 1.5.5.4 Calidad predictiva de la política adaptativa

Para validar que el sistema "elige bien lo siguiente", se evalúa su poder de predicción de respuestas y de dominio futuro. Métricas comunes son log-loss (pérdida de probabilidad), AUC/ROC para acierto de la próxima respuesta y exactitud o F1 en clasificación de dominio/no-dominio. En secuenciación adaptativa, también se puede medir el beneficio esperado o regret acumulado frente a una política de referencia. Estas métricas no sustituyen

a la evidencia de aprendizaje, pero aseguran que el motor de decisión es consistente y estable [10], [11].

#### 1.5.5.5 Equidad y robustez

Un sistema adaptativo debe ser justo y estable. La equidad se estudia con análisis DIF (ítems que favorecen a subgrupos), comparaciones de error/precisión y tasa de dominio entre perfiles, y auditorías de exposición de ítems. La robustez exige pruebas de sensibilidad a supuestos del modelo (unidimensionalidad, independencia local) y validación cruzada cuando se recalibra el banco. Por último, la transparencia en el uso de datos y la interpretabilidad de reportes para docentes son métricas de calidad percibida y confianza [7], [12].

#### 1.5.5.6 Reporte para la integración A$\leftrightarrow$B

Para cerrar el ciclo A$\leftrightarrow$B, el Componente B debe devolver un panel compacto: (i) precisión alcanzada ($SE(\hat{\theta})$,o intervalo de confianza de puntaje/total), (ii) eficiencia (ítems/tiempo vs. objetivo), (iii) progreso por habilidad (probabilidad de dominio y ganancia), (iv) calidad de predicción (AUC/logloss) y (v) equidad (DIF y exposición balanceada). Con ese resumen, el Componente A puede ajustar dificultad, apoyo y espaciado con criterio.

# 2 METODOLOGÍA

## 2.1 Enfoque y diseño de la investigación

En el desarrollo del Sistema de Evaluación Adaptativa (Componente B) se utilizó un enfoque de investigación cuantitativa en el que la evaluación era objetiva, numérica, reproducible y medible de las variables pedagógicas y computacionales. Este enfoque era lógico ya que se relaciona con el tipo de problema abordado, es decir, optimizar procesos de evaluación y validar un sistema de software fundamentado en modelos matemáticos, estadísticos y probabilísticos que centra la evaluación en niveles profundos del conocimiento del estudiante. El método cuantitativo permite la evaluación de la forma de actuar del motor adaptativo mediante la valoración de indicadores con los que se puede medir, como pueden ser la estimación de la habilidad latente del aprendiz ($\theta$), el ajuste en base a la precisión de las métricas como el error cuadrático medio (RMSE), la fiabilidad de las probabilidades analizada con la métrica de Brier Score; además tomando métricas concretas de la ingeniería de ciencias computacionales como la latencia de la respuesta del sistema, percentiles de tiempo de procesamiento (P50 y P95), y la tasa de peticiones por segundo (RPS). Estas métricas sirven para el diagnóstico en base a criterios cuantificables de la precisión, la eficiencia y la escalabilidad del sistema propuesto. La utilización de esta vertiente metodológica se apoya en la documentación especializada de los sistemas de aprendizaje adaptativo y la evaluación psicométrica, la cual establece que para la obtención de indicadores robustos de aprendizaje deben emplearse modelos estadísticos que permiten inferir variables latentes a partir de la evidencia empírica observable, particularmente en el caso de la Teoría de Respuesta al Ítem (IRT) y en los modelos bayesianos de rastreo de conocimiento [7], [8], [11], [12].

### 2.1.1 Clasificación de la investigación

Tomando en cuenta el marco metodológico que se ha seguido, esta investigación puede ser clasificada como una investigación tecnológica aplicada, la cual se halla orientada al diseño, a la validación y a la implementación de un artefacto de software funcional y operativo que tiene la finalidad de proporcionar una solución a un problema de práctica educativa en contextos locales, específicamente, la modulación adaptativa de evaluaciones mediante la integración de modelos de Machine Learning en la educación superior [4], [6].

### 2.1.2 Arquitectura experimental y estrategia de simulación

La arquitectura experimental se apoya en la simulación computacional, pues las limitaciones logísticas, éticas y operativas de llevar a cabo un elevado número de pruebas con estudiantes reales en una fase incipiente del desarrollo nos llevaron a adoptar esta opción metodológica. En este contexto, la simulación estocástica y los métodos Monte Carlo conforman una opción metodológica argumentada teóricamente, pero también muy extendida y validada en la literatura para evaluar sistemas adaptativos complejos [3], [6], [10].

Este modelo propone la construcción de un entorno de simulación en el que fueron modelados perfiles de estudiantes virtuales con ciertos parámetros psicométricos controlados, entre los que podemos encontrar el nivel de habilidad inicial ($\theta$), la consistencia de respuesta ante ítems de dificultad variable y el ratio de aprendizaje. Este entorno permite generar un elevado número de interacciones simuladas entre el sistema y perfiles de estudiantes heterogéneos, lo que permite estudiar la convergencia del algoritmo adaptativo, su comportamiento bajo distintas condiciones operativas y evaluar la robustez frente a situaciones adversas o excepcionales. De igual manera, la simulación computacional favoreció la validación operativa del sistema en situaciones de baja probabilidad de ocurrencia o difícilmente reproducibles en contextos reales de aplicación, tales como patrones de respuestas erráticas por parte de los estudiantes, ejecución de múltiples sesiones de evaluación de manera concurrente, y escenarios de escasez de ítems calibrados en el banco de preguntas. Todo ello contribuyó a proporcionar consistencia empírica a la evaluación de la tolerancia a fallos y la robustez estructural del motor adaptativo. Finalmente, el diseño metodológico propuesto permite establecer que esta fase corresponde a una validación algorítmica y técnica del sistema en condiciones controladas. La arquitectura del estudio contempla la ejecución de pruebas con usuarios reales para una fase posterior del proceso de investigación, una fase que estará orientada al análisis del impacto pedagógico efectivo del sistema y al estudio de factores cualitativos emergentes en contextos auténticos de aprendizaje.

En el siguiente apartado se presenta la relación detallada de las variables e indicadores de validación en la Tabla 2.1, la cual recoge la síntesis estructurada de los criterios cuantitativos que se han establecido para la evaluación sistemática del desempeño del motor adaptativo.

| Variable operacionalizada | Descripción conceptual | Función en el proceso de validación |
|---|---|---|
| $\theta$ (parámetro de habilidad estimado) | Inferencia del nivel latente de dominio cognitivo del estudiante | Evaluar la precisión diagnóstica del modelo psicométrico |
| RMSE / MAE | Cuantificación del error entre el parámetro real de habilidad y su estimación computacional | Medir exactitud predictiva del sistema implementado |
| Brier Score | Función de pérdida cuadrática aplicada a probabilidades predichas | Evaluar calidad y calibración de las predicciones probabilísticas |
| Latencia (ms) | Duración temporal del procesamiento de peticiones del sistema | Analizar rendimiento computacional y eficiencia algorítmica |
| RPS | Volumen de peticiones procesadas exitosamente por unidad temporal | Evaluar escalabilidad horizontal y capacidad de concurrencia |
| P50 / P95 | Percentiles de la distribución de tiempos de respuesta | Detectar degradación del rendimiento bajo condiciones de carga elevada |

Cuadro 2.1: Criterios cuantitativos utilizados para la validación del motor adaptativo

## 2.2 Tipo y diseño de la investigación

La investigación desarrollada se enmarca dentro del ámbito de la investigación tecnológica aplicada en la Ingeniería en Ciencias de la Computación. Esta clasificación responde a que el propósito central del trabajo no es la formulación de teorías abstractas, sino el diseño, la implementación y la validación de un artefacto computacional operativo, concretamente un Sistema de Evaluación Adaptativa orientado a la personalización del aprendizaje mediante el uso de modelos psicométricos y técnicas de aprendizaje automático. Este tipo de investigación tecnológica aplicada se halla caracterizada por la producción de conocimiento a partir de la construcción y la evaluación sistemática de soluciones software que logran dar respuesta a problemas reales, sin perder los principios de verificabilidad, reproducibilidad y rigor experimental que caracterizan a la ingeniería de software. Por tanto, el valor científico se encuentra tanto en la arquitectura del sistema como en las pruebas empíricas recogidas durante el proceso de validación. Desde esta óptica, diferentes trabajos en el campo del aprendizaje adaptativo y los sistemas de tutoría inteligente establecen que la evaluación de este tipo de sistemas debe fundamentarse en medidas cuantitativas objetivas (precisión diagnóstica, eficiencia algorítmica, calidad predictiva, entre otras) y no en aproximaciones meramente descriptivas [4], [5], [6]. Esta línea de trabajo refuerza la idoneidad del tipo de investigación escogida.

### 2.2.1 Diseño experimental basado en simulación

En relación con el diseño de la investigación, se optó por un diseño experimental ya que la investigación supone la manipulación controlada de variables independientes y la observación sistemática de sus efectos sobre variables dependientes relacionadas con el rendimiento del sistema. Las variables manipuladas incluyen el nivel de habilidad previo del estudiante, la consistencia en las respuestas, la dificultad de los ítems y la concurrencia de usuarios; mientras que las variables observadas son la convergencia de la estimación de habilidad, el error de medición, la calidad predictiva del modelo y el rendimiento computacional del sistema. El diseño experimental se implementó a través de simulación computacional, una técnica ampliamente empleada en investigaciones vinculadas con la ingeniería de software, los sistemas autoadaptativos y el aprendizaje adaptativo, particularmente en contextos donde la experimentación directa con usuarios reales se encuentra limitada por consideraciones éticas, logísticas o temporales [3], [6], [10]. La simulación posibilita la reproducción de escenarios complejos bajo condiciones controladas, lo que contribuye a fortalecer la validez interna del estudio y a garantizar la reproducibilidad de los experimentos. Con este fin, se desarrolló un simulador de estudiantes virtuales capaz de generar interacciones estocásticas con el sistema de evaluación adaptativa, siguiendo un enfoque de tipo Monte Carlo. Cada estudiante virtual fue modelado a partir de parámetros psicométricos previamente definidos, tales como la habilidad latente inicial ($\theta$), la probabilidad de dominio asociada a cada habilidad y la consistencia en las respuestas. Este planteamiento permite analizar el comportamiento del sistema frente a una amplia diversidad de perfiles de aprendizaje. Este enfoque ha sido ampliamente empleado para analizar la estabilidad, la equidad diagnóstica y la eficiencia de algoritmos de secuenciación adaptativa y de rastreo del conocimiento [10], [11].

### 2.2.2 Análisis transversal y longitudinal

El diseño experimental por simulación permitió llevar a cabo experimentos de tipo transversal y longitudinal. En el análisis transversal se observa la respuesta inmediata del sistema ante distintos perfiles de estudiantes, mientras que el análisis longitudinal permite simular la evolución temporal del aprendizaje considerando factores tales como la consolidación del conocimiento adquirido o el decaimiento progresivo del mismo. Este tipo de análisis es crítico para los sistemas de evaluación adaptativa porque permite evaluar la capacidad del modelo para detectar la pérdida gradual de dominio en las habilidades y recomendar intervenciones oportunas, tal como subrayan las publicaciones científicas relacionadas [4], [11]. El diseño también permitió evaluar el sistema en situaciones extremas poco reproducibles en ambientes educativos reales, como patrones de respuestas erráticas, escasez de ítems calibrados disponibles o ejecución simultánea de un elevado número de sesiones de evaluación concurrentes. La inclusión de estas pruebas contribuyó al análisis de la robustez, tolerancia a fallos y escalabilidad del motor adaptativo previo a su implementación en contextos académicos reales. La Tabla 2.2 presenta un resumen de los componentes centrales

del tipo y diseño de investigación adoptados junto con la justificación técnica y metodológica correspondiente.

| Elemento metodológico | Clasificación adoptada | Justificación técnica |
|---|---|---|
| Tipo de investigación | Tecnológica aplicada | Desarrollo y validación de un sistema software funcional |
| Diseño de investigación | Experimental | Manipulación controlada de variables y medición de efectos |
| Estrategia experimental | Simulación computacional | Reproducibilidad y control de escenarios complejos |
| Técnica de simulación | Monte Carlo | Evaluación estocástica de múltiples perfiles de estudiantes |
| Horizonte de análisis | Transversal y longitudinal | Evaluación inmediata y análisis temporal del aprendizaje |

Cuadro 2.2: Clasificación metodológica y justificación técnica del estudio

## 2.3 Método de investigación

Para llevar a cabo el desarrollo y validación del Sistema de Evaluación Adaptativa (Componente B) se utilizó el método de investigación hipotéticodeductivo, el cual resulta ser uno de los más extendidos en los ámbitos de investigación en ingeniería y ciencias de la computación cuando se desea analizar y validar la forma de comportamiento del sistema en base a una serie de supuestos teóricos formalizados. Así, resulta altamente consistente el uso de este método de investigación para el estudio de sistemas adaptativos en los que la parte de diseño algorítmico fue formulada a partir de modelos matemáticos y de modelos probabilísticos de los que se deduce la necesidad de contrastar empíricamente su validez mediante el método de experimentación controlada y reproducible. El método hipotéticodeductivo se define por ser un procedimiento que parte de la observación sistemática a partir de un problema, la formulación de hipótesis explicativas, la deducción de las consecuencias observables y la posterior comprobación experimental de éstas. Dentro de esta investigación, dicho enfoque hizo posible estructurar el desarrollo del motor adaptativo como proceso lógico y secuencial de forma que la teoría psicométrica subyacente, las decisiones de diseño de representación algorítmica y los resultados hallados durante el curso de validación [4], [6], [8] tuvieran coherencia.

### 2.3.1 Fase de observación y problematización

A través de la etapa de observación y problematización se identificaron las limitaciones recurrentes de los sistemas de evaluación tradicionales que se caracterizan por secuencias de ítems estáticos, criterios de calificación pragmáticos y escasa capacidad de adaptación a lo que realmente sabe el estudiante. La literatura especializada en aprendizaje adaptativo y sistemas de tutoría inteligente señala precisamente que estos sistemas suelen dar lugar

a evaluaciones ineficaces y diagnósticos imprecisos en contextos educativos con alta heterogeneidad en los perfiles de aprendizaje [4], [5]. Los resultados de este análisis dieron pie a la formulación de la hipótesis central de la investigación, consistente en que la integración de un modelo híbrido que combine la Teoría de Respuesta al Ítem (IRT) para obtener una estimación global de la habilidad latente del estudiante y los modelos bayesianos de rastreo de conocimiento para obtener un monitoreo más granular de las habilidades permite conseguir una mejora notable en la eficiencia, la precisión y la equidad diagnóstica de la evaluación frente a los métodos lineales o no adaptativos. Esta hipótesis se apoya en trabajos anteriores que advierten sobre la complementariedad en la combinación de los modelos psicométricos globales y de técnicas de rastreo probabilísticas del aprendizaje a nivel de habilidad [8], [10], [11].

### 2.3.2 Fase de deducción

En la fase de deducción, la hipótesis propuesta se tradujo en un conjunto de decisiones de diseño dirigido a orientar la implementación del motor adaptativo. Concretamente, se decidió utilizar el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem para estimar la habilidad latente, aplicando el método de estimación a posteriori esperada (EAP) con el objetivo de garantizar la estabilidad numérica y disminuir los sesgos en situaciones de información escasa. A su vez, se propuso un modelo bayesiano de rastreo de conocimiento con decaimiento temporal, cuyo objetivo es modelar la probabilidad de dominio de cada habilidad y la posibilidad de una pérdida progresiva de la misma en el tiempo. De estas decisiones se dedujeron una serie de consecuencias observables que podían ser evaluadas empíricamente, tales como:

- La convergencia progresiva de la estimación de la habilidad del estudiante.
- El error estándar de medición en decremento a medida que se administraran ítems informativos.
- La detección temprana de las brechas de conocimiento.
- La adaptación dinámica de ítems y actividades propuestas.

Se dedujo que el sistema debería conseguir niveles aceptables de precisión diagnóstica con un número reducido de ítems, a la vez que un adecuado rendimiento de cómputo en condiciones de concurrencia.

### 2.3.3 Fase de verificación experimental

La fase de verificación experimental se llevó a cabo mediante la administración de baterías de pruebas automatizadas y experimentos controlados fundamentados en simulación computacional. En esta etapa fue posible contrastar la lógica deducida de la propia hipótesis respecto de lo que sucedía en la realidad del comportamiento de la propuesta de trabajo. Los experimentos realizados incluyeron pruebas de convergencia de la habilidad

estimada, evaluaciones de eficiencia del número de ítems necesarios para la calibración del estudiante, pruebas de equidad diagnóstica mediante distintos perfiles de habilidad, y experimentaciones con la calidad predictiva de las probabilidades generadas mediante el modelo, lo que requirió ajustar algunas métricas como el Brier Score. El método hipotéticodeductivo permitió extender la validación del sistema más allá de su comportamiento algorítmico, incorporando también hipótesis relacionadas con la propuesta de trabajo como servicio software. En este sentido, se formularon y experimentaron supuestos relacionados con la estabilidad del sistema bajo carga, su comportamiento ante la presencia de fallos en situaciones de alta concurrencia y el cumplimiento de umbrales aceptables de latencia y escalabilidad, aspectos críticos que deben contemplarse en sistemas educativos, especialmente en aquellos que se desarrollan como artefactos basados en web [6].

### 2.3.4 Aporte del método a la rigurosidad científica

El uso del método hipotéticodeductivo determinó que la investigación se sustentara en un proceso lógico, verificable y replicable, en el cual cada decisión de diseño tiene su respaldo en una hipótesis explícita y cada resultado experimental realiza la validación o refutación de dicha hipótesis. De esta forma se hizo más palpable el rigor científico del trabajo, al igual que se garantizó que la propuesta del Sistema de Evaluación Adaptativa respondiera a un proceso sistemático de investigación propio de la Ingeniería en Ciencias de la Computación en lugar de decisiones determinadas de un modo empírico o arbitrario.

## 2.4 Población y Muestra

La definición de la población y la muestra en estudios de validación de sistemas adaptativos basados en simulación computacional requiere un enfoque diferenciado respecto a investigaciones empíricas con participantes humanos. En el presente estudio, la población objetivo y la muestra de validación se establecieron considerando tanto el contexto educativo al cual se orienta el sistema como las características metodológicas propias de la validación algorítmica mediante técnicas de simulación estocástica.

### 2.4.1 Población objetivo

La población objetivo del Sistema de Evaluación Adaptativa (Componente B) está constituida por estudiantes universitarios de nivel superior que cursan asignaturas de cálculo diferencial, en concreto, aquellos que se encuentran en la fase de aprendizaje del tema de las derivadas y sus aplicaciones. Esta población presenta una heterogeneidad muy elevada en cuanto a conocimientos previos, ritmos de aprendizaje y estrategias de estudio, aspectos que justifican la necesidad de sistemas de evaluación personalizados y adaptativos. Desde una perspectiva psicométrica, esta población puede representarse mediante un continuo de habilidad latente () que refleja el nivel de dominio del contenido matemático evaluado. En

el marco de la Teoría de Respuesta al Ítem (IRT), la habilidad latente en poblaciones universitarias típicamente se distribuye en un rango aproximado de [-3.0, +3.0] en la escala logit, donde los valores negativos representan estudiantes con conocimientos insuficientes, los valores cercanos a cero corresponden a estudiantes de nivel medio, y los valores positivos indican un dominio avanzado o experto del tema [7], [8]. La delimitación de esta población objetivo resulta fundamental para interpretar adecuadamente los resultados de la validación y establecer los límites de generalización del sistema desarrollado. Si bien el presente estudio se realizó mediante simulación computacional, la caracterización explícita de la población objetivo guía tanto el diseño de los perfiles de estudiantes virtuales como la futura implementación del sistema en contextos educativos reales.

### 2.4.2 Muestra de validación

Dadas las limitaciones éticas, logísticas y operativas asociadas a la experimentación intensiva con estudiantes reales en etapas tempranas de desarrollo de sistemas adaptativos, se optó por realizar la validación inicial mediante simulación computacional con estudiantes virtuales, enfoque ampliamente aceptado en la literatura especializada en aprendizaje adaptativo y sistemas de tutoría inteligente [3], [6], [10]. La muestra de validación estuvo conformada por N = 10 perfiles de estudiantes virtuales, cada uno caracterizado por un conjunto de parámetros psicométricos controlados y conocidos a priori. Estos perfiles fueron diseñados para cubrir de manera sistemática el espectro de habilidad latente de la población objetivo, permitiendo evaluar el comportamiento del sistema adaptativo ante distintos niveles de conocimiento.

- **Habilidad latente real ($\theta$):** Distribuida uniformemente en el intervalo $[-2{,}0, +2{,}0]$, asignando un valor específico a cada perfil: $\{-2{,}0, -1{,}5, -1{,}0, -0{,}5, 0{,}0, +0{,}5, +1{,}0, +1{,}5, +2{,}0\}$, además de un décimo valor aleatorio dentro del intervalo con el fin de incrementar la variabilidad del conjunto de estudiantes simulados. Esta distribución permite representar estudiantes con bajo dominio ($\theta < -1{,}0$), dominio medio ($-1{,}0 \leq \theta \leq +1{,}0$) y dominio alto ($\theta > +1{,}0$).

- **Probabilidad de dominio inicial por habilidad (*Mastery*):** Modelada de forma correlacionada con la habilidad latente, siguiendo la relación $p_{\text{mastery}} \approx (\theta + 2)/4$, incorporando una variación estocástica de tipo gaussiano con desviación estándar $\sigma = 0{,}15$ para reflejar heterogeneidad realista entre estudiantes. Los valores resultantes fueron posteriormente normalizados en el intervalo $[0{,}1, 0{,}9]$.

- **Consistencia de respuesta:** Parámetro que representa la probabilidad de que el estudiante responda de forma coherente con su nivel de habilidad, distribuido uniformemente en el intervalo $[0{,}80, 0{,}95]$, donde valores cercanos a $1{,}0$ caracterizan estudiantes altamente consistentes, mientras que valores inferiores modelan variabilidad en el desempeño debida a errores accidentales, distracciones u otros factores contextuales.

- **Tasa de aprendizaje (*learning rate*):** Parámetro que modela la capacidad del estudiante para adquirir conocimiento a medida que progresa la sesión de evaluación, distribuido uniformemente en el intervalo $[0{,}10, 0{,}20]$. Este parámetro permite simular el efecto de la práctica y la asimilación progresiva de habilidades durante la interacción con el sistema.

- **Factor de fatiga:** Parámetro que representa el incremento progresivo del tiempo de respuesta como consecuencia de la fatiga cognitiva, distribuido uniformemente en el intervalo $[0{,}01, 0{,}03]$ por ítem administrado.

- **Fundamentación teórica de la parametrización:** La parametrización de los estudiantes virtuales se fundamentó en modelos teóricos de la Teoría de Respuesta al Ítem y en estudios empíricos sobre patrones de respuesta en evaluaciones adaptativas, garantizando que el comportamiento simulado fuera coherente con observaciones reales en contextos educativos [8], [10], [11].

### 2.4.3 Banco de items

El banco de ítems utilizado para la validación del sistema estuvo conformado por 200 ítems de opción múltiple diseñados para evaluar conocimientos sobre derivadas. Cada ítem fue caracterizado mediante el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem, incluyendo los siguientes parámetros psicométricos:

a) **Parámetro de discriminación ($a$):** Distribuido siguiendo una distribución log-normal con media $\mu = 0{,}3$ y desviación estándar $\sigma = 0{,}4$, resultando en valores comprendidos en el rango $[0{,}5, 2{,}5]$ tras la aplicación de límites de truncamiento. Este parámetro refleja la capacidad del ítem para diferenciar entre estudiantes con distintos niveles de habilidad.

b) **Parámetro de dificultad ($b$):** Distribuido uniformemente en el intervalo $[-3{,}0, +3{,}0]$, garantizando una cobertura amplia del continuo de habilidad. La distribución consideró aproximadamente un $33\,\%$ de ítems fáciles ($b < -0{,}6$), un $33\,\%$ de ítems de dificultad media ($-0{,}6 \leq b \leq +0{,}6$) y un $33\,\%$ de ítems difíciles ($b > +0{,}6$).

c) **Parámetro de adivinanza ($c$):** Distribuido uniformemente en el intervalo $[0{,}0, 0{,}25]$, representando la probabilidad de que un estudiante responda correctamente por azar en ítems de opción múltiple.

Los 200 ítems fueron distribuidos equitativamente entre dos habilidades específicas (skills) del dominio de derivadas: regla de la potencia (100 ítems) y regla de la cadena (100 ítems). Esta distribución balanceada aseguró la disponibilidad de ítems suficientes para la evaluación adaptativa de ambas habilidades durante las sesiones simuladas. Es importante señalar que, dado el carácter de validación algorítmica del presente estudio, los parámetros IRT de los ítems fueron generados de forma sintética mediante procedimientos estocásticos controlados, en lugar de ser calibrados empíricamente con datos reales de estudiantes. Si

bien esta es una práctica estándar en fases tempranas de desarrollo de sistemas adaptativos [3], [6], se reconoce como una limitación que será abordada en fases posteriores del proyecto mediante la calibración del banco de ítems con datos reales.

### 2.4.4 Justificación del tamaño muestral

La determinación del tamaño muestral en estudios de validación mediante simulación computacional responde a criterios distintos de aquellos utilizados en investigaciones con participantes humanos. En lugar de fundamentarse en cálculos de potencia estadística para detectar diferencias entre grupos, el tamaño muestral en simulaciones se orienta a garantizar la cobertura representativa del espacio de parámetros y la estabilidad de las estimaciones obtenidas mediante métodos Monte Carlo [3], [10]. La literatura especializada en evaluación de sistemas de testing adaptativo computarizado (CAT) y algoritmos de selección de ítems recomienda un tamaño muestral mínimo de N  10 perfiles de estudiantes virtuales para validaciones iniciales de tipo algorítmico, siempre que estos perfiles cubran de manera sistemática el rango de habilidad de interés y presenten heterogeneidad en sus características de respuesta [10], [11]. Este tamaño permite evaluar la estabilidad del algoritmo, la convergencia de las estimaciones y la ausencia de sesgos sistemáticos en distintos niveles de habilidad. En el presente estudio, el tamaño muestral de N = 10 fue considerado suficiente para los siguientes propósitos metodológicos:

- **Evaluación de convergencia:** Analizar si el algoritmo de estimación de habilidad mediante EAP converge de forma estable hacia el valor verdadero de $\theta$ en distintos niveles del continuo de habilidad.
- **Análisis de equidad diagnóstica:** Verificar que el sistema no presente sesgos sistemáticos en la precisión de las estimaciones entre estudiantes con bajo, medio y alto rendimiento.
- **Evaluación de eficiencia:** Determinar el número promedio de ítems requeridos para alcanzar niveles aceptables de precisión diagnóstica, definidos como $SE(\theta) \leq 0{,}4$, en distintos perfiles de estudiantes.
- **Detección de fallos algorítmicos:** Identificar posibles errores lógicos, condiciones de borde no controladas o comportamientos anómalos del sistema bajo escenarios diversos.

Adicionalmente, cada perfil de estudiante virtual fue sometido a sesiones de evaluación de hasta 20 ítems, generando un total de aproximadamente 200 interacciones ítem-estudiante registradas. Este volumen de datos resulta suficiente para calcular métricas agregadas con errores estándar aceptables y realizar análisis de sensibilidad ante distintas condiciones de operación del sistema. Es importante destacar que, si bien el tamaño muestral de N = 10 resulta adecuado para la validación técnica y algorítmica del sistema, no es suficiente para realizar inferencias estadísticas generalizables a la población objetivo de estudiantes reales.

Esta fase de validación corresponde a una evaluación de tipo técnico, orientada a verificar el correcto funcionamiento del motor adaptativo antes de su despliegue en contextos educativos reales. La validación con estudiantes humanos, que requerirá tamaños muestrales mayores determinados mediante análisis de potencia estadística, se plantea como una etapa posterior del proyecto. Finalmente, la adopción de simulación computacional como estrategia de validación inicial presenta ventajas metodológicas significativas, tales como la capacidad de controlar rigurosamente las variables del experimento, la posibilidad de reproducir exactamente las mismas condiciones en múltiples ejecuciones (reproducibilidad) y la evaluación del sistema ante escenarios extremos o poco probables que serían difíciles de observar en contextos reales. Estas características fortalecen la validez interna del estudio y permiten una evaluación exhaustiva del comportamiento del sistema antes de su uso con estudiantes reales [3], [6], [10].

## 2.5 Variables de la Investigación

La identificación, operacionalización y clasificación de las variables de estudio constituyen elementos fundamentales en el diseño experimental de investigaciones cuantitativas en ingeniería de software, particularmente en el contexto de sistemas adaptativos basados en modelos psicométricos y técnicas de aprendizaje automático. En el presente estudio, las variables fueron definidas siguiendo los principios de la Teoría de Respuesta al Ítem y los modelos bayesianos de rastreo de conocimiento, asegurando su medibilidad, reproducibilidad y coherencia con el marco teórico adoptado [7], [8]. Las variables se clasificaron en tres categorías principales: variables independientes, correspondientes a los parámetros controlados o manipulados durante la simulación; variables dependientes, que representan las salidas o mediciones generadas por el Sistema de Evaluación Adaptativa; y variables de control, que permanecieron constantes durante los experimentos para aislar los efectos de las variables independientes sobre las dependientes. Esta clasificación permite establecer relaciones causales claras y facilita la interpretación de los resultados obtenidos durante la fase de validación [6]. La definición explícita de las variables y su operacionalización resulta esencial para garantizar la reproducibilidad del estudio, facilitar la interpretación de los resultados y posibilitar futuras réplicas o extensiones de la investigación en contextos similares.

### 2.5.1 Variables independientes

Las variables independientes corresponden a los parámetros psicométricos y comportamentales de los estudiantes virtuales, así como a las características de los ítems administrados. Estas variables fueron manipuladas de forma controlada durante la simulación con el propósito de evaluar su impacto sobre el desempeño del motor adaptativo.

#### 2.5.1.1 Habilidad latente real del estudiante $\theta$

La habilidad latente constituye la variable independiente principal del estudio. Representa el nivel verdadero de conocimiento del estudiante en el dominio evaluado, expresado en la escala logit de la Teoría de Respuesta al Ítem. Esta variable fue operacionalizada mediante valores numéricos reales distribuidos uniformemente en el rango [-2.0, +2.0], donde valores negativos representan estudiantes con conocimientos insuficientes, valores cercanos a cero corresponden a estudiantes de nivel medio, y valores positivos indican dominio avanzado del contenido [7], [8]. La distribución uniforme de los valores de asignados a los estudiantes virtuales permite evaluar el comportamiento del sistema adaptativo en todo el espectro de habilidad relevante, evitando sesgos hacia estudiantes de un nivel específico y facilitando el análisis de equidad diagnóstica.

#### 2.5.1.2 Parámetros IRT del ítem

#### 2.5.1.3 Consistencia de respuesta

#### 2.5.1.4 Tasa de aprendizaje (learning rate)

#### 2.5.1.5 Factor de fatiga

### 2.5.2 Variables dependientes

#### 2.5.2.1 Habilidad estimada ($\hat{\theta}$)

#### 2.5.2.2 Error estándar de la estimación ($SE(\theta)$)

#### 2.5.2.3 Error de estimación absoluto ($|\theta - \hat{\theta}|$)

#### 2.5.2.4 Probabilidad de dominio por habilidad ($p_{\text{mastery}}$)

#### 2.5.2.5 Brier Score

#### 2.5.2.6 Latencia de respuesta del sistema

#### 2.5.2.7 Número de ítems administrados hasta convergencia

### 2.5.3 Variables de control

#### 2.5.3.1 Parámetros del modelo BKT por habilidad

#### 2.5.3.2 Configuración del algoritmo EAP

#### 2.5.3.3 Umbrales de decisión

#### 2.5.3.4 Parámetros de decay temporal

### 2.5.4 Síntesis de variables

La Tabla 2.3 presenta una síntesis de las variables de la investigación, clasificadas según su rol en el diseño experimental, incluyendo su descripción, unidad de medida y rango de valores observados o asignados.

| Tipo | Variable | Descripción | Unidad | Rango |
|---|---|---|---|---|
| Independiente | $\theta$ (habilidad real) | Nivel verdadero de conocimiento | Escala logit | $[-2{,}0, +$ |
| Independiente | $a$ (discriminación) | Capacidad discriminativa del ítem | Adimensional | $[0{,}5, 2{,}5$ |
| Independiente | $b$ (dificultad) | Nivel de dificultad del ítem | Escala logit | $[-3{,}0, +$ |
| Independiente | $c$ (adivinanza) | Probabilidad de acierto por azar | Probabilidad | $[0{,}0, 0{,}2$ |
| Independiente | Consistencia | Coherencia en las respuestas | Probabilidad | $[0{,}80, 0{,}$ |
| Independiente | *Learning rate* | Tasa de aprendizaje incremental | Por ítem | $[0{,}10, 0{,}$ |
| Independiente | Factor de fatiga | Incremento de tiempo por fatiga | Por ítem | $[0{,}01, 0{,}$ |
| Dependiente | $\hat{\theta}$ (habilidad estimada) | Estimación EAP de la habilidad | Escala logit | $[-4{,}0, +$ |
| Dependiente | $SE(\hat{\theta})$ | Error estándar de la estimación | Escala logit | $[0{,}2, 1{,}0$ |
| Dependiente | $\lvert\theta - \hat{\theta}\rvert$ | Error de estimación absoluto | Escala logit | $[0{,}0, 2{,}0$ |
| Dependiente | $p_{\text{mastery}}$ | Probabilidad de dominio por habilidad | Probabilidad | $[0{,}0, 1{,}0$ |
| Dependiente | Brier Score | Calibración predictiva | Error cuadrático | $[0{,}0, 1{,}0$ |
| Dependiente | Latencia | Tiempo de respuesta del sistema | Milisegundos | $[50, 500$ |
| Dependiente | $N$ ítems de convergencia | Ítems requeridos hasta $SE(\hat{\theta}) \leq 0{,}4$ | Cantidad | $[5, 20]$ |
| Control | Parámetros BKT | $p_{L0}, p_T, p_G, p_S, p_F$ | Probabilidades | Fijos po |
| Control | Configuración EAP | Grid, prior $N(0, 1)$ | – | Fijos |
| Control | Umbrales | $\tau$, $SE$ objetivo, límites | – | Fijos |

Cuadro 2.3: Definición y clasificación de variables del estudio

## 2.6 Marco metodológico de desarrollo

El desarrollo del Sistema de Evaluación Adaptativa (Componente B) se llevó a cabo siguiendo una metodología propia de la Ingeniería del Software, lo que hizo posible una construcción sistemática, controlada y alineada con buenas prácticas de desarrollo. Bajo esta lógica, se adoptó la metodología ágil SCRUM como marco de gestión del proceso de desarrollo, complementando el método de investigación hipotéticodeductivo descrito anteriormente. Vale la pena aclarar aquí un punto que puede generar confusión: SCRUM no fue empleado como método de investigación científica en sí mismo, sino como un mecanismo práctico para organizar, planificar y dar seguimiento al trabajo técnico que implicaba construir el sistema.

### 2.6.1 Justificación de la elección metodológica

La decisión de trabajar con SCRUM tiene que ver directamente con la naturaleza del sistema que se estaba desarrollando. El motor adaptativo integra varios módulos que dependen

unos de otros: modelos psicométricos, lógica de selección adaptativa de ítems, persistencia del estado del estudiante, instrumentación de métricas y mecanismos de validación. Estos componentes necesitaban ciclos cortos de desarrollo, prueba y ajuste para poder integrarse correctamente. Es como armar un mecanismo complejo donde cada pieza debe encajar con precisión, pero solo se puede verificar que funciona una vez que las partes están conectadas. SCRUM dio la flexibilidad necesaria para ir incorporando funcionalidades poco a poco y ver en tiempo real cómo afectaban al comportamiento global del sistema [13], [14]. Hay otra razón de peso para elegir metodologías ágiles cuando se trabaja con sistemas basados en inteligencia artificial y aprendizaje automático. La incertidumbre es alta: no siempre se puede anticipar cómo va a comportarse un modelo psicométrico bajo condiciones reales, o qué impacto tendrá modificar ciertos parámetros de convergencia. Los requisitos técnicos también suelen evolucionar conforme se van realizando experimentos y se obtienen resultados inesperados. SCRUM permite ajustar el plan de desarrollo según lo que va mostrando la evidencia empírica en cada iteración, lo que reduce bastante el riesgo de tomar decisiones de diseño que después resulten desconectadas de los resultados experimentales [6].

### 2.6.2 Estructura y organización de los sprints

El marco SCRUM que se aplicó en este proyecto se organizó mediante sprints de una semana cada uno, con objetivos técnicos específicos y entregables que se podían verificar. Elegir una semana no fue casualidad: es un período que permite ver resultados tangibles sin tener que esperar demasiado, pero a la vez da tiempo suficiente para implementar funcionalidades que cumplan con estándares mínimos de calidad. El desarrollo de cada iteración obedecía a una progresión definida de las siguientes tareas:

- **Planificación del sprint:** Definir los objetivos técnicos y seleccionar las tareas del backlog que se abordarían.
- **Desarrollo iterativo:** Implementar las funcionalidades que se habían priorizado.
- **Pruebas unitarias:** Verificar que cada componente individual funcionara como debía.
- **Pruebas de integración:** Comprobar que los módulos interactuaran correctamente entre sí.
- **Validación funcional:** Confirmar que se cumplieran los requisitos técnicos y algorítmicos establecidos.

Este proceso garantizaba que cada incremento tuviera un nivel mínimo de calidad antes de integrarse al sistema base. La idea era evitar acumular deuda técnica que después pudiera comprometer la estabilidad del motor adaptativo cuando el sistema creciera en complejidad.

### 2.6.3 Adaptación al contexto académico

Hay que señalar algo importante: la forma en que se usó SCRUM aquí no es exactamente igual a como se usa en un proyecto comercial típico. En una empresa, cada sprint busca

entregar valor tangible al cliente o al usuario final. En este caso, cada sprint se centraba más bien en validar técnica y algorítmicamente el sistema, siguiendo las hipótesis de investigación que se habían planteado. Los entregables no se medían tanto por funcionalidades listas para producción, sino por componentes validados empíricamente que confirmaban o cuestionaban aspectos específicos del diseño propuesto.
Esta adaptación hizo que el proceso de desarrollo estuviera muy conectado con el proceso investigativo, sin sacrificar el rigor metodológico ni perder la trazabilidad de las decisiones técnicas. Cada decisión de diseño, cada ajuste que se hacía en los algoritmos, cada refactorización importante quedaba documentada y justificada según los resultados experimentales que se iban obteniendo. El desarrollo técnico no era un fin en sí mismo, sino más bien una forma de responder las preguntas de investigación que se habían formulado al inicio.
La Tabla 2.4 presenta los roles principales y los artefactos de SCRUM que se consideraron durante el desarrollo del Componente B, adaptados específicamente al contexto de un proyecto académico en Ingeniería en Ciencias de la Computación.

| **Elemento SCRUM** | **Descripción** | **Aplicación en el proyecto** |
|---|---|---|
| Product Owner | Responsable de priorizar requisitos y objetivos | Definición de funcionalidades del motor adaptativo |
| Scrum Master | Facilitador del proceso ágil | Gestión del flujo de desarrollo y resolución de bloqueos |
| Equipo de desarrollo | Responsable de implementar el producto | Diseño e implementación del motor, simulador y pruebas |
| Product Backlog | Lista priorizada de requisitos | Funcionalidades técnicas y experimentales |
| Sprint Backlog | Tareas seleccionadas para el sprint | Implementaciones específicas por iteración |
| Incremento | Versión funcional del producto | Versiones sucesivas del motor adaptativo |

Cuadro 2.4: Elementos SCRUM aplicados al desarrollo del Sistema de Evaluación Adaptativa

### 2.6.4 Planificación progresiva y mejora continua

Un rasgo que definió singularmente la utilización de dicha técnica fue la planificación incremental de los sprints y no una planificación inamovible y totalmente anticipada. Esta forma de proceder permitió que los resultados obtenidos como producto del sprint permitieran tomar decisiones sobre el diseño y la priorización de las tareas que deberían desarrollarse después del correspondiente sprint. Con el propósito de ofrecer un ejemplo concreto, en el caso de que en el contexto de un sprint quedaran evidenciados problemas de convergencia por parte del algoritmo de selección adaptativa de ítems en relación con un grupo de estudiantes con niveles de habilidad muy bajos, la programación del sprint siguiente podría establecerse tomando como posición de partida aquella información para actualizar la programación que pudiera incluir cambios en los criterios de selección como modificaciones en

el recalibrado de los parámetros del modelo inicial.
Este enfoque iterativo hizo posibles varios tipos de mejoras que resultaron fundamentales:

- **Ajustes algorítmicos:** Refinar los parámetros del modelo de Rasch y los criterios de convergencia según lo que mostraban los experimentos.
- **Optimización del rendimiento:** Mejorar la eficiencia computacional del motor de selección de ítems cuando se identificaban cuellos de botella.
- **Refactorizaciones estructurales:** Reorganizar el código para que fuera más fácil de mantener y extender conforme el sistema crecía en complejidad.

Todas estas mejoras se basaron en la evidencia empírica que se iba obteniendo durante la validación del sistema. No se trataba de hacer cambios porque sonaban bien en teoría, sino porque las pruebas mostraban que eran necesarios.
La Tabla 2.5 detalla la distribución temporal de los sprints que se desarrollaron a lo largo del ciclo de construcción del Componente B, destacando los objetivos técnicos que se alcanzaron en cada etapa del proyecto.

| **Sprint** | **Objetivo principal** | **Resultados obtenidos** |
|---|---|---|
| Sprint 0–1 | Investigación y diseño inicial | Selección de frameworks, definición de arquitectura y contratos JSON |
| Sprint 2–3 | Diseño del modelo adaptativo | Implementación del modelo híbrido IRT+BKT |
| Sprint 4–5 | Implementación algorítmica | Estimación EAP, selección adaptativa de ítems y *decay* temporal |
| Sprint 6 | Validación experimental | Simulación, pruebas automatizadas y pruebas de carga |
| Sprint 7 | Refactorización y documentación | Optimización del código, documentación técnica y cierre del desarrollo |

Cuadro 2.5: Planificación incremental de sprints para el desarrollo del motor adaptativo

### 2.6.5 Resultados de la aplicación de SCRUM

Usar SCRUM contribuyó bastante a manejar la complejidad que implicaba desarrollar este sistema. Entre los beneficios más evidentes están:

- **Identificación temprana de errores:** Detectar problemas en etapas tempranas mediante la validación continua, cuando corregirlos resulta significativamente menos costoso en términos de tiempo y esfuerzo.
- **Validación incremental:** Probar a fondo cada funcionalidad antes de pasar a la siguiente, asegurando tener una base sólida y estable para seguir construyendo.
- **Mejoras basadas en datos:** Fundamentar los cambios del sistema en evidencia experimental concreta, no en especulaciones sobre cómo debería comportarse.

- **Trazabilidad completa:** Mantener una alineación constante entre los objetivos de investigación, las decisiones de diseño y los resultados que se iban obteniendo mediante la estructura iterativa.

Esta trazabilidad tiene un valor especial en un contexto académico, donde poder reproducir el trabajo y justificar rigurosamente las decisiones técnicas son aspectos centrales del proceso investigativo. Cada sprint generó documentación detallada que permitiría a otros investigadores entender no solo qué se implementó, sino por qué se tomaron determinadas decisiones de diseño. El marco metodológico adoptado garantizó que el Sistema de Evaluación Adaptativa se desarrollara siguiendo principios de calidad de software, mantenibilidad y extensibilidad. Estos aspectos son fundamentales tanto para la futura integración del sistema con otros componentes del ecosistema de aprendizaje como para su eventual implementación en entornos educativos reales. La combinación del método hipotético-deductivo para la investigación y SCRUM para el desarrollo técnico resultó efectiva, permitiendo mantener el rigor científico mientras se construía un artefacto computacional que realmente funciona [13], [14].

## 2.7 Técnicas e Instrumentos de Recolección

La recolección de información para validar el Sistema de Evaluación Adaptativa (Componente B) se apoyó en técnicas propias de la Ingeniería en Ciencias de la Computación, las cuales permitieron obtener datos objetivos, reproducibles y que provienen directamente de la ejecución del sistema. Dado el carácter algorítmico, experimental y computacional de esta investigación, no se utilizaron encuestas ni instrumentos cualitativos tradicionales. En su lugar, se utilizaron simulación computacional, generación de datos sintéticos, telemetría automática y pruebas controladas de rendimiento, métodos ampliamente reconocidos en la evaluación de sistemas adaptativos, sistemas de tutoría inteligente y software basado en inteligencia artificial [4], [6], [10].
Estas técnicas posibilitaron la recolección de información tanto del comportamiento pedagógico del motor adaptativo como de su funcionamiento computacional como servicio software. La información recabada caracteriza adecuadamente el funcionamiento interno del sistema, permitiendo su análisis cuantitativo bajo criterios de precisión diagnóstica, eficiencia algorítmica, estabilidad operativa y escalabilidad, aspectos considerados fundamentales en la validación de sistemas auto-adaptativos [6].

### 2.7.1 Simulación estocástica de estudiantes virtuales

Como técnica principal de recopilación se utilizó la simulación estocástica de estudiantes, que se implementó mediante un software específico desarrollado para este propósito: el simulador de estudiantes virtuales. Este simulador crea agentes artificiales que se parametrizan según perfiles psicométricos definidos, que incluyen el nivel de habilidad latente inicial $(\theta)$, la consistencia en las respuestas, la probabilidad de acierto al azar y la tasa de apren-

dizaje. Estos parámetros posibilitan modelar una amplia variedad de comportamientos de aprendizaje, siguiendo los supuestos de la Teoría de Respuesta al Ítem y los modelos de rastreo de conocimiento [8], [10], [11]. Con la simulación se recolectó un volumen considerable de interacciones controladas entre los estudiantes virtuales y el motor adaptativo. Esto permitió analizar cómo converge la estimación de habilidad a medida que se administran más ítems, cómo va disminuyendo el error estándar de medición, y en qué medida el diagnóstico resulta equitativo cuando se aplica a distintos perfiles de estudiantes. La simulación también sirvió para reproducir de forma sistemática situaciones extremas que rara vez se encuentran en la práctica educativa cotidiana: respuestas erráticas que no siguen un patrón predecible, casos de aprendizaje acelerado donde el estudiante avanza muy rápido, o situaciones de estancamiento prolongado donde no se observa progreso significativo. Este tipo de escenarios son muy difíciles de estudiar con estudiantes reales, no solo por las obvias implicaciones éticas de exponer a los estudiantes a evaluaciones poco apropiadas, sino también por la complejidad práctica de controlar todas las variables en un entorno educativo auténtico.

### 2.7.2 Generación de datos sintéticos

De forma complementaria a la simulación, se emplearon técnicas de generación de datos sintéticos con el objetivo de evaluar la robustez del sistema ante distintas condiciones operativas. Mediante el uso de perfiles psicométricos y secuencias de interacción controladas, se sometió al motor adaptativo a escenarios específicamente diseñados para poner a prueba sus mecanismos de estimación, selección adaptativa y actualización del estado del estudiante. Esta estrategia resulta particularmente eficaz en la validación de sistemas adaptativos complejos, dado que la diversidad de situaciones del mundo real es difícil de abarcar completamente en las primeras etapas de desarrollo [3], [6].

### 2.7.3 Sistema de registro y telemetría automática

Para el registro de información se diseñó un sistema de telemetría automática basado en archivos de auditoría estructurados en formato JSON. Este sistema registra de forma secuencial e inmutable toda interacción que procesa el motor adaptativo: la selección del ítem, la respuesta del estudiante, la estimación de habilidad latente, la probabilidad de dominio por habilidad y la recomendación que genera el motor de inferencia. Estos registros son la fuente primaria para analizar posteriormente el comportamiento del sistema, a la vez que aseguran la trazabilidad completa de las decisiones algorítmicas que se implementan. Los archivos de auditoría permiten obtener métricas de desempeño bastante detalladas, pero más allá de eso, hacen posible reconstruir sesiones completas de forma determinista usando mecanismos de replay. Esto tiene un valor metodológico importante porque asegura que los experimentos sean verificables y replicables, dos aspectos que resultan básicos en cualquier investigación rigurosa de ingeniería de software y sistemas auto-adaptativos [6].

### 2.7.4 Pruebas de carga y concurrencia

Como técnica de recolección orientada al desempeño computacional, se llevaron a cabo pruebas de carga y concurrencia. Para ello se utilizaron herramientas de simulación de usuarios concurrentes que permitieron generar peticiones simultáneas al servicio de evaluación adaptativa. Durante estas pruebas se capturaron métricas de latencia, tasa de peticiones procesadas por segundo (RPS), estabilidad del sistema y comportamiento bajo condiciones de estrés. Estas métricas resultan esenciales para valorar la viabilidad del despliegue del sistema en entornos de aprendizaje auténticos con múltiples usuarios concurrentes.

### 2.7.5 Síntesis de las técnicas empleadas

La combinación de estas técnicas permitió obtener una caracterización completa del comportamiento del Sistema de Evaluación Adaptativa, tanto desde la perspectiva algorítmica como desde el funcionamiento del sistema software. La Tabla 2.6 presenta las principales técnicas de recolección de información utilizadas en el estudio, el tipo de datos obtenidos y el objetivo metodológico correspondiente.

| **Técnica** | **Instrumento** | **Datos recolectados** | **Propósito metod** |
|---|---|---|---|
| Simulación estocástica | Simulador de estudiantes virtuales | Respuestas simuladas, convergencia de $\theta$, error estándar | Evaluar precisión dad del modelo |
| Datos sintéticos | Perfiles psicométricos parametrizados | Escenarios controlados de aprendizaje | Probar robustez miento límite |
| Telemetría automática | Logs de auditoría en formato JSON | Historial de sesiones, métricas internas | Trazabilidad y aná do |
| Replay determinista | Reconstrucción desde logs | Secuencias completas de interacción | Verificabilidad y re |
| Pruebas de carga | Simulación de usuarios concurrentes | Latencia, RPS, estabilidad | Evaluar escalabi sempeño |

Cuadro 2.6: Técnicas, instrumentos y datos utilizados en la validación experimental

El conjunto de técnicas e instrumentos de recolección empleados permitió recopilar información válida, estructurada y vinculada directamente con el comportamiento real del sistema, evitando sesgos derivados de mediciones subjetivas o indirectas. Esta práctica de recolección de datos se encuentra bien establecida en la literatura sobre evaluación de sistemas adaptativos e ingeniería de software, constituyendo un enfoque con alto grado de rigor metodológico [4], [6].

La información recolectada mediante estos instrumentos constituye la base empírica sobre la cual se fundamenta el análisis estadístico y experimental que se desarrolla en las secciones subsiguientes del marco metodológico.

## 2.8 Actividades y productos del proyecto

La elaboración del Sistema de Evaluación Adaptativa (Componente B) utilizó una serie de actividades técnicas organizadas en forma de actividades secuenciadas y planificadas para cumplir, de forma progresiva, con los objetivos específicos establecidos en el Plan de Trabajo de Integración Curricular. Dichas actividades se plantearon como acciones de carácter concreto y demostrable, coherentes con el enfoque cuantitativo, el diseño de experimentos y el marco metodológico de desarrollo ágil presentados en las secciones anteriores, garantizando la existencia de una trazabilidad del desarrollo entre los objetivos que se fijan, las decisiones técnicas adoptadas y los resultados finalmente conseguidos. Desde una visión metodológica, el trabajo de las actividades se realizó siguiendo principios propios de la Ingeniería en Ciencias de la Computación mediante los cuales cada uno de los objetivos se traduce en tareas de análisis, diseño, implementación y validación. Este planteamiento garantiza que el cumplimiento de los objetivos no se reduzca a formulaciones teóricas sino que se concrete en componentes software funcionales, evaluables y respaldados por la evidencia empírica que, tal y como sugieren los estudios en desarrollo de sistemas adaptativos y de tutoría inteligente [4], [6], resulta necesario obtener. Los procesos de desarrollo se concibieron como procesos incrementales e iterativos, y los productos que se obtenían en cada conjunto de actividades devuelven la información a las decisiones que los responsables toman en la siguiente iteración del proceso. Con ello se facilitaba la escalación del sistema desde un primer diseño conceptual de él a un motor adaptativo totalmente operativo, testado mediante simulación computacional, pruebas automáticas y análisis de rendimiento. Este enfoque es muy adecuado para aquellos proyectos donde se pueden combinar de formas complejas modelos psicométricos y técnicas de inteligencia artificial, ya que en estos proyectos la adecuada sintonía de los algoritmos depende de la evidencia que se obtenga durante su experimentación [8], [10], [11].

### 2.8.1 Actividades por objetivo específico

#### 2.8.1.1 Objetivo 1: Análisis de herramientas de inteligencia artificial y aprendizaje automático

Respecto al primer objetivo específico, enmarcado en el análisis de herramientas de inteligencia artificial y aprendizaje automático aplicadas a la evaluación adaptativa, se llevaron a cabo actividades de revisión sistemática del estado del arte del aprendizaje adaptativo, de los sistemas de tutoría inteligente y de la psicometría computacional. En dicho análisis se incluyeron actividades específicas como la comparación de modelos de Teoría de Respuesta al Ítem y técnicas de rastreo de conocimiento, las cuales evaluaron distintos criterios como la precisión diagnóstica, la interpretabilidad y la viabilidad computacional. De estas actividades se extrajo de manera justificada la selección de un modelo híbrido basado en IRT (3PL) y BKT, combinación que es avalada por la literatura como una alternativa efectiva

para la evaluación adaptativa [8], [10], [11], [12].

#### 2.8.1.2 Objetivo 2: Diseño del sistema de evaluación progresiva personalizada

El segundo objetivo específico se centró en el diseño de un sistema de evaluación progresiva fundamentada en la personalización adaptativa. Para su consecución se llevaron a cabo actividades de diseño de la arquitectura del motor adaptativo, definición de contratos de comunicación entre componentes y modelado de la lógica de selección adaptativa de ítems. Estas actividades dieron lugar a la especificación de reglas de parada, criterios de actualización del estado del estudiante y mecanismos de interoperabilidad con otros componentes del ecosistema de aprendizaje, lo que permitió garantizar un diseño modular, extensible y ajustado a buenas prácticas de ingeniería de software [6].

#### 2.8.1.3 Objetivo 3: Implementación de modelos para el análisis y seguimiento del aprendizaje

En lo que respecta al objetivo relacionado con la implementación de modelos para el análisis y seguimiento del aprendizaje, se llevaron a cabo actividades técnicas de codificación del modelo IRT con estimación EAP y del modelo bayesiano de rastreo de conocimiento con decaimiento temporal. Las implementaciones de ambos modelos fueron integradas en el motor adaptativo junto con mecanismos para calcular métricas de desempeño y de aprendizaje en tiempo real, lo que permitió generar indicadores objetivos acerca de la evolución del conocimiento del estudiante. La correcta implementación de estos modelos fue un aspecto central dada la relación directa entre ésta y las capacidades del sistema en términos de inferencias y adaptaciones adecuadas [8], [11].

#### 2.8.1.4 Objetivo 4: Validación mediante pruebas funcionales y experimentales

El cuarto objetivo específico se centró en la validación del sistema mediante pruebas funcionales y experimentales. Para ello se construyó un simulador de estudiantes virtuales que permitió llevar a cabo interacciones controladas con el motor adaptativo y que facilitó la realización de pruebas de convergencia de la habilidad estimada, análisis de eficiencia del número de ítems administrados, evaluación de la equidad diagnóstica entre perfiles de aprendizaje y medición de la calidad predictiva de las probabilidades generadas por el modelo. Como complemento, se llevaron a cabo pruebas de carga y concurrencia como forma de evaluar el comportamiento del sistema como servicio software en condiciones de estrés, siendo este un aspecto clave para el futuro despliegue en entornos educativos reales [6], [10].

### 2.8.2 Productos y evidencias técnicas generadas

Las actividades desarrolladas para cada objetivo generaron productos y evidencias técnicas concretas en forma de artefactos de diseño, módulos de software funcionales, registros de simulación, reportes de pruebas automatizadas y métricas de rendimiento. Esta producción de evidencias fue la que permitió comprobar de forma objetiva el cumplimiento de los objetivos específicos y a la vez facilitó la evaluación del avance del proyecto a lo largo del tiempo. La Tabla 2.7 muestra una síntesis de la relación entre objetivos específicos del proyecto, principales actividades desarrolladas y productos o evidencias generadas, donde se evidencia la trazabilidad metodológica entre lo planificado y lo ejecutado.

| **Objetivo específico** | **Actividades desarrolladas** | **Productos / evidencias** |
|---|---|---|
| Analizar herramientas de IA aplicables a la evaluación adaptativa | Revisión del estado del arte en aprendizaje adaptativo, ITS e IRT. Análisis comparativo de modelos psicométricos y de rastreo de conocimiento. | Selección fundamentada del modelo híbrido IRT (3PL) + BKT |
| Diseñar un sistema de evaluación progresiva y personalizada | Diseño de la arquitectura del motor adaptativo y definición de contratos de comunicación y reglas de selección de ítems. | Arquitectura del Componente B y esquemas JSON |
| Implementar modelos de análisis y seguimiento del aprendizaje | Implementación del modelo IRT con EAP y del modelo BKT con decaimiento temporal. Integración de métricas. | Motor adaptativo funcional y módulos de cálculo |
| Validar el sistema mediante pruebas funcionales y experimentales | Simulación de estudiantes virtuales, pruebas de convergencia, eficiencia, equidad y pruebas de carga. | Resultados de simulación, reportes de pruebas y métricas de rendimiento |

Cuadro 2.7: Correspondencia entre objetivos específicos, actividades y productos generados

## 2.9 Técnicas de Análisis de la Información

La actividad de análisis de la información obtenida durante la validación del Sistema de Evaluación Adaptativa (Componente B) se llevó a cabo utilizando técnicas del ámbito cuantitativo de la Ingeniería en Ciencias de la Computación, que sirven para evaluar no sólo el comportamiento algorítmico del motor adaptativo, sino también el comportamiento computacional que tiene lugar en la ejecución del sistema software. El análisis constituye una etapa importante del proceso metodológico porque sirve para contrastar empíricamente las hipótesis formuladas y para comprobar la consecución de los objetivos específicos de la investigación. Las técnicas de análisis escogidas están en línea con el enfoque cuantitativo y el diseño experimental que se han adoptado en esta investigación, dando preferencia al uso de métricas objetivas, numéricas, verificables, reproducibles y comparables. En este sentido, el análisis implementado se centra no solo en una interpretación descriptiva de los resultados sino que busca identificar patrones, evaluar la estabilidad del sistema ante diferentes escenarios y medir la precisión diagnóstica y la eficiencia operativa, tal y como

viene prescrito en la literatura que trata temas de aprendizaje adaptativo y sistemas auto-adaptativos [4], [6].

### 2.9.1 Análisis del rendimiento algorítmico

Desde el punto de vista del rendimiento algorítmico, el análisis realizado se centró en evaluar la precisión del modelo híbrido implementado considerando métricas de error ampliamente utilizadas en psicometría computacional. Dentro de estas métricas se encuentran el error cuadrático medio (RMSE) y el error absoluto medio (MAE), que se calculan considerando la diferencia entre la habilidad real de los estudiantes simulados y la habilidad estimada por el modelo adaptativo. Estas métricas posibilitan evaluar de modo concreto el nivel de precisión del sistema en la estimación de estados latentes de conocimiento, que es precisamente la cuestión central de los sistemas que están fundamentados en Teoría de Respuesta al Ítem y en modelos de rastreo del conocimiento [7], [8], [11]. De manera complementaria, el análisis también incluyó la observación de la evolución del error estándar de medición correspondiente a la estimación de la habilidad latente. Este análisis permitió observar si el modelo iba convergiendo a medida que se van administrando ítems informativos, así como si el sistema era capaz de reducir la incertidumbre diagnóstica con el avance del proceso de interacción con el estudiante. La reducción sostenida de este error constituye un buen indicador de la eficiencia de los sistemas de evaluación adaptativa, dado que se interpreta como la capacidad del motor para proporcionar un diagnóstico preciso con un número reducido de ítems, optimizando así el proceso de evaluación [8], [10]. La calidad predictiva del sistema fue analizada a partir del Brier Score, una métrica ampliamente usada para el análisis de calibración de probabilidades obtenidas a partir de modelos probabilísticos. En el caso presente, el Brier Score permitió la comparación entre las probabilidades de respuesta correcta predichas por el sistema frente a los resultados observados en la simulación. El Brier Score hizo posible la evaluación del grado de alineación entre lo predicho por el sistema y el comportamiento real del estudiante simulado en relación con el resultado observado. Cuanto más bajo sea el valor correspondiente, más adecuada será la calibración probabilística, un aspecto muy a tener en cuenta para poder obtener decisiones adaptativas confiables [11], [12]. Junto con el Brier Score, también se implementó un análisis longitudinal del aprendizaje simulado dirigido a comprobar la capacidad del sistema para detectar los cambios en el dominio de las habilidades a lo largo del tiempo contemplando medidas de adquisición progresiva de los conocimientos y fenómenos del tipo de decaimiento o pérdida del dominio del mismo tipo, que permitiría comprobar la sensibilidad del modelo frente a los cambios temporales en el rendimiento del estudiante. Este tipo de análisis es muy relevante en sistemas de evaluación adaptativa que intentan proporcionar retroalimentación continua y personalizada [10], [11].

### 2.9.2 Análisis del rendimiento computacional

Desde el punto de vista del análisis del rendimiento computacional, fueron utilizadas métricas de ingeniería de software orientadas a evaluar el comportamiento del sistema bajo condiciones de carga. Entre dichas métricas se encuentran la latencia de respuesta (en milisegundos), los percentiles de tiempo de respuesta (P50 y P95) y la tasa de peticiones por segundo procesadas (RPS). Estos indicadores fueron analizados a partir de los datos obtenidos durante las pruebas de carga y concurrencia, permitiendo así identificar cuellos de botella y evaluar la escalabilidad del sistema antes de su despliegue en entornos educativos reales [6]. El análisis de las métricas indicadas estableció límites de rendimiento aceptables para el sistema, tanto por lo que respecta a la experiencia del usuario final como por los requerimientos técnicos de plataformas educativas basadas en servicios web. Se comprobó especialmente que el sistema contara con tiempos de respuesta estables y previsibles bajo situaciones de alta concurrencia, fundamental para garantizar su viabilidad operativa.

### 2.9.3 Herramientas y reproducibilidad del análisis

Las técnicas de análisis se implementaron utilizando herramientas y librerías estándar del ecosistema Python, tales como módulos estadísticos para el cálculo de métricas descriptivas, funciones matemáticas para la estimación de errores y procedimientos automatizados para el tratamiento de registros de telemetría. De esta forma se garantiza la transparencia del proceso analítico y se facilita la reproducibilidad de los resultados, dos principios de vital importancia en la investigación en ingeniería de software y sistemas auto-adaptativos [6]. Por último, los resultados obtenidos mediante estas técnicas de análisis fueron interpretados a partir de criterios de aceptación predefinidos resumidos en la Tabla 2.8, tales como niveles máximos de error admisibles, reducción esperada del error estándar de medición y umbrales aceptables de latencia y estabilidad. Estos criterios permitieron evaluar de forma objetiva el grado en que los objetivos del estudio fueron cumplidos y contribuir a sustentar las conclusiones que más adelante se presentan.

| **Dimensión analizada** | **Métrica** | **Descripción** | **Propósito del análi** |
|---|---|---|---|
| Precisión diagnóstica | RMSE / MAE | Error entre habilidad real y estimada | Evaluar exactitud de |
| Convergencia | Error estándar de medición | Nivel de incertidumbre en la estimación de $\theta$ | Analizar eficiencia a |
| Calidad predictiva | Brier Score | Calibración de probabilidades predichas | Validar confiabilidad delo |
| Rendimiento | Latencia (ms) | Tiempo de respuesta del sistema | Evaluar experiencia rio |
| Escalabilidad | RPS, P50/P95 | Capacidad bajo concurrencia | Analizar viabilidad o |

Cuadro 2.8: Dimensiones y métricas utilizadas para la evaluación del desempeño del motor adaptativo

## 2.10 Criterios de Validación y Aceptación

## 2.11 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

### 2.11.1 Resultados

### 2.11.2 Conclusiones

### 2.11.3 Recomendaciones

# REFERENCIAS BIBLIOGRÁFICAS

# BIBLIOGRAFÍA


[1] M. Zapata Ros, «IA generativa y ChatGPT en Educación: Un reto para la evaluación y £una nueva pedagogía?» *Revista Paraguaya de Educación a Distancia*, vol. 5, n.º 1, págs. 12-44, 2024. DOI: 10.56152/reped2024-vol5num1-art2

[2] R. Juárez Cádiz, «PathRAG application in adaptive learning with generative AI for inclusive and sustainable education,» *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 29, n.º 1, 2026. DOI: 10.5944/ried.45378

[3] G. C. Tenorio-Sepúlveda, A. Soberanes-Martín y M. Martínez-Reyes, «Diseño instruccional con aprendizaje adaptativo de un curso en línea: Redacción de protocolos de investigación,» *Revista de Gestión Universitaria*, vol. 2, n.º 3, págs. 9-16, mar. de 2018.

[4] N. Carbonell Bernal y M. Á. Hernández Prados, «Impacto de los Sistemas de Tutoría Inteligente. Una revisión sistemática,» *EDUTEC. Revista Electrónica de Tecnología Educativa*, n.º 89, págs. 121-132, sep. de 2024. DOI: 10.21556/edutec.2024.89.3025

[5] M. H. Rodríguez Chávez, «Sistemas de tutoría inteligente y su aplicación en la educación superior,» *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, vol. 12, n.º 22, e175, 2021. DOI: 10.23913/ride.v11i22.848

[6] O. Gheibi, D. Weyns y F. Quin, «Applying Machine Learning in Self-adaptive Systems: A Systematic Literature Review,» *ACM Transactions on Autonomous and Adaptive Systems*, vol. 15, n.º 3, Article 9, 2021. DOI: 10.1145/3469440

[7] M. D. Hidalgo-Montesinos y B. F. French, «Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas,» *Revista de Psicología Clínica con Niños y Adolescentes*, vol. 3, n.º 2, págs. 13-21, jul. de 2016.

[8] H. F. Attorresi, G. S. Lozzia, F. J. P. Abal, M. S. Galibert y M. E. Aguerri, «Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos,» *Revista Argentina de Clínica Psicológica*, vol. 18, n.º 2, págs. 179-188, ago. de 2009.

[9] F. J. P. Abal, G. S. Lozzia, M. E. Aguerri, M. S. Galibert y H. F. Attorresi, «La escasa aplicación de la teoría de respuesta al ítem en tests de ejecución típica,» *Revista Colombiana de Psicología*, vol. 19, n.º 1, págs. 111-122, 2010.

[10] Y. Hicke, *Knowledge Tracing Challenge: Optimal Activity Sequencing for Students*, arXiv:2311.14707v1, 2023.

[11] S. Xu, M. Sun, W. Fang, K. Chen, H. Luo y P. X. W. Zou, «A Bayesian-based knowledge tracing model for improving safety training outcomes in construction: An adaptive learning framework,» *Developments in the Built Environment*, vol. 13, pág. 100 111, 2023.

[12] A. Psychogyiopoulos, N. Smits y L. A. van der Ark, «Estimating the Joint Item-Score Density Using an Unrestricted Latent Class Model,» *Journal of Computerized Adaptive Testing*, vol. 12, n.º 3, págs. 136-151, jul. de 2025. DOI: 10.7333/2507-1203136

[13] E. Hernández-Salazar y C. A. Beltrán, «SCRUM, un enfoque práctico de metodología ágil para la ingeniería de software,» *Revista Tecnología, Investigación y Academia (TIA)*, vol. 8, n.º 2, págs. 61-73, 2020.

[14] K. Schwaber y J. Sutherland, *La Guía Scrum: La guía definitiva de Scrum Las reglas del juego*, Versión 2020. Licencia Creative Commons Attribution Share-Alike 4.0, 2020. dirección: https://scrumguides.org

[15] L. Vargas Peña, «Estudio comparativo de modelos adaptativos y análisis de su uso en un ambiente de m-Learning implementando la técnica adaptativa de la Teoría de Respuesta al Ítem (IRT),» Tesis de mtría., Universidad Rey Juan Carlos, Madrid, España, 2012.

[16] J. A. Sarango y M. Villamar, «Aplicación de metodologías de desarrollo de software en proyectos de ingeniería,» Escuela Superior Politécnica del Litoral, Ecuador, 2022.

# 3 ANEXOS

## ANEXO I: Aplicación Móvil

## ANEXO II: Aplicación Web

## ANEXO III: Página Web