

Clusterização Iterativa Baseada em Densidade em Rastreamento de Conhecimento para Identificação de Perfis e Lacunas de Aprendizagem

Taciana Vasconcelos¹, William Oliveira da Costa e Silva², Luiz Rodrigues³,
Mario Antonio Pessoa Santos⁴, Tyagi Lima⁴, Armando M. Toda⁵,
Gabriel Alves¹, Rafael Ferreira Mello¹

¹ UFRPE – Recife – PE – Brasil ² UFRR – Boa Vista – RR – Brasil

³ UTFPR – Apucarana – PR – Brasil ⁴ Cogna Educação S.A. – São Paulo – SP – Brasil

⁵ USP – São Carlos – SP – Brasil

{taciana.vasconcelos, gabriel.alves, rafael.mello}@ufrpe.br

williamcostac@gmail.com, luizrodrigues@utfpr.edu.br, armando.toda@usp.br

{mario.a.santos, tyagi.lima}@cogna.com.br

Abstract. *Knowledge Tracing (KT) techniques allow monitoring student proficiency, but transforming them into interpretable profiles remains a challenge at scale. This paper presents a two-stage density-based clustering method to identify student profiles and their learning gaps from KT data. The method was evaluated with 19,048 students from Calculus I. In the first stage, the full dataset undergoes dimensionality reduction with UMAP and clustering via HDBSCAN to detect dense student cores. The second stage is executed exclusively over students classified as noise (unassigned in the previous step); reapplying the algorithm allows isolating secondary profiles previously hidden in the total dataset. To mitigate the ceiling effect without suppressing between-group contrast, a between-cluster normalization anchored on the empirical range of proficiency means (0.53 to 0.99) is applied. The method produced 10 cohesive profiles, covering 91.1% of the students, with high internal consistency and revealing conceptual dependencies across topics. The proposed approach shows promising potential in mapping pedagogically actionable profiles for targeted interventions.*

Resumo. *Técnicas de Rastreamento de Conhecimento (KT) permitem acompanhar a proficiência dos estudantes, mas transformá-las em perfis interpretáveis ainda é um desafio em larga escala. Este artigo apresenta um método de clusterização baseado em densidade em dois estágios para identificar perfis de estudantes e suas lacunas de aprendizagem a partir de dados de KT. O método foi avaliado com 19.048 estudantes de Cálculo I. No primeiro estágio, o conjunto completo passa por uma redução de dimensionalidade com UMAP e agrupamento via HDBSCAN para detectar os núcleos densos de estudantes. O segundo estágio é executado exclusivamente sobre os estudantes classificados como ruído (não atribuídos no estágio anterior); a reaplicação do algoritmo permite isolar perfis secundários antes ocultos no conjunto total. Para mitigar*

o efeito teto sem suprimir o contraste intergrupos, aplica-se uma normalização inter-cluster ancorada na amplitude real das médias de proficiência (0,53 a 0,99). O método produziu 10 perfis coesos, cobrindo 91,1% dos estudantes, com alta consistência interna e revelando dependências conceituais entre os tópicos. A abordagem apresenta potencial promissor no mapeamento de perfis pedagogicamente acionáveis para futuras intervenções direcionadas.

1. Introdução

Dados de interação gerados por plataformas digitais de ensino, quando analisados com técnicas de *Learning Analytics*, apresentam potencial significativo para revelar padrões de aprendizagem e fundamentar intervenções pedagógicas baseadas em evidências [Maschio et al. 2018]. Por exemplo, modelos de Rastreamento de Conhecimento (*Knowledge Tracing* – KT) estimam continuamente a probabilidade de domínio (i.e., *mastery* ou proficiência) de cada estudante sobre Componentes de Conhecimento (*Knowledge Components* – KCs) específicos [Corbett and Anderson 1994]. Contudo, extrair perfis interpretáveis a partir dessas estimativas contínuas para orientar ações institucionais personalizadas continua sendo um desafio em larga escala [Neto et al. 2021].

Uma possível solução para auxiliar na interpretação dos dados são algoritmos de clusterização, como o *K-means* e o HDBSCAN [Oliveira et al. 2022]. Esses algoritmos podem auxiliar na extração de perfis interpretáveis a partir das trajetórias de aprendizagem estimadas por modelos de KT, agrupando estudantes com padrões de proficiência semelhantes em *clusters* semanticamente coerentes. No entanto, uma das limitações na clusterização de estimativas de KT com algoritmos baseados em centroides, como o *K-means*, é a tendência de gerar agrupamentos esparsos e heterogêneos [Dutt et al. 2015]. Como esses algoritmos forçam a alocação de todos os indivíduos, os *clusters* resultantes podem agregar estudantes com perfis de proficiência e ritmos de estudo divergentes. Essa dispersão geométrica tende a comprometer o propósito central da personalização discente, dificultando a estruturação de intervenções pedagógicas coordenadas e focadas em necessidades homogêneas [Tapalova and Zhiyenbayeva 2022].

Uma alternativa à clusterização baseada em centroides é a abordagem baseada em densidade. Algoritmos como o HDBSCAN contornam as limitações dos métodos baseados em centroides ao isolar núcleos de alta similaridade e classificar os demais pontos como ruído [Campello et al. 2013]. A consequência dessa especificidade, porém, é um potencial descarte de dados, uma vez que uma parcela expressiva de estudantes pode ficar sem diagnóstico ou cobertura pedagógica por não ser alocada a nenhum grupo [Dutt et al. 2015]. Soma-se a isso o fato de que avaliações remotas frequentemente sofrem de efeito teto, concentrando as estimativas de proficiência no topo da escala, o que pode mascarar assimetrias conceituais e aproximar de forma artificial perfis estudantis que possuem necessidades pedagógicas distintas [Chyung et al. 2020].

Para mitigar essas limitações, este trabalho propõe um *pipeline* não supervisionado em dois estágios para a identificação de perfis de aprendizagem e de lacunas relativas de conhecimento em larga escala. A partir das estimativas geradas por um modelo de KT, o método combina redução de dimensionalidade via UMAP [Becht et al. 2019] e aplicações sequenciais do algoritmo HDBSCAN. O primeiro estágio isola os núcleos discentes densos e coesos; o segundo atua exclusivamente sobre o ruído remanescente,

revelando perfis secundários que seriam ignorados em análises convencionais de estágio único. Por fim, uma normalização global das médias inter-*cluster* visa mitigar o efeito teto, ampliando o contraste visual e pedagógico das vulnerabilidades de aprendizagem.

Neste contexto, as principais contribuições deste artigo são:

- Um *pipeline* em dois estágios de clusterização por densidade que reduz o descarte de estudantes (*noise*) em bases contínuas de rastreamento de conhecimento;
- Uma abordagem de normalização inter-*cluster* que recupera a interpretabilidade pedagógica das lacunas relativas de conhecimento sob a presença de efeito teto;
- A validação de consistência interna e de construto do método proposto utilizando dados reais de 19.048 estudantes da disciplina de Cálculo Diferencial e Integral I.

2. Trabalhos Relacionados

Considerando o uso de algoritmos de agrupamento para a identificação de características educacionais digitais, [Oliveira et al. 2022] utilizaram o *K-means* para agrupar o desempenho acadêmico de 6.500 estudantes, combinado às interações em um Ambiente Virtual de Aprendizagem (AVA), visando identificar fatores que auxiliem na prevenção da evasão escolar. A partir dos resultados, os autores identificaram perfis de engajamento distintos graças à clusterização. Contudo, essa análise baseou-se no desempenho acumulado geral dos estudantes, desconsiderando a progressão contínua e as lacunas dinâmicas de aprendizagem mapeadas por modelos de KT.

Semelhantemente, [Ramos et al. 2020] utilizaram o *K-means* e o agrupamento hierárquico para identificar perfis de estudantes baseados em dados de interação em fóruns e sistemas de uma universidade pública brasileira. A partir das análises, foram identificados três grandes grupos (baixa, média e alta interação). No entanto, de forma análoga a [Oliveira et al. 2022], o estudo não considerou a proficiência evolutiva dos estudantes gerada por modelos de KT. Além disso, a análise baseou-se em uma amostra de centenas de estudantes, o que limita a representatividade e a generalização dos achados para bases populacionais de larga escala [Wohlin et al. 2012].

No cenário internacional de Mineração de Dados Educacionais (*Educational Data Mining* – EDM), a aplicação de algoritmos de agrupamento é amplamente consolidada para caracterizar trajetórias e subsidiar a personalização de conteúdos [Dutt et al. 2015, Abdelrahman et al. 2023]. Contudo, as abordagens baseadas em centroides (como o *K-means*) sofrem com a sensibilidade a *outliers* e forçam a alocação de estudantes com perfis atípicos, gerando agrupamentos artificiais e altamente heterogêneos. Por outro lado, algoritmos baseados em densidade (como o HDBSCAN), embora preservem a coesão dos perfis, resultam em um descarte massivo de dados (*noise*), deixando uma parcela significativa de alunos sem diagnóstico [Dutt et al. 2015].

A combinação entre estimativas de KT e clusterização é promissora [Abdelrahman et al. 2023], mas extrair perfis interpretáveis e inclusivos em larga escala continua sendo um desafio em aberto [Neto et al. 2021]. Diante dessas lacunas, este trabalho investiga um *pipeline* em dois estágios projetado para resgatar o ruído e aplica uma normalização global inter-*cluster* para restituir o contraste pedagógico afetado pelo efeito teto. Essa formulação é estruturada e avaliada sob a ótica de três questões de pesquisa (QP).

2.1. Questões de Pesquisa

QUESTÃO DE PESQUISA 1 (QP1):

Em que medida as estimativas de proficiência por componente de conhecimento, geradas por um modelo de KT, oferecem base discriminativa adequada para métodos não supervisionados de clusterização?

Modelos de KT estimam a evolução do aprendizado ao longo do tempo e têm ampla validação em sistemas tutores inteligentes. Contudo, sua aplicação em bases de dados massivas impõe desafios de escalabilidade, tratamento de esparsidade e generalização para populações heterogêneas [Corbett and Anderson 1994], e poucos estudos examinam se as saídas contínuas desses modelos servem como preditores confiáveis para algoritmos não supervisionados [Abdelrahman et al. 2023]. Investigamos, portanto, não a qualidade preditiva do KT em si, mas se suas estimativas oferecem base discriminativa útil para a clusterização.

QUESTÃO DE PESQUISA 2 (QP2):

Como o uso de algoritmos de clusterização baseados em densidade em dois estágios, aplicados às estimativas de KT, permite mitigar o ruído e identificar perfis de proficiência coesos?

Algoritmos como o HDBSCAN dispensam a especificação prévia do número de grupos, detectam geometrias arbitrárias e tratam dados ruidosos, características promissoras para o contexto educacional [Costa et al. 2025]. Todavia, ao lidar com saídas de KT, a continuidade das trajetórias acadêmicas frequentemente faz com que uma proporção elevada de estudantes seja rotulada como ruído. Para contornar essa limitação, propõe-se o agrupamento em dois estágios: primeiro sobre a base completa e, sequencialmente, sobre os estudantes não alocados. Essa estratégia visa resgatar indivíduos dispersos sem fragmentar os núcleos densos iniciais, consolidando perfis de proficiência pedagogicamente acionáveis.

QUESTÃO DE PESQUISA 3 (QP3):

Em que medida a normalização global das médias de proficiência entre clusters permite mitigar o efeito teto e revelar lacunas relativas de aprendizagem?

O efeito teto, caracterizado pela compressão das proficiências no topo da escala, decorre comumente de exames pouco discriminativos, avaliações remotas ou vieses de amostragem [Chyung et al. 2020]. Na análise de perfis, esse fenômeno aglutina as médias dos grupos próximas ao limite máximo, ocultando contrastes conceituais intergrupos [Abdelrahman et al. 2023]. A abordagem proposta aplica a normalização Min-Max sobre a matriz de médias dos *clusters*, reescalando-as para o intervalo $[0, 1]$. Esta questão avalia o potencial dessa transformação em revelar lacunas de conhecimento que permaneceriam imperceptíveis sob métricas absolutas [Chyung et al. 2020].

3. Método

Este estudo utiliza dados secundários de interações discentes (videoaulas, questionários e conteúdos textuais) coletados em uma plataforma digital de ensino superior online brasileira. O modelo de avaliação das disciplinas adota aprovação com nota mínima de 60

pontos em uma escala de 100, estruturada em prova final somativa (50% da nota), engajamento com materiais didáticos (30%) e questionários formativos assíncronos com múltiplas tentativas permitidas (20%).

Os dados de interação analisados compreendem uma base com 19.048 estudantes e 25 KCs da disciplina de *Cálculo Diferencial e Integral I* (CDI-I). Para fins deste estudo, o modelo de KT é adotado como um componente independente de geração de dados. Conceitualmente, esse modelo monitora de forma contínua a probabilidade de domínio (*mastery* ou proficiência) de cada estudante em cada componente do conhecimento. A arquitetura do modelo de KT subjacente adota o *framework* SAINT com codificação posicional rotativa de tempo e ordem (*Time-and-Order Rotational Positional Encoding* – TO-RoPE) [Costa et al. 2026]. Essa modelagem trata a sequência e o tempo contínuo como propriedades geométricas do espaço latente, mitigando limitações de extrapolação. Sua alta acurácia preditiva foi previamente homologada e validada em larga escala [Costa et al. 2026], fornecendo as estimativas de proficiência que alimentam o nosso *pipeline* de clusterização.

Essas estimativas foram estruturadas diretamente em uma matriz de proficiência $M \in \mathbb{R}^{n \times p}$, onde n representa o total de estudantes e p a quantidade de KCs. Devido à interação de todos os estudantes com a totalidade dos KCs, a matriz resultante não apresenta esparsidade, o que dispensou o uso de técnicas de imputação de valores ausentes. Em seguida, os valores foram normalizados via Min-Max no intervalo $[0, 1]$. Descartou-se a padronização por Z-Score para evitar que a equalização artificial das variâncias distorcesse as distâncias geométricas de proficiência discente.

3.1. Redução de Dimensionalidade

A alta dimensionalidade da matriz M (p KCs) pode degradar o desempenho de algoritmos de clusterização baseados em densidade devido à compressão das distâncias métricas, fenômeno conhecido como a maldição da dimensionalidade [Bellman 1957]. Para contornar essa limitação, aplicou-se o método *Uniform Manifold Approximation and Projection* (UMAP) [McInnes et al. 2018]. A escolha do UMAP, em detrimento de técnicas tradicionais como PCA e t-SNE, fundamenta-se em sua capacidade de preservar simultaneamente as estruturas locais e globais do conjunto de dados. Enquanto o PCA falha em capturar relações não lineares típicas das trajetórias de aprendizagem, o t-SNE tende a desorganizar o posicionamento global dos elementos em favor de agrupamentos locais.

No Estágio 1, o UMAP projeta a base completa para mapear as relações globais de similaridade de proficiência. No Estágio 2, uma nova projeção independente é executada exclusivamente sobre o subconjunto de ruído remanescente. Ao remover as populações densas dominantes do Estágio 1, reconfigura-se o espaço latente em torno da variação *interna* do ruído, revelando subgrupos secundários que estavam geometricamente ocultos no espaço global.

A dimensionalidade do UMAP foi fixada em 5 componentes, adotando $min_dist = 0$ para maximizar a compactação local de áreas densas, conforme recomendado para tarefas de agrupamento [McInnes et al. 2018]. Os demais hiperparâmetros foram selecionados via busca em grade (*grid search*) independente, conforme detalhado na Tabela 1.

Tabela 1. Espaço de busca da busca em grade para o UMAP.

Hiperparâmetro	Valores explorados
<i>n_neighbors</i>	50, 100, 150, 200, 250
<i>min_dist</i>	0.00, 0.01, 0.05
<i>metric</i>	euclidiana, cosseno

3.2. Clusterização em Dois Estágios

Para a identificação dos perfis de aprendizagem, adotou-se o HDBSCAN [Campello et al. 2013]. Porém, a aplicação direta desse algoritmo às estimativas de proficiência tende a produzir uma proporção elevada de ruído devido ao caráter contínuo do aprendizado [Abdelrahman et al. 2023]. Para mitigar esse descarte sem fragmentar os agrupamentos principais, propõe-se uma **estratégia de busca iterativa em dois estágios**:

1. **Estágio 1 — Núcleos densos:** O algoritmo HDBSCAN é aplicado às coordenadas de todos os estudantes na projeção de cinco dimensões. Configura-se o algoritmo com parâmetros restritivos para priorizar a coesão dos grupos principais. Os indivíduos não alocados a nenhum *cluster* são retidos temporariamente sob o rótulo de ruído.
2. **Estágio 2 — Resgate do ruído:** Uma nova projeção é gerada exclusivamente sobre os estudantes não atribuídos no Estágio 1, reconfigurando o espaço em torno da variação interna desse subconjunto. O HDBSCAN é então aplicado a essa projeção, discriminando subgrupos estáveis que eram geometricamente indiferenciáveis no espaço global. Os novos *clusters* são reindexados sequencialmente para evitar colisões com os do Estágio 1.

A parametrização do HDBSCAN envolveu uma busca em grade independente em cada estágio, explorando o método de seleção (`method` $\in \{eom, leaf\}$), vizinhança mínima (`min_samples` $\in \{5, 10, 15\}$), limiar de distância (`cluster_selection_epsilon` $\in \{0, 00, 0, 01, 0, 02, 0, 05, 0, 10\}$) e tamanho mínimo do cluster (`min_cluster_size` $\in \{1\%, 2, 5\%, 5\%, 10\%\}$). No algoritmo, o `min_cluster_size` define o tamanho mínimo para formação de um grupo, o `min_samples` controla a sensibilidade ao ruído, os métodos *leaf* e *eom* ditam a seleção dos *clusters* (foco em folhas homogêneas vs. estabilidade global) e o `cluster_selection_epsilon` (ϵ) evita a fragmentação excessiva ao fundir grupos geometricamente muito próximos.

Para o `min_cluster_size`, estabeleceu-se adicionalmente um piso de 30 indivíduos. Essa escolha justifica-se pedagogicamente por equivaler ao tamanho médio de uma turma de graduação no ensino superior brasileiro, assegurando viabilidade operacional para ações institucionais coordenadas (como alocação de tutorias).

3.3. Avaliação, Seleção e Normalização das Médias

A seleção dos modelos ótimos fundamentou-se na métrica DBCV (*Density-Based Clustering Validation*) [Moulavi et al. 2014], calculada de forma independente para os *clusters* de cada estágio no espaço de cinco componentes do UMAP. A métrica de Silhouette [Rousseeuw 1987] e o índice Davies-Bouldin (DB) [Davies and Bouldin 1979]

foram extraídos de forma complementar, diretamente no mesmo *embedding*, além do monitoramento da taxa de ruído residual. Para fins de reprodutibilidade, fixou-se `random_state=42` e `n_jobs=1` na execução do UMAP.

As estimativas absolutas de proficiência concentraram-se majoritariamente próximas ao teto da escala, possivelmente porque as avaliações de CDI-I ocorreram em regime remoto e sem supervisão direta, o que reduziu a variância e o poder discriminatório das pontuações originais. Na seção de Resultados, observa-se que a menor média absoluta de proficiência registrada entre os *clusters* criados foi de 0,53 e a maior foi de 0,99. Embora o patamar inferior (0,53) represente pedagogicamente um domínio insuficiente para os critérios de aprovação, sua proximidade numérica com os valores superiores atenua o contraste matemático intergrupos, justificando a necessidade da normalização proposta.

Para neutralizar essa distorção e expor as assimetrias conceituais latentes, extraiu-se a proficiência média de cada *cluster* por KC e aplicou-se uma normalização Min-Max intergrupos diretamente sobre essa matriz de médias. Esse procedimento expande os valores observados para o intervalo completo $[0, 1]$, restabelecendo o contraste estrito necessário para expor as assimetrias conceituais latentes e isolar vulnerabilidades relativas de aprendizagem que estavam ocultas sob a alta pontuação absoluta global.

4. Resultados

As buscas em grade selecionaram os parâmetros consolidados na Tabela 2, que se dividem entre a projeção espacial e a clusterização por densidade. Os números distintos de vizinhos (`n_neighbors`) entre os estágios (50 no E1 e 150 no E2) refletem as diferentes escalas estruturais das duas subpopulações discentes analisadas.

Tabela 2. Configurações selecionadas pela busca em grade para a disciplina CDI-I.

Estágio	<code>n_neighbors</code>	<code>min_dist</code>	<code>metric</code>	<code>min_cluster_size</code>	<code>min_samples</code>	<code>method</code>	<code>cluster_selection_epsilon</code>
E1	50	0.00	euclidiana	476 (2.5%)	10	leaf	0.05
E2	150	0.00	euclidiana	446 (5.0%)	10	eom	0.05

4.1. Discriminabilidade das Estimativas e Eficiência do Agrupamento (QP1)

O aumento do tamanho mínimo de agrupamento (`min_cluster_size`) para 5,0% no Estágio 2 justifica-se pela esparsa distribuição geométrica do subconjunto de ruído remanescente. Em regiões de menor densidade, exige-se um limiar de agrupamento proporcionalmente maior para consolidar perfis estáveis, mitigando a proliferação de microgrupos espúrios no espaço latente. A evolução volumétrica da base e as respectivas métricas de qualidade do agrupamento por estágio estão consolidadas na Tabela 3.

Os resultados revelam que o Estágio 1 isolou seis *clusters* altamente coesos, mas rotulou 46,9% da base discente como ruído pedagógico. Ao reconfigurar o espaço latente sobre esse grupo, o Estágio 2 resgatou 7.237 estudantes em quatro novos agrupamentos estáveis, reduzindo o descarte populacional final para apenas 8,9% (1.698 alunos) e alcançando uma cobertura de 91,1% da turma.

As métricas de qualidade (Tabela 3) validam estatisticamente a robustez do *pipeline*: os coeficientes DBCV positivos confirmam gradientes internos de densidade,

Tabela 3. Evolução volumétrica e métricas de qualidade dos agrupamentos por estágio.

Estágio	Alunos Proc.	Clusters (k)	Alunos Clust.	Ruído	DBCV	Silhouette	DB
E1	19.048	6	10.113	8.935 (46,9%)	0,265	0,314	0,744
E2	8.935	4	7.237	1.698 (19,0%*)	0,445	0,450	0,901
Final	19.048	10	17.350	1.698 (8,9%)	–	–	–

*Percentual calculado sobre os 8.935 estudantes processados no respectivo estágio.

o índice de Davies-Bouldin abaixo de 1,0 indica separação no espaço latente e os valores de Silhouette alinham-se ao esperado para geometrias arbitrárias do HDBSCAN [Campello et al. 2013]. Essa coesão matemática sugere que o ruído preliminar do Estágio 1 não representava apenas dados dispersos aleatórios, mas sim subpopulações estáveis com comportamentos de proficiência homogêneos, cujas distribuições detalhadas por KC constam no mapa de calor absoluto (Figura 2(a)).

4.2. Perfis Pedagógicos (QP2)

Os 10 perfis estabelecem um espectro hierárquico de desempenho, estendendo-se dos grupos de alta proficiência ($C2_E1$ e $C0_E2$) até aqueles de acentuada vulnerabilidade ($C4_E1$ e $C1_E2$). Para sanar as limitações de legibilidade visual decorrentes da densidade de KCs nas figuras de mapeamento, apresenta-se na Tabela 4 uma síntese estruturada que caracteriza cada agrupamento discente e suas respectivas vulnerabilidades conceituais dominantes.

Tabela 4. Caracterização sintética dos perfis pedagógicos identificados na disciplina de CDI-I.

Perfil / Cluster	Estágio	Estudantes (n)	% Base	Caracterização Pedagógica e Lacunas de Conhecimento (KCs)
$C2_E1$	E1	4.191	22,0%	Altíssima proficiência global; domínio pleno e homogêneo de todos os KCs.
$C0_E2$	E2	3.242	17,0%	Alta proficiência global; leve queda apenas em KCs de derivação implícita.
$C1_E1$	E1	755	4,0%	Bom desempenho geral; vulnerabilidade severa na Regra de L'Hôpital.
$C0_E1$	E1	554	2,9%	Perfil atípico; erro de Guttman (domínio de concavidade sem pré-requisitos).
$C2_E2$	E2	1.092	5,7%	Desempenho intermediário-alto; vulnerabilidade em Regra de L'Hôpital e limites.
$C3_E2$	E2	908	4,8%	Desempenho intermediário; oscilação e lacunas pontuais em derivadas.
$C3_E1$	E1	942	4,9%	Desempenho intermediário-baixo; vulnerabilidades dispersas em limites e derivadas.
$C5_E1$	E1	491	2,6%	Lacunas severas em KCs básicos (definição de limite, assíntotas e continuidade).
$C1_E2$	E2	1.995	10,5%	Lacunas severas em KCs básicos (limites) e regras iniciais de diferenciação.
$C4_E1$	E1	3.172	16,7%	Altíssima vulnerabilidade global; baixíssima proficiência em todos os KCs da disciplina.

A inspeção da matriz normalizada revela duas topologias principais de lacunas conceituais discentes: (i) **Dificuldades em KCs básicos** ($C5_E1$ e $C1_E2$), caracterizadas por baixa proficiência em conceitos fundamentais como *assíntotas horizontais e verticais*, *definição de limite* e *definição de função contínua*, fragilidades estruturais que comprometem o avanço no restante da disciplina e que estavam camufladas pelas estimativas absolutas; (ii) **Dificuldades em KCs avançados** ($C1_E1$ e $C2_E2$), em que os estudantes apresentam bom aproveitamento nos fundamentos de cálculo, mas registram quedas acentuadas em tópicos como *formas indeterminadas* e *regra de L'Hôpital*, demandando suporte focado nessas ferramentas específicas; e (iii) **Comportamento Atípico** ($C0_E1$), caracterizado pelo alto domínio do *teste da segunda derivada para concavidade* na ausência de competências preliminares, assemelhando-se aos erros de Guttman [Meijer 1994, Reichenheim et al. 2026] e sugerindo quebras locais na hierarquia pedagógica esperada.

A distribuição da proficiência global por *cluster*, ordenada pela mediana (Figura 1), confirma esse espectro e evidencia que os agrupamentos de ambos os estágios se intercalam em diferentes níveis de proficiência.

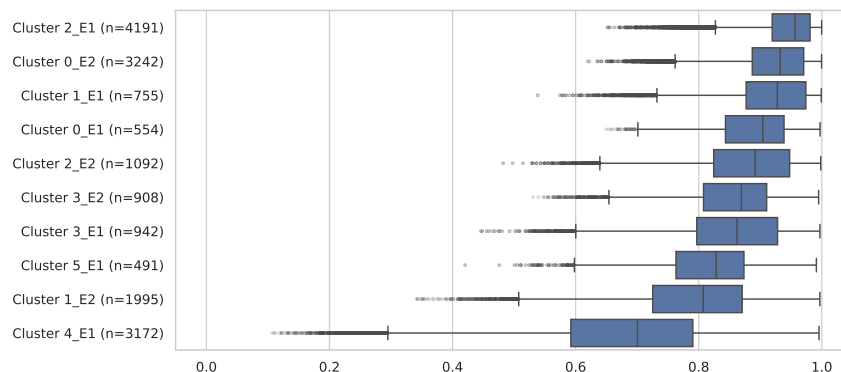


Figura 1. Distribuição da proficiência global por *cluster*, com ordenação crescente pelas medianas de proficiência discente.

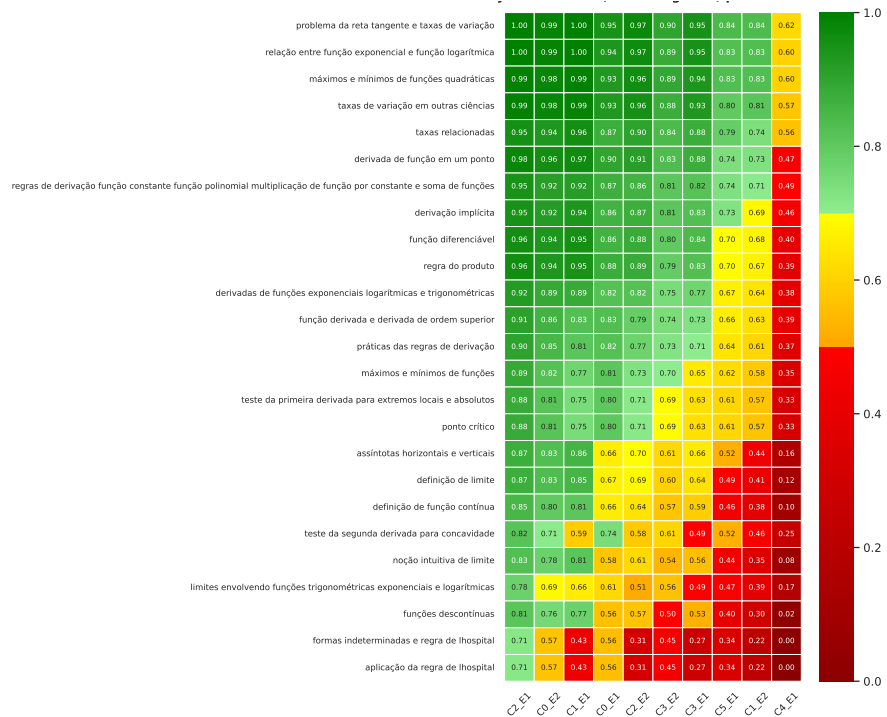
4.3. Normalização das Variáveis (QP3)

Os mapas de calor de proficiência (Figura 2) evidenciam o impacto prático da normalização na legibilidade pedagógica dos dados. No mapa absoluto (Figura 2(a)), o adensamento dos valores na faixa estreita entre 0,53 e 0,99 ilustra a compressão decorrente do efeito teto, gerando tonalidades homogêneas e baixíssimo contraste visual para diferenciar o domínio dos KCs entre os grupos.

Após a aplicação da normalização global inter-*cluster* (Figura 2(b)), as lacunas relativas de aprendizagem tornam-se plenamente visíveis. O reescalonamento das médias com base na amplitude empírica preserva o contraste original e revela quedas de desempenho em KCs específicos dentro de subpopulações que antes apresentavam médias absolutas nominalmente altas, explicitando variações de conteúdo ocultas nos valores absolutos institucionais.



(a) Heatmap de proficiência absoluto por cluster.



(b) Heatmap de proficiência normalizado.

Figura 2. Comparativo dos perfis de proficiência por cluster antes e após a normalização global para mitigação do efeito teto.

5. Discussão

Os resultados obtidos fornecem indícios consistentes sobre a relevância de converter as trajetórias contínuas de estimativas de KT em perfis discretos pedagogicamente acionáveis. Os achados são discutidos a seguir à luz das três questões de pesquisa (QP) investigadas.

Na QP1, os dados sugerem que a ausência de esparsidade na matriz M combinada à projeção espacial do UMAP gerou um espaço latente no qual as distâncias geométricas refletem de forma consistente a similaridade conceitual dos estudantes. A segmentação de grupos com trajetórias pedagógicas divergentes pelo HDBSCAN apoia a premissa de que as estimativas geradas pelo modelo de KT preservam a variância e o sinal estatístico necessários para fundamentar algoritmos baseados em densidade geométrica, indicando viabilidade para análises em larga escala.

A QP2 investiga a **topologia e a cobertura dos dados**. A arquitetura em dois estágios proposta buscou contornar uma limitação geométrica intrínseca ao HDBSCAN: o descarte elevado de indivíduos em dados contínuos. O Estágio 1 do *pipeline* é matematicamente equivalente à aplicação clássica do HDBSCAN de estágio único (nosso primeiro *baseline* comparativo). Sob essa configuração convencional, as restrições de densidade exigiriam classificar 46,9% da base total de estudantes como ruído, o que é pedagogicamente inviável para ações coordenadas de nivelamento. O acoplamento do Estágio 2 resolveu essa restrição geométrica ao reconfigurar o espaço latente sobre esse contingente disperso, resgatando 7.237 discentes em quatro novos perfis altamente coesos e reduzindo o descarte populacional final para apenas 8,9%.

Essa dinâmica de agrupamento em dois estágios contrasta vantajosamente com algoritmos baseados em centroides, como o *K-means* (utilizado em [Oliveira et al. 2022, Ramos et al. 2020]). Embora o *K-means* ofereça 100% de cobertura populacional ao forçar a alocação de todas as instâncias, essa imposição geométrica gera agrupamentos com altíssima variância interna, misturando estudantes com ritmos e dificuldades pedagógicas divergentes em um mesmo grupo, o que tende a comprometer a eficácia de intervenções personalizadas.

A QP3 foca na **interpretabilidade pedagógica e no contraste métrico**. Enquanto a QP2 define a composição discente de cada grupo, a QP3 auxilia a determinar como esses perfis se diferenciam na prática. Diante do achatamento métrico causado pelo regime de avaliações remotas [Chyung et al. 2020], a normalização global proposta expandiu o intervalo original de $[0, 53; 0, 99]$ para $[0, 1]$, ampliando o contraste matemático para expor as assimetrias conceituais latentes [Abdelrahman et al. 2023]. Sem isso, perfis com médias absolutas altas seriam interpretados como regulares, mascarando vulnerabilidades críticas que os valores absolutos ocultavam [Chyung et al. 2020].

A integração dessas soluções gera implicações práticas relevantes para a gestão de aprendizagem e a tomada de decisões institucionais. Como o efeito teto se manifestou nas notas homologadas, a validação do *pipeline* fundamentou-se na consistência topológica do construto e na identificação de dependências conceituais. Enquanto os indicadores administrativos e operacionais sugeriam um cenário de aprovação e desempenho homogêneos (médias absolutas acima de 0, 53), a normalização inter-*cluster* revelou que o grupo *CI_E2* acumulava vulnerabilidades severas em KCs estruturais

básicos (como *definição de limite e função contínua*), ao passo que o *C0_E1* apresentava quebras locais de pré-requisitos características de erros de Guttman [Meijer 1994, Reichenheim et al. 2026].

A emersão desses perfis customizados confere indícios de validade de construto à abordagem baseada em dados secundários descaracterizados [Wohlin et al. 2012], indicando sua capacidade de sinalizar vulnerabilidades de aprendizagem camufladas pelo efeito teto [Chyung et al. 2020]. Para a coordenação de CDI-I, esse mapeamento sistemático oferece subsídios práticos para estruturar ações de nivelamento direcionadas e mitigar o risco de evasão tardia ou reprovação em disciplinas subsequentes de alta dependência conceitual, como Cálculo II, aproximando os indicadores administrativos de desempenho do domínio cognitivo real dos estudantes. Contudo, ressalta-se o caráter diagnóstico-exploratório do presente estudo; a comprovação da eficácia pedagógica direta dessas trilhas personalizadas na aprendizagem discente requer a condução de ensaios controlados aleatorizados (*Randomized Controlled Trials* – RCTs), os quais estão mapeados como trabalhos futuros.

6. Limitações e considerações finais

Este estudo apresenta limitações relativas ao escopo dos achados. A normalização Min-Max inter-*cluster* aplicada atua de forma relativa sobre a matriz de médias; logo, atribuições de baixo domínio indicam desvantagem frente ao grupo discente contemporâneo, e não lacunas absolutas, visto que o menor patamar absoluto observado foi de 0,53. Sob o efeito teto, esse mapeamento relativo ganha importância prática ao atuar como indicador antecedente de defasagem oculta antes do avanço curricular discente. Ademais, a análise restringe-se a uma única disciplina de exatas com avaliações assíncronas. Como trabalhos futuros, projeta-se aplicar o *pipeline* em disciplinas de outras áreas do conhecimento para testar sua reprodutibilidade, submeter o ecossistema a testes comparativos frente a outros algoritmos de agrupamento baseados em densidade e conduzir ensaios controlados aleatorizados (RCTs) para avaliar empiricamente o impacto pedagógico de intervenções personalizadas de nivelamento baseadas nos perfis mapeados.

Cuidados Éticos

Este estudo utiliza dados secundários institucionalmente anonimizados, estando dispensado de avaliação pelo sistema CEP/CONEP conforme a Resolução CNS 510/2016 (Art. 1º, Parágrafo Único, Inciso V) e em estrita conformidade com a LGPD.

Agradecimentos

Os autores agradecem à COGNA pelo suporte institucional e viabilização técnica deste estudo.

Declaração sobre o Uso de Inteligência Artificial

Para além do escopo metodológico explicitado nos experimentos, ferramentas de Inteligência Artificial Generativa foram utilizadas estritamente no auxílio à redação, refinamento gramatical e organização textual deste documento. Os autores revisaram integralmente o texto gerado e assumem total responsabilidade científica e editorial pelo conteúdo final.

Referências

- Abdelrahman, G., Wang, Q., and Nunes, B. (2023). Knowledge tracing: A survey. *ACM Computing Surveys*, 55(11):1–37.
- Becht, E., McInnes, L., Healy, J., Dutertre, C.-A., Kwok, I. W. H., Ng, L. G., Ginhoux, F., and Newell, E. W. (2019). Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology*, 37(1):38–44.
- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining (PAKDD)*, pages 160–172. Springer.
- Chyung, S. Y., Hutchinson, D., and Shamsy, J. A. (2020). Evidence-based survey design: Ceiling effects associated with response scales. *Performance Improvement*, 59(6):6–13.
- Corbett, A. T. and Anderson, J. R. (1994). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modeling and User-Adapted Interaction*, 4:253–278.
- Costa, J. A. R., Lima, G. D. O., Araújo, R. D., and Dorça, F. A. (2025). Exploring self-regulated learning in virtual environments: An experimental clustering-based approach. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 469–482. SBC.
- Costa, W., Pereira, F. D., Alves, G., Nascimento, A., Rodrigues, L., Santos, M. A. P., Lima, T. M., and Mello, R. F. (2026). Time as a geometric property of the latent space: An empirical study of Time-and-Order RoPE in Knowledge Tracing. In *Proceedings of the 19th International Conference on Educational Data Mining*, pages 317–328, Seoul, Republic of Korea. International Educational Data Mining Society.
- Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227.
- Dutt, A., Aghabozrgi, S., Ismail, M. A. B., and Mahrooeian, H. (2015). Clustering algorithms applied in educational data mining. *International Journal of Information and Electronics Engineering*, 5(2):112.
- Maschio, P., Vieira, M. A., Costa, N., de Melo, S., and Júnior, C. P. (2018). Um panorama acerca da mineração de dados educacionais no brasil. In *Anais do XXIX Simpósio Brasileiro de Informática na Educação*. SBC.
- McInnes, L., Healy, J., and Melville, J. (2018). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.
- Meijer, R. R. (1994). The number of guttman errors as a simple and powerful person-fit statistic. *Applied Psychological Measurement*, 18(4):311–314.
- Moulavi, D., Jaskowiak, P. A., Campello, R. J. G. B., Zimek, A., and Sander, J. (2014). Density-based clustering validation. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 839–847. SIAM.
- Neto, M. V. G., Vasconcelos, G. C., and Zanchettin, C. (2021). Mineração de dados aplicada à predição do desempenho de escolas e técnicas de interpretabilidade dos

- modelos. In *Simpósio Brasileiro de Informática na Educação (SBIE)*, pages 773–782. SBC.
- Oliveira, P. L. S. d., Rodrigues, R. L., Ramos, J. L. C., and Silva, J. C. S. (2022). Identificação de pesquisas e análise de algoritmos de clusterização para a descoberta de perfis de engajamento. *Revista Brasileira de Informática na Educação*, 30:1–19.
- Ramos, J., Santos, L., Silva, J., and Rodrigues, R. (2020). Identificação de perfis de interação de estudantes de educação a distância por meio de técnicas de agrupamentos. In *Anais do XXXI Simpósio Brasileiro de Informática na Educação*, pages 932–941, Porto Alegre, RS, Brasil. SBC.
- Reichenheim, M. E., Moraes, C. L. d., and Bastos, J. L. (2026). Gráficos de erro de Guttman: uma abordagem visual para a análise de escalabilidade. *Revista de Saúde Pública*, 60:e1.
- Rousseeuw, P. J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65.
- Tapalova, O. and Zhiyenbayeva, N. (2022). Artificial intelligence in education: AIED for personalised learning pathways. *Electronic Journal of e-Learning*, 20(5):639–653.
- Wohlin, C., Runeson, P., Höst, M., Ohlsson, M. C., Regnell, B., and Wesslén, A. (2012). *Experimentation in Software Engineering*, volume 236. Springer.