

## Segmentação de perfis de participação em formações continuadas por clusterização a partir de microdados educacionais do SGF e Moodle

\*Leonardo Cruz de Andrade, Sandro Ricardo de Souza,  
Almir Carletti Neto, Sérgio Nery Simões

Instituto Federal do Espírito Santo (Ifes)  
Serra – ES – Brasil

leonardocdea@gmail.com, sandro.fisica@gmail.com,  
almir.carletti.neto@gmail.com, sergio@ifes.edu.br

**Abstract.** *This paper segments participation profiles in continuing professional development programs through clustering, integrating microdata from a Training Management System with Moodle behavioral traces. A two-route architecture combines dimensionality reduction (PCA) and K-means clustering across 191,412 enrollments. Route B (34,985 cases) yielded two profiles: Full Engagement (87.7%) and Intermediate Friction (12.3%). Route A yielded four profiles: Express Certification (45.3%), Chronic Dropout (38.2%), Organic Engagement (14.6%), and Resilience with Friction (1.8%). The study contributes an analytical strategy for segmenting participation in large-scale public education.*

**Resumo.** *Este artigo segmenta perfis de participação em formações continuadas por clusterização, integrando microdados do Sistema de Gestão da Formação com rastros comportamentais do Moodle. Uma arquitetura em duas rotas combina redução de dimensionalidade (PCA) e clusterização por K-means em 191.412 inscrições. A Rota B (34.985 casos) revelou dois perfis: Engajamento Pleno (87,7%) e Atrito Intermediário (12,3%). A Rota A revelou quatro perfis: Certificação Expressa (45,3%), Evasão Crônica (38,2%), Engajamento Orgânico (14,6%) e Resiliência com Fricção (1,8%). O trabalho contribui com uma estratégia analítica para segmentar a participação em redes públicas de ensino.*

### 1. Introdução

A análise da participação de profissionais da educação em ações de formação continuada, em redes públicas de grande escala, envolve dificuldades metodológicas específicas. Em contextos marcados por diversidade funcional, heterogeneidade territorial e uso crescente de plataformas digitais, indicadores administrativos isolados, como inscrição, acesso ou certificação, tendem a representar apenas parte do percurso formativo, limitando a compreensão de engajamento, permanência e desempenho [de La Hoz-Ruiz et al. 2024]. Esse problema ganha relevância à medida que a formação continuada assume papel estratégico nas políticas educacionais e passa a operar em ecossistemas institucionais orientados por dados. Sistemas de gestão concentram informações cadastrais, funcionais e administrativas, enquanto ambientes virtuais de aprendizagem registram rastros transacionais e comportamentais; a articulação analítica entre dados institucionais e rastros

---

\*Os autores agradecem à FAPES/UnAC (Nº FAPES 1068/2023 P 2023-SGLQ7, Nº SIA-FEM 2023-SGLQ7) pelo apoio financeiro dado por meio do Sistema UniversidadES.

digitais pode ampliar a compreensão dos modos de participação na política formativa [Aulakh et al. 2023, Paolucci et al. 2024, Gamage et al. 2022].

Neste contexto, o problema de pesquisa que este trabalho busca abordar é como segmentar e caracterizar os perfis de participação de profissionais da educação em programas de formação continuada, considerando a complexidade e a heterogeneidade dos dados disponíveis, e como esses perfis podem informar intervenções pedagógicas e de gestão mais direcionadas. A relevância deste estudo reside na capacidade de transformar grandes volumes de microdados institucionais em conhecimento acionável, subsidiando políticas educacionais mais eficazes e personalizadas [Cerezo et al. 2024, de La Hoz-Ruiz et al. 2024, Paolucci et al. 2024].

Para tanto, este trabalho analisa a segmentação de perfis de participação dos profissionais da educação em programas de formação continuada, construídos pelo Centro de Formação dos Profissionais da Educação do Espírito Santo (Cefope), institucionalizado pela Secretaria de Estado da Educação do Espírito Santo (Sedu/ES), que operacionaliza a política de formação continuada da rede pública estadual [ESPÍRITO SANTO. Secretaria da Educação 2026]. A análise integra microdados administrativos do Sistema de Gestão de Formação (SGF) e rastros transacionais do ambiente virtual de aprendizagem (Moodle) do Cefope, com base no processo de Descoberta de Conhecimento em Bases de Dados (KDD) e em técnicas de redução de dimensionalidade e clusterização, examinando padrões de engajamento, permanência, desempenho e distribuição territorial [Fayyad et al. 1996, Han et al. 2012, Pecuchová and Drlík 2024].

A principal contribuição do artigo para a Informática na Educação reside na proposição de arquitetura analítica em duas rotas para interpretar a heterogeneidade da participação em formações continuadas. A Rota A descreve a estrutura global da política, incorporando inscrições com baixa densidade observacional, enquanto a Rota B focaliza os casos com engajamento pedagógico registrado no ambiente virtual, permitindo identificar perfis ativos de participação e sua projeção territorial. Com isso, o trabalho busca ampliar a inteligibilidade analítica de bases educacionais institucionais e produzir subsídios baseados em evidências para pesquisa e gestão educacional orientada por dados [Cerezo et al. 2024, de La Hoz-Ruiz et al. 2024, Paolucci et al. 2024].

## 2. Fundamentos

A formação continuada de profissionais da educação constitui dimensão do desenvolvimento profissional, formalizada no campo normativo brasileiro pela Resolução CNE/CP nº 1/2020 (posteriormente revogada pela Resolução CNE/CP nº 4/2024) [BRASIL. Conselho Nacional de Educação 2020, BRASIL. Conselho Nacional de Educação 2024, Brito and Nunes 2025]. Em redes públicas de grande escala, essa discussão exige reconhecer a diversidade de públicos e contextos institucionais – professores, pedagogos e gestores participam a partir de trajetórias e condições de trabalho distintas [Gatti 2021, Brito and Nunes 2025] –, heterogeneidade que indicadores agregados como matrícula ou certificação tendem a ocultar [de La Hoz-Ruiz et al. 2024] e que a mediação tecnológica agrava, já que a participação observável depende de familiaridade com plataformas, condições de acesso e desenho pedagógico [Huang et al. 2024, Gamage et al. 2022], num contexto reforçado pelas transformações recentes da educação básica brasileira, como as mu-

danças no Ensino Médio (Lei n. 14.945/2024) e as diretrizes de integração curricular da educação digital [BRASIL 2024, BRASIL. Conselho Nacional de Educação 2022, BRASIL. Conselho Nacional de Educação 2025].

A ampliação do uso de plataformas digitais intensificou a produção de registros sobre ensino, aprendizagem e participação [Cerezo et al. 2024, Aulakh et al. 2023], dando origem a duas tradições centrais: a Mineração de Dados Educacionais (EDM), voltada à identificação de padrões e relações latentes por métodos computacionais e estatísticos, e a Learning Analytics (LA), que enfatiza a compreensão de trajetórias e o apoio à decisão em contextos educacionais [Baker and Yacef 2009, Siemens 2013] – abordagens complementares, já que a primeira modela e descobre padrões e a segunda interpreta esses resultados em função de finalidades pedagógicas e institucionais [Siemens 2013, Cerezo et al. 2024]. Nesse quadro, a Ciência de Dados Educacionais (EDS) opera como enquadramento conceitual mais abrangente, articulando tratamento e modelagem de dados a exigências de inteligibilidade pedagógica e permitindo integrar registros administrativos, atributos territoriais e rastros comportamentais em uma mesma arquitetura analítica [Cerezo et al. 2024, Paolucci et al. 2024] – convergência que sustenta o presente estudo, ao requerer tanto mecanismos de descoberta de estruturas latentes [Fayyad et al. 1996, Han et al. 2012] quanto critérios de interpretação relacionados a funções profissionais, territórios e uso das plataformas [Cerezo et al. 2024].

A operacionalização dessa perspectiva segue o paradigma da Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* – KDD), adequado a contextos educacionais em que os registros, produzidos por rotinas administrativas, pedagógicas e tecnológicas distintas, exigem tratamento prévio à modelagem [Fayyad et al. 1996, Han et al. 2012, Aulakh et al. 2023]; a redução de dimensionalidade – via Análise de Componentes Principais (PCA), que condensa a variância da base em dimensões ortogonais e reduz colinearidades [Greenacre et al. 2022, Borges et al. 2018, Pecuchová and Drlík 2024] – prepara o espaço para a etapa de clusterização [Han et al. 2012, Ramos et al. 2020, Pecuchová and Drlík 2024], na qual o K-means foi selecionado por seu equilíbrio entre simplicidade computacional, escalabilidade e interpretabilidade via centróides, em detrimento de alternativas como agrupamento hierárquico (custoso em grandes bases) e DBSCAN (parametrização complexa, menor interpretabilidade), com adequação validada internamente pelo Método do Cotovelo, coeficiente de silhueta e pelos índices de Calinski–Harabasz [Caliński and Harabasz 1974] e Davies–Bouldin [Davies and Bouldin 1979].

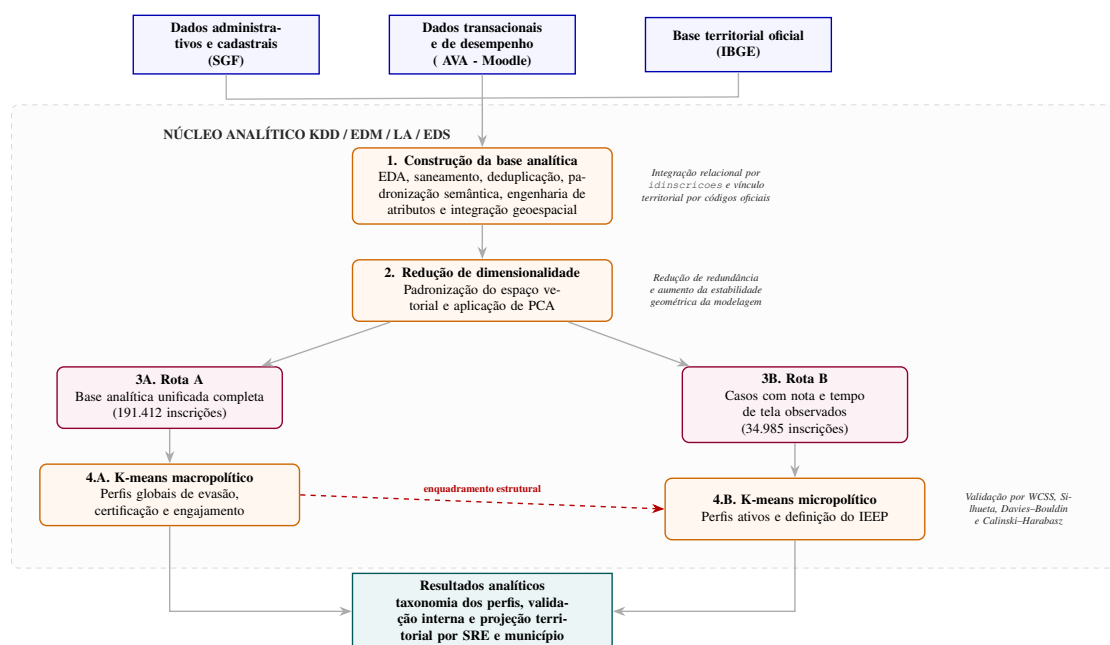
### 3. Metodologia

Esta pesquisa caracteriza-se como um estudo observacional, retrospectivo e analítico, de natureza exploratória e abordagem quantitativa, fundamentado na Mineração de Dados Educacionais (EDM) [Baker and Yacef 2009], na Learning Analytics (LA) [Siemens 2013] e na Ciência de Dados Educacionais (EDS) [Cerezo et al. 2024]. O percurso metodológico foi organizado segundo o paradigma da Descoberta de Conhecimento em Bases de Dados (KDD), com foco na identificação de padrões em registros secundários produzidos no sistema institucional de gestão da formação e no Ambiente Virtual de Aprendizagem Moodle [Fayyad et al. 1996, Han et al. 2012]. A população investigada compreendeu profissionais da educação da rede pública estadual do Espírito Santo, com inscrições em ofertas formativas realizadas entre 2023 e 2025. A

base analisada reuniu aproximadamente 80 mil participantes únicos e 191.412 inscrições válidas. A unidade observacional adotada foi a inscrição em oferta formativa, identificada por *idinscricoes*, chave derivada da concatenação entre o identificador desidentificado do profissional e o identificador normalizado da oferta, o que permitiu integrar as bases sem circulação de identificadores pessoais diretos no pipeline analítico [BRASIL 2018, BRASIL. Conselho Nacional de Saúde 2016].

A Figura 1 sintetiza o encadeamento das etapas de integração, redução de dimensionalidade, modelagem em duas rotas e projeção territorial adotadas no estudo. A construção da base analítica foi processada em ambiente *Python*<sup>1</sup> e estruturada em quatro etapas fundamentais.

**Figura 1. Arquitetura metodológica da pesquisa: integração de bases, modelagem em duas rotas e projeção territorial.**



**Fonte:** Elaborado pelos autores (2026).

Na primeira etapa, procedeu-se à análise exploratória e ao saneamento das bases. Na Base de Inscrições (SGF), variáveis operacionais como *Cargo*, *Funcao* e *Tipo\_Instituicao* apresentaram ausência simétrica de 1,50% (6.750 registros), sendo imputadas categoricamente como "NÃO INFORMADO / PÚBLICO EXTERNO" para preservar a volumetria. As variáveis *SRE* e *Município* foram tratadas via *Fuzzy Matching* com dados oficiais do IBGE. Na Base de Resultados (Moodle), detectou-se uma esparsidade crítica de 56,06% na variável *nota\_final\_na\_formacao*. Em vez de exclusão, essa ausência foi interpretada metodologicamente como indicativo de evasão, sendo imputada logicamente com nota e progresso 0,0, acompanhada de *flags* binárias de auditoria (*Shadow Matrix*: *flg\_nota\_imputada* e *flg\_tempo\_imputado*). Adicionalmente, a variável bruta *tempo\_total\_de\_acesso* foi higienizada via expressões regulares e desdobrada nas *features* contínuas *minutos\_totais* e *dias\_permanencia*. Tais premissas alinham-se à

<sup>1</sup>Código-fonte completo disponível na Seção "Disponibilidade de Artefatos".

literatura, que destaca a importância de documentar os efeitos da imputação sobre análises de agrupamento [Huang and Keller 2025, Marcolino et al. 2025, Alam et al. 2023, Alfredo et al. 2024].

Na segunda etapa, voltada à engenharia de atributos e preparação para a modelagem não supervisionada, adotou-se uma segregação estrita entre variáveis contínuas e categóricas. Atributos demográficos e geográficos foram preservados na *Analytical Base Table* (ABT) exclusivamente para a análise *post-hoc* de perfilamento, evitando a esparsidade do *One-Hot Encoding* e prevenindo vieses espaciais. O espaço vetorial submetido aos algoritmos foi estruturado em oito dimensões contínuas e ordinais (Tabela 1).

Na terceira etapa, previamente à redução de dimensionalidade (PCA), aplicaram-se três transformações matemáticas sequenciais para garantir a validade da distância euclidiana: (i) Winsorização isolada ( $Q3 + 1,5 \times IQR$ ) para contenção de *outliers* de tempo de tela, calculada sobre dados reais para evitar distorções pela massa de zeros da evasão; (ii) Transformação logarítmica  $\ln(x + 1)$  nas variáveis temporais para comprimir a assimetria de cauda longa; e (iii) Escalonamento padrão (Z-score) de todas as variáveis contínuas. A aplicação subsequente do PCA otimizou a estrutura do espaço vetorial, tornando comparáveis as escalas, reduzindo redundâncias e gerando uma representação sintética [Borges et al. 2018, Greenacre et al. 2022, Hamina 2024]. Na quarta etapa, aplicou-se o algoritmo K-means, com validação interna baseada em WCSS, coeficiente de silhueta [Rousseeuw 1987], índice de Davies–Bouldin [Davies and Bouldin 1979] e índice de Calinski–Harabasz [Caliński and Harabasz 1974], além da comparação integrada dessas métricas com a interpretabilidade dos agrupamentos [Priyanto et al. 2025].

**Tabela 1. Variáveis do espaço vetorial submetido à clusterização K-Means**

| Variável                           | Tipo       | Pré-processamento                                        | Rota  |
|------------------------------------|------------|----------------------------------------------------------|-------|
| <i>nota_final_kdd</i>              | Contínua   | Z-score                                                  | A e B |
| <i>progresso_kdd</i>               | Contínua   | Z-score                                                  | A e B |
| <i>minutos_totais_log1p</i>        | Contínua   | Z-score, Winsorização ( $IQR \times 1,5$ ), $\ln(x + 1)$ | A e B |
| <i>dias_permanencia_log1p</i>      | Contínua   | Z-score, Winsorização ( $IQR \times 1,5$ ), $\ln(x + 1)$ | A e B |
| <i>dias_duracao_curso</i>          | Contínua   | Z-score                                                  | A e B |
| <i>qtd_tentativas</i>              | Discreta   | Z-score                                                  | A e B |
| <i>Idade</i>                       | Contínua   | Z-score                                                  | A e B |
| <i>Escolaridade_Ordinal</i>        | Ordinal    | Z-score                                                  | A e B |
| <i>flg_nota_imputada</i>           | Binária    | <i>Shadow Matrix</i> (1 = Imputado, 0 = Real)            | A     |
| <i>flg_tempo_imputado</i>          | Binária    | <i>Shadow Matrix</i> (1 = Imputado, 0 = Real)            | A     |
| <i>SRE, Macro_Cargo, Município</i> | Categórica | Excluídas do K-Means (Análise <i>Post-Hoc</i> )          | –     |

**Fonte:** Elaborado pelos autores (2026).

**Nota:** As variáveis foram pré-processadas rigorosamente para otimizar a convergência do algoritmo K-Means e mitigar a distorção por grandezas escalares.

A modelagem foi organizada em duas rotas complementares, definidas em função da densidade observacional das inscrições. A Rota A considerou a totalidade da base analítica unificada, com 191.412 inscrições, permitindo examinar conjuntamente evasão, baixa observabilidade transacional e engajamento na política de formação continuada. A Rota B, por sua vez, restringiu a análise às 34.985 inscrições com nota e tempo de tela registrados no ambiente virtual. Essa restrição foi motivada pela alta esparsidade de

56,06% na variável *nota\_final\_na\_formacao* e de 56,05% em *progresso\_na\_formacao* na base transacional do Moodle, conforme detalhado na etapa de análise exploratória. Essa separação teve função metodológica explícita: distinguir, de um lado, a estrutura global da política, inclusive sob forte presença de registros com baixa densidade observacional, e, de outro, os perfis de participação ativa com evidências transacionais observadas no ambiente virtual, reduzindo a influência analítica de casos marcados por ausência de informação em variáveis centrais [Huang and Keller 2025, Marcolino et al. 2025].

Na *Rota B*, o espaço vetorial foi composto por oito variáveis centrais de desempenho, esforço e trajetória formativa. Após a aplicação do PCA, o conjunto foi reduzido a sete componentes, preservando 97,87% da variância explicada e definindo um subespaço ortogonal de análise [Greenacre et al. 2022]. Sobre esse subespaço, a solução mais parcimoniosa foi  $k = 2$ , selecionada com base na convergência entre as métricas internas de qualidade e a interpretabilidade dos agrupamentos [Priyanto et al. 2025]. Na *Rota A*, a solução adotada foi  $k = 4$ , definida a partir da convergência entre as métricas de validação interna e da interpretabilidade substantiva dos agrupamentos, em consonância com recomendações da literatura de EDM que enfatizam a avaliação de clusters por múltiplos critérios e sua relevância analítica para o problema investigado [Pecuchová and Drlík 2024, Lu et al. 2025].

Na etapa final, os resultados da *Rota B* foram projetados territorialmente em duas escalas de agregação, 11 superintendências regionais e 78 municípios [Chew et al. 2020]. Para essa finalidade, definimos o **Índice de Engajamento e Eficiência Proporcional (IEEP)**, calculado como a proporção de profissionais ativos classificados no perfil de maior desempenho da *Rota B* em relação ao total de profissionais ativos da localidade. Esta nomenclatura foi adotada para melhor refletir a natureza da medida como uma proporção de alto desempenho dentro da subamostra ativa, evitando conotações mais amplas de "excelência" e enfatizando a eficiência no engajamento. O indicador foi examinado em sua forma bruta e em sua forma estabilizada, considerando a utilidade de abordagens bayesianas e de procedimentos de suavização para lidar com incerteza, heterogeneidade espacial e flutuações em pequenas áreas, ampliando a comparabilidade territorial [Wang et al. 2024, You and Hidioglou 2023].

#### 4. Resultados e Discussão

Os resultados indicam que a política de formação continuada analisada não se distribui sobre uma população homogênea. Na base analítica unificada, composta por 191.412 inscrições válidas, a articulação entre registros de desempenho e rastros transacionais ampliou a visibilidade sobre padrões de participação e engajamento que tendem a permanecer pouco evidentes quando a análise se apoia apenas em indicadores administrativos mais sintéticos [Porto and Battestin 2025].

Na *Rota B*, restrita às 34.985 inscrições com nota e tempo de tela registrados nos logs do ambiente virtual, a modelagem identificou dois perfis principais de participação ativa. A definição do hiperparâmetro  $k = 2$  foi estabelecida após varredura geométrica sistemática no intervalo  $k \in [2, 8]$  e triangulação de métricas de validação interna no subespaço ortogonal<sup>2</sup>. A análise comparativa das métricas de validação interna revelou uma

<sup>2</sup>Tabela completa de triangulação de métricas ( $k = 2$  a  $8$ ) da Rota B disponível na Seção "Disponibilidade de Artefatos".

divergência característica de dados educacionais empíricos, exigindo o balanceamento entre:

- **Qualidade Geométrica Global:** O Coeficiente de Silhueta atingiu seu pico máximo (0,3770) em  $k = 2$ , indicando a melhor separação global e a menor sobreposição de fronteiras possível para a rede.
- **Densidade Local vs. Parcimônia:** Índices complementares de dispersão (Calinski-Harabasz e Davies-Bouldin) apontaram otimização matemática em  $k = 6$  (10573,0 e 1,1275, respectivamente). Contudo, essa configuração reduziu drasticamente a Silhueta (0,2785), sacrificando a clareza das fronteiras globais em favor de subníveis de alta densidade interna.

A decisão por  $k = 2$  fundamentou-se no princípio da parcimônia e na interpretabilidade educacional. A adoção de partições maiores (como  $k = 6$ ) resultaria em *overfitting* semântico — a fragmentação da rede em micro-perfis pedagogicamente indistinguíveis na prática. Ao restringir a solução a dois agrupamentos, o modelo garantiu alta utilidade analítica, polarizando a rede em regimes de engajamento distintos e acionáveis, interpretados como: o **Perfil 0** (“Engajamento Pleno e Alta Eficiência”) e o **Perfil 1** (“Atrito Intermediário e Evasão Parcial”).

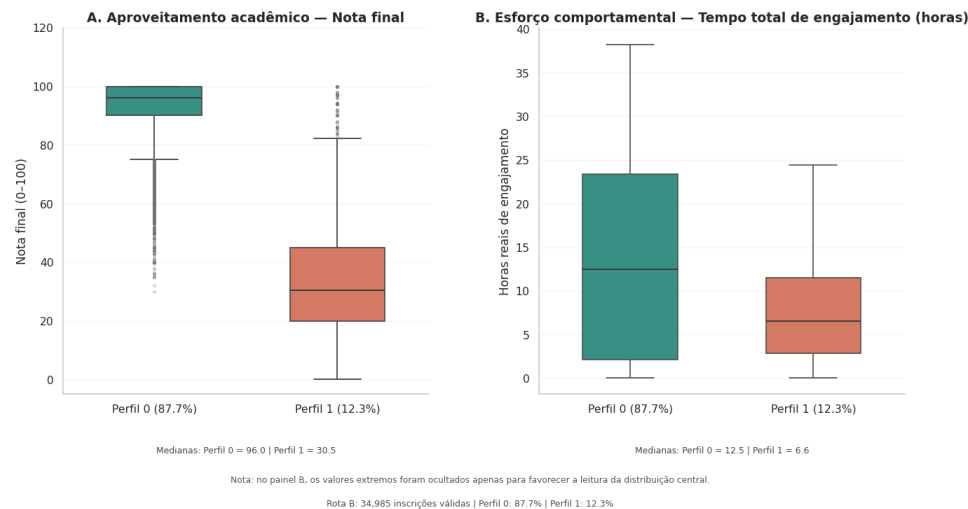
A análise das cargas da PCA<sup>3</sup> revelou que o primeiro componente principal (PC1) foi fortemente influenciado pelo desempenho acadêmico (*nota\_final\_kdd*, com carga de 0,638) e pelas métricas de tempo de permanência (*minutos\_totais\_log1p* e *dias\_permanencia\_log1p*, com cargas de 0,268 e 0,262, respectivamente). Essa configuração vetorial confere sustentação empírica à separação geométrica, demonstrando matematicamente a correlação entre o sucesso na plataforma e o esforço temporal investido.

A partição final revelou uma forte assimetria entre os grupos, aqui tratados como rótulos analíticos de maior fluidez e maior fricção: o primeiro concentrou 30.698 inscrições (87,7%), enquanto o segundo reuniu 4.287 inscrições (12,3%). Embora a separação geométrica entre os grupos tenha sido moderada, a solução binária mostrou-se suficiente para distinguir dois regimes de participação observados nos dados com utilidade analítica, uma vez que, em contextos educacionais, a interpretação dos agrupamentos depende não apenas da separação geométrica, mas também de sua coerência substantiva e de sua utilidade para identificar padrões comportamentais relevantes [Pecuchová and Drlík 2024, Lu et al. 2025]. Essa distinção torna-se mais inteligível quando se observam diretamente as distribuições centrais representadas na Figura 2 e as estatísticas descritivas consolidadas na Tabela 2.

No painel A da Figura 2, observa-se nítida separação entre os perfis quanto ao aproveitamento acadêmico (mediana de 96,0 no **Perfil 0** vs. 30,5 no **Perfil 1**); no painel B, a diferença no esforço comportamental permanece visível, porém menos acentuada (12,5 vs. 6,6 horas de engajamento). Esses resultados indicam que a segmentação está associada sobretudo à combinação entre desempenho e intensidade de engajamento observável no Moodle, em consonância com estudos que relacionam maior envolvimento com recursos da plataforma a melhores resultados acadêmicos [Costa et al. 2025, Velu et al. 2023].

<sup>3</sup>Tabela completa de cargas dos sete componentes principais da Rota B disponível na Seção "Disponibilidade de Artefatos".

**Figura 2. Distribuição do aproveitamento acadêmico (A) e do engajamento temporal por perfil ativo (B) na Rota B.**



**Fonte:** Elaborado pelos autores (2026).  
**Nota:** O painel A apresenta a distribuição da nota final por perfil ativo, e o painel B mostra a distribuição do tempo total de engajamento em horas. As medianas evidenciam contraste acentuado de desempenho e uma diferença moderada de esforço temporal entre os agrupamentos.

A tradução dessa taxonomia bidimensional para o planejamento de políticas públicas permite superar a padronização das ofertas formativas, orientando as estratégias de intervenção sintetizadas na Tabela 2. A fricção sistêmica observada no **Perfil 1**, em particular, fornece insumos diretos para o redesenho instrucional do ambiente virtual: a concentração de dificuldades operacionais ou cognitivas indica a necessidade de fragmentação do conteúdo (*microlearning*), diversificação dos formatos midiáticos e melhoria na usabilidade da plataforma, garantindo que o atrito tecnológico não atue como limitador do engajamento pedagógico [Alfredo et al. 2024].

**Tabela 2. Estatísticas descritivas e recomendações de intervenção por perfil (Rota B - Moodle)**

| Cluster | Perfil                          | N (%)          | Nota Final  | Progresso (%) | Horas Engaj. | Tent.     | Recomendação de Intervenção                                                                                                                                |
|---------|---------------------------------|----------------|-------------|---------------|--------------|-----------|------------------------------------------------------------------------------------------------------------------------------------------------------------|
| C0      | Engaj. Pleno e Alta Eficiência  | 30.698 (87,7%) | 96,0 ± 9,5  | 100,0 ± 6,7   | 12,5 ± 11,4  | 1,0 ± 0,3 | <b>Aceleração e Multiplicação:</b> trilhas autoinstrucionais, certificação acelerada por competência, atuação como multiplicadores ou em tutoria de pares. |
| C1      | Atrito Interm. e Evasão Parcial | 4.287 (12,3%)  | 30,5 ± 20,6 | 44,8 ± 19,5   | 6,6 ± 7,0    | 1,0 ± 0,8 | <b>Suporte e Monitoramento:</b> alertas precoces ( <i>early warning systems</i> ), tutoria ativa e reforço de suporte técnico para letramento digital.     |

**Fonte:** Elaborado pelos autores (2026).  
**Nota:** As medianas evidenciam a sustentação empírica dos rótulos; as recomendações traduzem as características comportamentais de cada perfil em estratégias de intervenção pedagógica.

Com a identificação, interpretação e análise dos perfis na *Rota A*, que incorporou a totalidade das 191.412 inscrições, inclusive aquelas com baixa densidade observacional, a solução com quatro clusters permitiu descrever a estrutura global da política. Após a redução de sete variáveis para cinco componentes principais, com preservação de 96,31%

da variância, a revalidação geométrica confirmou  $k = 4$  como a solução mais consistente.

A análise das cargas da PCA<sup>4</sup> revelou que o primeiro componente principal (PC1) foi fortemente influenciado pelo desempenho acadêmico (*nota\_final\_kdd*, com carga de 0,538) e pelas métricas de tempo de permanência (*minutos\_totais\_logIp* e *dias\_permanencia\_logIp*, com cargas de 0,459 e 0,445, respectivamente). Essa colinearidade estrutural no PC1 confirma matematicamente que a ausência de rastros (evasão) colapsa todas as métricas de engajamento de forma simultânea. Para esta rede global, a topologia de quatro agrupamentos ( $k = 4$ ) foi selecionada a partir de uma varredura sistemática no intervalo  $k \in [2, 8]$ , triangulando os índices de Silhueta, Calinski-Harabasz e Davies-Bouldin. A comparação entre as alternativas evidenciou que soluções menores ( $k \leq 3$ ) apresentavam subajuste espacial, falhando em isolar a densa massa de evasão crônica dos padrões de baixo engajamento ativo. Em contrapartida, partições superiores ( $k \geq 5$ ) maximizavam a densidade local, mas fragmentavam os perfis ativos em micro-grupos redundantes, gerando *overfitting* semântico.

A arquitetura com  $k = 4$  representou o ponto de convergência ideal para a Rota A, fundamentada na otimização simultânea das métricas de validação interna no espaço ortogonal<sup>5</sup>. O coeficiente de Silhueta atingiu seu valor máximo (0,4533), indicando a melhor coesão e separação dos agrupamentos. Na mesma direção, o índice de Calinski-Harabasz alcançou seu pico (111.918,7) e o de Davies-Bouldin registrou seu mínimo global (0,7790). Essa convergência matemática unânime fornece uma justificativa robusta para a parametrização, pois alinha qualidade geométrica — ao isolar perfeitamente a massa de dados imputados —, parcimônia contra a fragmentação excessiva e alta interpretabilidade educacional, mapeando um espectro claro de atrito e sucesso na política pública.

A Tabela 3 apresenta a taxonomia global da Rota A, com estatísticas descritivas e diretrizes de intervenção pedagógica para os quatro perfis identificados a partir da combinação entre progressão administrativa, padrões de acesso e observabilidade transacional: *Certificação Expressa* (45,3%), *Evasão Crônica* (38,2%), *Engajamento Orgânico* (14,6%) e *Resiliência com Fricção* (1,8%). A soma de *Certificação Expressa* e *Evasão Crônica* alcança 83,5% do universo analisado, enquanto apenas 16,4% das inscrições se distribuem entre *Engajamento Orgânico* e *Resiliência com Fricção* (diferença residual por arredondamento decimal). Essa taxonomia evidencia que certificação formal e engajamento pedagogicamente observável não são dimensões equivalentes: a elevada proporção de *Certificação Expressa* sugere que o progresso administrativo pode coexistir com baixa densidade de rastros digitais, enquanto a *Evasão Crônica* reflete o peso estrutural da inatividade no ambiente virtual.

A leitura articulada das duas rotas foi central para a interpretação dos achados: a *Rota A* forneceu o enquadramento estrutural da política, tornando visíveis o peso da evasão e a dissociação entre certificação e rastros digitais, enquanto a *Rota B* permitiu examinar com maior controle analítico os perfis efetivamente ativos no ambiente virtual. Assim, as duas rotas não produziram diagnósticos concorrentes, mas escalas complementares de

<sup>4</sup>Tabela completa de cargas dos cinco componentes principais da Rota A disponível na Seção "Disponibilidade de Artefatos".

<sup>5</sup>Tabela completa de triangulação de métricas ( $k = 2$  a  $8$ ) da Rota A disponível na Seção "Disponibilidade de Artefatos".

**Tabela 3. Estatísticas descritivas e diretrizes de intervenção por perfil (Rota A - Moodle)**

| Cluster | Perfil                  | N (%)          | Nota Final  | Progresso (%) | Horas Engaj. | Tent.     | Diretrizes de Intervenção                                                                                                                               |
|---------|-------------------------|----------------|-------------|---------------|--------------|-----------|---------------------------------------------------------------------------------------------------------------------------------------------------------|
| C0      | Certificação Expressa   | 86.775 (45,3%) | 91,0 ± 13,4 | 100,0 ± 10,4  | 0,0 ± 1,7    | 1,0 ± 0,0 | <b>Governança e Auditoria:</b> revisão das regras de certificação; métricas compostas de engajamento real [de La Hoz-Ruiz et al. 2024, Lu et al. 2025]. |
| C1      | Evasão Crônica          | 73.175 (38,2%) | 0,0 ± 7,9   | 0,0 ± 14,9    | 0,0 ± 1,0    | 1,0 ± 0,0 | <b>Alerta Precoce:</b> suporte técnico proativo nas primeiras semanas para mitigar barreiras de acesso [Porto and Battestin 2025].                      |
| C2      | Engajamento Orgânico    | 27.977 (14,6%) | 94,0 ± 33,4 | 100,0 ± 29,0  | 14,4 ± 10,8  | 1,0 ± 0,1 | <b>Parâmetro de Controle:</b> linha de base para calibrar alertas precoces; não demanda correção de rota.                                               |
| C3      | Resiliência com Fricção | 3.485 (1,8%)   | 78,0 ± 36,1 | 87,5 ± 36,8   | 0,0 ± 4,6    | 4,0 ± 0,0 | <b>Tutoria Ativa:</b> monitoramento contínuo e redesenho de módulos com sobrecarga cognitiva [Pecuchová and Drlík 2024].                                |

**Fonte:** Elaborado pelos autores (2026), a partir do algoritmo K-means aplicado à base global de inscrições.

observação do mesmo fenômeno, traduzindo essas desigualdades comportamentais em evidências acionáveis para a formulação de políticas de intervenção focalizadas em acompanhamento pedagógico e gestão institucional [Siemens 2013, Cerezo et al. 2024].

Passando para a análise da projeção territorial, especificamente da *Rota B*, cujos registros transacionais dos profissionais da educação nos cursos do Moodle foram efetivamente observados, verifica-se que o engajamento ativo também se distribui de forma desigual no espaço [Chew et al. 2020]. O **Índice de Engajamento e Eficiência Proporcional (IEEP)** permitiu comparar localidades a partir da proporção de profissionais ativos situados no perfil de melhor desempenho relativo, deslocando a análise de totais absolutos para uma medida proporcional de engajamento e eficiência. Como esse indicador deriva exclusivamente da subamostra ativa da *Rota B*, ele não deve ser interpretado como medida da qualidade global da política em cada território, mas como composição relativa dos perfis ativos observados.

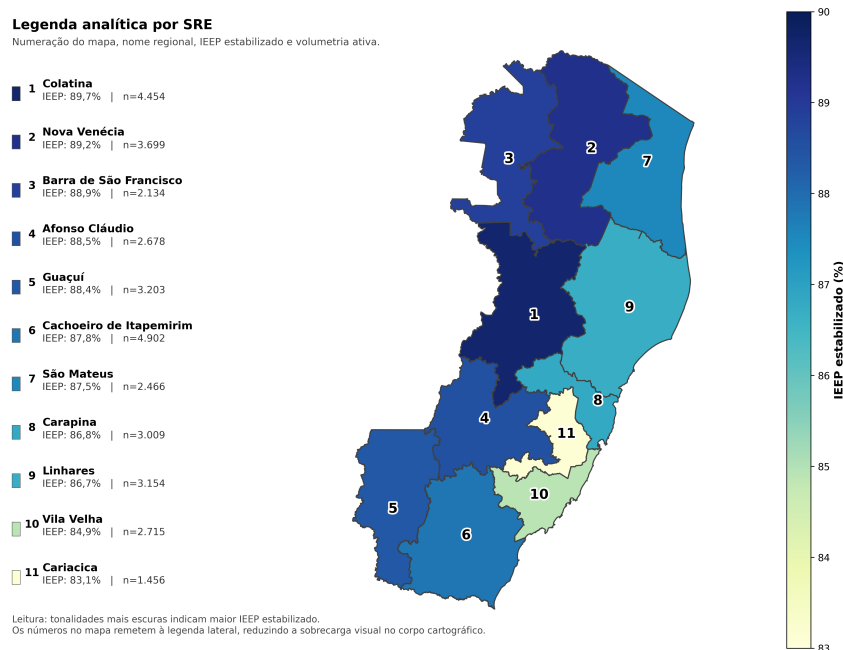
A Figura 3 mostra que esse desempenho relativo não se distribui homogeneamente entre as superintendências regionais. A legenda analítica e a intensidade cromática do mapa indicam valores mais elevados do **IEEP** estabilizado em Colatina (89,68), Nova Venécia (89,25) e Barra de São Francisco (88,86), enquanto Cariacica (83,13), Vila Velha (84,91) e Carapina (86,82) ocupam as menores posições.

Essa configuração mostrada na Figura sugere a presença de um gradiente territorial no interior da subamostra ativa da *Rota B* [Chew et al. 2020]. A leitura conjunta entre o valor do indicador e volumetria ativa reforça esse ponto, pois regionais com **IEEP** mais alto não coincidem necessariamente com as maiores bases ativas, o que confirma a natureza proporcional do índice. A comparação entre a taxa bruta e a taxa suavizada mostrou ajuste médio de aproximadamente 0,26 ponto percentual, com diferença máxima inferior a 1 ponto percentual. Isso indica que a estabilização operou como mecanismo de robustez inferencial, e não como reconfiguração substantiva da hierarquia territorial.

Na decomposição microrregional, observou-se heterogeneidade interna pouco visível nas médias regionais. A SRE de Cachoeiro de Itapemirim, por exemplo, permaneceu

em posição intermediária-alta no ranking macrorregional (87,83), mas concentrou simultaneamente o município com melhor desempenho relativo do estado, Jerônimo Monteiro (93,27), além de outras localidades com resultados inferiores. Esse contraste indica que a distribuição territorial do engajamento não é uniforme e que inferências baseadas apenas em agregados regionais devem ser interpretadas com cautela, uma vez que o nível de agregação pode ocultar heterogeneidades internas relevantes [Kakampakou et al. 2025].

**Figura 3. Distribuição macrorregional do IEEP estabilizado por superintendência regional de educação (SRE) do estado do Espírito Santo.**



**Fonte:** Elaborado pelos autores (2026), versão estabilizada do IEEP aplicada à Rota B.

**Nota:** A coloração representa o IEEP suavizado por SRE, enquanto a legenda lateral explicita a ordenação regional e a volumetria da subamostra ativa (*n*), favorecendo a comparação territorial do desempenho relativo na Rota B.

Em conjunto, os resultados indicam que a política de formação continuada analisada é atravessada por diferentes regimes de participação e por desigualdades territoriais relevantes. Nessa perspectiva, a contribuição da modelagem não se restringe à identificação de agrupamentos, mas à produção de evidências capazes de subsidiar a focalização territorial, desenho pedagógico diferenciado e estratégias de acompanhamento mais compatíveis com a heterogeneidade funcional e espacial da rede [de La Hoz-Ruiz et al. 2024].

#### 4.1. Limitações e ameaças à validade

Os resultados devem ser interpretados à luz de limitações inerentes à base utilizada, sobretudo da natureza indireta dos indicadores comportamentais e da dependência de sua validade em relação ao contexto de coleta, o que pode confundir baixa participação efetiva com baixa evidência registrada no ambiente virtual [Winter et al. 2024, Hakimi et al. 2021, Huang and Keller 2025]. Há também limitações de construto: tempo de tela, progresso e nota final operam como *proxies* de engajamento e desempenho sem esgotar a complexidade dos processos formativos, e os perfis identificados devem ser lidos como categorias analíticas inferidas dos dados, não como classes naturais dos sujeitos. Por fim, a generalização dos achados requer cautela, já que os resulta-

dos refletem características institucionais, tecnológicas e territoriais da rede investigada [Siemens 2013, de La Hoz-Ruiz et al. 2024].

## 5. Cuidados Éticos

A pesquisa utilizou exclusivamente dados secundários desidentificados na origem, em conformidade com a Lei Geral de Proteção de Dados Pessoais e com a Resolução CNS n. 510/2016 [BRASIL 2018, BRASIL. Conselho Nacional de Saúde 2016]. As chaves primárias foram substituídas, antes do processamento analítico, por identificadores criptográficos, impedindo a circulação de dados pessoais diretos ao longo do pipeline e reforçando a proteção da privacidade e da identidade dos participantes. O tratamento dos dados ausentes, a criação de variáveis derivadas e a produção de indicadores foram conduzidos com rastreabilidade metodológica e transparência analítica, preservando a responsabilidade interpretativa sobre os resultados [Huang and Keller 2025, Hakimi et al. 2021, Alfredo et al. 2024]. O estudo foi aprovado pelo Comitê de Ética em Pesquisa (CEP) do IFES, conforme Certificado de Apresentação para Apreciação Ética (CAAE): 94696925.5.0000.5072.

## 6. Considerações Finais

Este estudo demonstrou que a compreensão da política de formação continuada em larga escala transcende a análise de métricas administrativas isoladas. A integração entre registros institucionais e rastros transacionais do Moodle revelou a heterogeneidade intrínseca da participação, permanência e desempenho dos profissionais da educação, bem como suas assimetrias territoriais [Porto and Battestin 2025]. Metodologicamente, o trabalho consolidou a viabilidade de articular o paradigma KDD com PCA, K-means e análise espacial em uma base institucional de grande volume [Fayyad et al. 1996, Greenacre et al. 2022, Chew et al. 2020]. A arquitetura analítica em duas rotas, contribuição metodológica central, revelou-se fundamental para desvendar essa complexidade: a *Rota A* ofereceu visão macropolítica da estrutura global da política, inclusive sob forte presença de baixa densidade observacional, identificando perfis como "Certificação Expressa" e "Evasão Crônica"; a *Rota B* aprofundou a análise nos casos com observabilidade transacional robusta, identificando perfis ativos com distintos níveis de fluidez e fricção e embasando intervenções pedagógicas direcionadas. Essa distinção elucidou o papel crítico do tratamento de dados ausentes, transformando sua ausência em indicador comportamental relevante [Huang and Keller 2025, Marcolino et al. 2025]. Para futuras agendas, sugere-se a exploração de séries históricas mais completas, modelagem longitudinal de trajetórias e algoritmos de clusterização alternativos, ampliando a auditabilidade e a comparabilidade das análises [Chew et al. 2020].

## Disponibilidade de Artefatos

O código-fonte completo do pipeline analítico – incluindo os scripts de pré-processamento, engenharia de atributos, redução de dimensionalidade (PCA) e clusterização (K-means) –, bem como as tabelas complementares de cargas dos componentes principais e de triangulação de métricas de validação interna ( $k = 2$  a 8) das Rotas A e B não incluídas no corpo do artigo por limitação de espaço, estão disponíveis publicamente em: <https://github.com/FormacaoPerfis/pipeline>.

## Declaração sobre uso de Inteligência Artificial

Ferramentas de inteligência artificial generativa foram utilizadas exclusivamente como apoio à revisão linguística, à reorganização textual e ao ajuste estilístico do manuscrito. O conteúdo final foi criticamente revisto pelos autores, que assumem integral responsabilidade por sua redação, análise e interpretação.

## Referências

- Alam, S., Ayub, M. S., Arora, S., and Khan, M. A. (2023). An investigation of the imputation techniques for missing values in ordinal data enhancing clustering and classification analysis validity. *Decision Analytics Journal*, 9:100341. Disponível em: <https://doi.org/10.1016/j.dajour.2023.100341>. Acesso em: 17 mar. 2026.
- Alfredo, R., Echeverria, V., Jin, Y., Yan, L., Swiecki, Z., Gašević, D., and Martinez-Maldonado, R. (2024). Human-centred learning analytics and AI in education: A systematic literature review. *Computers and Education: Artificial Intelligence*, 6:100215. Disponível em: <https://doi.org/10.1016/j.caeai.2024.100215>. Acesso em: 17 mar. 2026.
- Aulakh, K., Roul, R. K., and Kaushal, M. (2023). E-learning enhancement through educational data mining with Covid-19 outbreak period in backdrop: A review. *International Journal of Educational Development*, 101:102814. Disponível em: <https://doi.org/10.1016/j.ijedudev.2023.102814>. Acesso em: 17 fev. 2026.
- Baker, R. S. and Yacef, K. (2009). The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1(1):3–17. Disponível em: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8>. Acesso em: 17 fev. 2026.
- Borges, V. R. P., Esteves, S. L., Araújo, P. D. N., Oliveira, L. C., and Holanda, M. (2018). Using principal component analysis to support students' performance prediction and data analysis. In *Anais do XXIX Simpósio Brasileiro de Informática na Educação (SBIE 2018)*, pages 1383–1392, Fortaleza. SBC. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/8099>. Acesso em: 17 fev. 2026.
- BRASIL (2018). Lei n. 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD). Disponível em: [https://www.planalto.gov.br/ccivil\\_03/\\_ato2015-2018/2018/lei/l13709.htm](https://www.planalto.gov.br/ccivil_03/_ato2015-2018/2018/lei/l13709.htm). Acesso em: 17 fev. 2026.
- BRASIL (2024). Lei n. 14.945, de 31 de julho de 2024. Altera a Lei n. 9.394/1996 e define diretrizes para o ensino médio. Disponível em: [https://www.planalto.gov.br/ccivil\\_03/\\_ato2023-2026/2024/lei/l14945.htm](https://www.planalto.gov.br/ccivil_03/_ato2023-2026/2024/lei/l14945.htm). Acesso em: 17 fev. 2026.
- BRASIL. Conselho Nacional de Educação (2020). Resolução CNE/CP n. 1, de 27 de outubro de 2020. Dispõe sobre as Diretrizes Curriculares Nacionais para a Formação Continuada de Professores da Educação Básica. Disponível em: [https://portal.mec.gov.br/index.php?option=com\\_docman&view=download&al](https://portal.mec.gov.br/index.php?option=com_docman&view=download&al)

ias=164841-rcp001-20&category\_slug=outubro-2020-pdf&Itemid=30192. Acesso em: 17 fev. 2026.

BRASIL. Conselho Nacional de Educação (2022). Parecer CNE/CEB n. 2, de 17 de fevereiro de 2022. Normas sobre Computação na Educação Básica – complemento à BNCC. Disponível em: [https://portal.mec.gov.br/index.php?option=com\\_docman&view=download&alias=235511-pceb002-22&category\\_slug=fevereiro-2022-pdf&Itemid=30192](https://portal.mec.gov.br/index.php?option=com_docman&view=download&alias=235511-pceb002-22&category_slug=fevereiro-2022-pdf&Itemid=30192). Acesso em: 17 fev. 2026.

BRASIL. Conselho Nacional de Educação (2024). Resolução CNE/CP n. 4, de 29 de maio de 2024. Dispõe sobre as Diretrizes Curriculares Nacionais para a Formação Inicial em Nível Superior de Profissionais do Magistério da Educação Escolar Básica (cursos de licenciatura, cursos de formação pedagógica para graduados não licenciados e cursos de segunda licenciatura). Disponível em: [https://www.gov.br/mec/pt-br/cne/pdf/resolucoes-do-cne/cp/2024/rcp004\\_24.pdf](https://www.gov.br/mec/pt-br/cne/pdf/resolucoes-do-cne/cp/2024/rcp004_24.pdf). Acesso em: 17 fev. 2026.

BRASIL. Conselho Nacional de Educação (2025). Resolução CNE/CEB n. 2, de 21 de março de 2025. Institui diretrizes operacionais nacionais sobre o uso de dispositivos digitais e a integração curricular da educação digital e midiática. Disponível em: [https://www.gov.br/mec/pt-br/cne/2025/marco/rceb002\\_25.pdf](https://www.gov.br/mec/pt-br/cne/2025/marco/rceb002_25.pdf). Acesso em: 17 fev. 2026.

BRASIL. Conselho Nacional de Saúde (2016). Resolução n. 510, de 7 de abril de 2016. Dispõe sobre as normas aplicáveis a pesquisas em ciências humanas e sociais. Disponível em: <https://www.gov.br/conselho-nacional-de-saude/pt-br/atos-normativos/resolucoes/2016/resolucao-no-510.pdf/@download/file>. Acesso em: 17 fev. 2026.

Brito, R. d. S. and Nunes, C. P. (2025). Políticas de formação continuada de professores no brasil: avanços e retrocessos. *Revista Brasileira de Política e Administração da Educação*, 41(1). Disponível em: <https://seer.ufrgs.br/index.php/rbpa/article/view/141276>. Acesso em: 17 fev. 2026.

Caliński, T. and Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics - Theory and Methods*, 3(1):1–27. Disponível em: <https://doi.org/10.1080/03610927408827101>. Acesso em: 11 mar. 2026.

Cerezo, R., Lara, J. A., Azevedo, R., and Romero, C. (2024). Reviewing the differences between learning analytics and educational data mining: Towards educational data science. *Computers in Human Behavior*, 154:108155. Disponível em: <https://doi.org/10.1016/j.chb.2024.108155>. Acesso em: 17 fev. 2026.

Chew, B., Satpathy, A., and Wong, E. (2020). Geospatial analyses to determine academic success factors in California’s K-12 education. *Annals of GIS*, 26(2):81–100. Disponível em: <https://doi.org/10.1080/19475683.2020.1739141>. Acesso em: 17 fev. 2026.

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 *Anais do XXXVI Simpósio Brasileiro de Informática na Educação (SBIE*

- 2025). SBC. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/38444>. Acesso em: 20 mar. 2026.
- Davies, D. L. and Bouldin, D. W. (1979). A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(2):224–227. Disponível em: <https://doi.org/10.1109/TPAMI.1979.4766909>. Acesso em: 11 mar. 2026.
- de La Hoz-Ruiz, J., Khalil, M., Domingo Segovia, J., and Liu, Q. (2024). Learning analytics for enhanced professional capital development: a systematic review. *Frontiers in Psychology*, 15:1302658. Disponível em: <https://doi.org/10.3389/fpsyg.2024.1302658>. Acesso em: 18 fev. 2026.
- ESPÍRITO SANTO. Secretaria da Educação (2026). Diretrizes operacionais da formação continuada dos profissionais da educação: 2026. Documento institucional do Cefope/SEDU. Disponível em: <https://sedu.es.gov.br/Media/sedu/pdf%20e%20Arquivos/DIRETRIZES%20OPERACIONAIS%20DA%20FORMA%C3%87%C3%83O%20CONTINUADA%20DOS%20PROFISSIONAIS%20DA%20EDUCA%C3%87%C3%83O%20%202026.pdf>. Acesso em: 23 abr. 2026.
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17(3):37–54. Disponível em: <https://ojs.aaai.org/index.php/aimagazine/article/view/1230>. Acesso em: 17 fev. 2026.
- Gamage, S. H. P. W., Ayres, J. R., and Behrend, M. B. (2022). A systematic review on trends in using moodle for teaching and learning. *International Journal of STEM Education*, 9(1):9. Disponível em: <https://doi.org/10.1186/s40594-021-00323-x>. Acesso em: 19 fev. 2026.
- Gatti, B. A. (2021). Formação de professores no brasil: Políticas e programas. *Revista Paradigma*, 42(Extra 2):1–17. Disponível em: <https://mail.revistaparadigma.com.br/index.php/paradigma/article/view/1044>. Acesso em: 17 fev. 2026.
- Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D’Enza, A., Markos, A., and Tuzhilina, E. (2022). Principal component analysis. *Nature Reviews Methods Primers*, 2:100. Disponível em: <https://www.nature.com/articles/s43586-022-00184-w>. Acesso em: 20 mar. 2026.
- Hakimi, L., Eynon, R., and Murphy, V. A. (2021). The ethics of using digital trace data in education: A thematic review of the research landscape. *Review of Educational Research*, 91(5):671–717. Disponível em: <https://doi.org/10.3102/00346543211020116>. Acesso em: 17 mar. 2026.
- Hamina, S. (2024). Data reduction using principal component analysis: theoretical underpinnings and practical applications in public health. *Journal of Contemporary Medical Education*, 14(1):1–14. Disponível em: <https://www.jcmedu.org/jcmedu-articles/data-reduction-using-principal-component-analysis-theoretical-underpinnings-and-practical-application-s-in-public-health-105124.html>. Acesso em: 19 fev. 2026.

- Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, Waltham, MA, 3 edition. Disponível em: <https://www.sciencedirect.com/book/9780123814791/data-mining-concepts-and-techniques>.
- Huang, F. and Keller, B. (2025). Working with missing data in large-scale assessments. *Large-scale Assessments in Education*, 13:13. Disponível em: <https://doi.org/10.1186/s40536-025-00248-9>. Acesso em: 17 mar. 2026.
- Huang, L., Liang, M., Xiong, Y., Wu, X., and Lim, C. P. (2024). A systematic review of technology-enabled teacher professional development during covid-19 pandemic. *Computers & Education*, 223:105168. Disponível em: <https://doi.org/10.1016/j.compedu.2024.105168>. Acesso em: 18 fev. 2026.
- Kakampakou, L., Stokes, J., Hoehn, A., de Kamps, M., Lawniczak, W., Arnold, K. F., Hensor, E. M. A., Heppenstall, A. J., and Gilthorpe, M. S. (2025). Simulating hierarchical data to assess the utility of ecological versus multilevel analyses in obtaining individual-level causal effects. *BMC Medical Research Methodology*, 25:79. Disponível em: <https://link.springer.com/article/10.1186/s12874-025-02504-6>. Acesso em: 20 mar. 2026.
- Lu, Y., Yeom, S., Maktoubian, J., Rahman, M. M., and Kim, S.-H. (2025). Improve student risk prediction with clustering techniques: A systematic review in education data mining. *Education Sciences*, 15(12):1695. Disponível em: <https://doi.org/10.3390/educsci15121695>. Acesso em: 11 mar. 2026.
- Marcolino, M. R., Porto, T. R., Primo, T. T., Targino, R., Ramos, V., Queiroga, E. M., Munoz, R., and Cechinel, C. (2025). Student dropout prediction through machine learning optimization: insights from Moodle log data. *Scientific Reports*, 15:9840. Disponível em: <https://doi.org/10.1038/s41598-025-93918-1>. Acesso em: 18 fev. 2026.
- Paolucci, C., Vancini, S., Bex II, R. T., Cavanaugh, C., Salama, C., and de Araujo, Z. (2024). A review of learning analytics opportunities and challenges for K-12 education. *Heliyon*, 10:e25767. Disponível em: <https://doi.org/10.1016/j.heliyon.2024.e25767>. Acesso em: 20 mar. 2026.
- Pecuchová, J. and Drlík, M. (2024). Enhancing the early student dropout prediction model: A comprehensive analysis of clustering algorithms. *IEEE Access*, 12:159342–159367. Disponível em: <https://ieeexplore.ieee.org/document/10736593>. Acesso em: 18 fev. 2026.
- Porto, B. and Battestin, V. (2025). Learning Analytics no Moodle: desenvolvimento e aplicação de um painel de indicadores para apoio à prática pedagógica. *Revista EaD & Tecnologias Digitais na Educação*, 13(20):350–366. Disponível em: <https://ojs.ufgd.edu.br/ead/article/view/20594>. Acesso em: 18 fev. 2026.
- Priyanto, E., Berlilana, and Tahyudin, I. (2025). Enhancing clustering performance through benchmarking of dimensionality reduction techniques on educational data. *Jurnal Teknik Informatika (JUTIF)*, 6(2):641–654. Disponível em: [847](https://jutif.</a></p></div><div data-bbox=)

[if.unsoed.ac.id/index.php/jurnal/article/view/4297](http://if.unsoed.ac.id/index.php/jurnal/article/view/4297). Acesso em: 18 fev. 2026.

Ramos, J. L. C., Santos, L. F. L., Silva, J. C. S., and Rodrigues, R. L. (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 (SBIE 2020)*, pages 932–941. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/12849>. Acesso em: 18 fev. 2026.

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. Disponível em: [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7). Acesso em: 25 abr. 2026.

Siemens, G. (2013). Learning analytics: The emergence of a discipline. *American Behavioral Scientist*, 57(10):1380–1400. Disponível em: <https://doi.org/10.1177/0002764213498851>. Acesso em: 19 fev. 2026.

Velu, S., Gangathulasi, J., Dineshkumar, K. S. A., and Sankayya, M. (2023). Design of self-regulated learning framework for professional development program through learning analytics. In *2023 ASEE Annual Conference & Exposition*. Disponível em: <https://peer.asee.org/design-of-self-regulated-learning-framework-for-professional-development-program-through-learning-analytics>. Acesso em: 19 fev. 2026.

Wang, Y., Chen, X., and Xue, F. (2024). A review of bayesian spatiotemporal models in spatial epidemiology. *ISPRS International Journal of Geo-Information*, 13(3):97. Disponível em: <https://doi.org/10.3390/ijgi13030097>. Acesso em: 20 mar. 2026.

Winter, M., Mordel, J., Mendzheritskaya, J., Biedermann, D., Ciordas-Hertel, G.-P., Hahnel, C., Bengs, D., Wolter, I., Goldhammer, F., Drachsler, H., Artelt, C., and Horz, H. (2024). Behavioral trace data in an online learning environment as indicators of learning engagement in university students. *Frontiers in Psychology*, 15:1396881. Disponível em: <https://doi.org/10.3389/fpsyg.2024.1396881>. Acesso em: 19 fev. 2026.

You, Y. and Hidirolou, M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, 39(4):571–590. Disponível em: <https://doi.org/10.2478/JOS-2023-0026>. Acesso em: 20 mar. 2026.