Revista Brasileira de Informática na Educação – RBIE
Brazilian Journal of Computers in Education
(ISSN online: 2317-6121; print: 1414-5685)
https://sol.sbc.org.br/journals/index.php/rbie

Submission: dd/Mmm/yyyy; 1st round notif.: dd/Mmm/yyyy; New version: dd/Mmm/yyyy; 2nd round notif.: dd/Mmm/yyyy;
Camera ready: dd/Mmm/yyyy; Edition review: dd/Mmm/yyyy; Available online: dd/Mmm/yyyy; Published: dd/Mmm/yyyy;

Modelo Híbrido de Inteligência Artificial para Análise de Eficiência Educacional: Integração de Aprendizado Não Supervisionado e Análise Envoltória de Dados

Title: Hybrid Artificial Intelligence Model for Educational Efficiency Analysis: Integration of Unsupervised Learning and Data Envelopment Analysis

Título: Modelo híbrido de inteligencia artificial para el análisis de la eficiencia educativa: integración del aprendizaje no supervisado y el análisis envolvente de datos

Francisco José Cardoso da Conceição
Universidade Estadual do Ceará
ORCID: 0009-0009-2687-6450
fco.cardoso@aluno.uece.br

Camila Campos Colares das Dores
Universidade Estadual do Ceará
ORCID: 0000-0002-2619-665X
camilacamposcolares@gmail.com

Gustavo Augusto Lima de Campos
Universidade Estadual do Ceará
ORCID: 0000-0001-7175-9071
gustavo.campos@uece.br

Rafael Lopes Gomes
Universidade Estadual do Ceará
ORCID: 0000-0001-7922-0695
rafa.lopes@uece.br

Resumo

A baixa eficiência dos sistemas educacionais públicos no Brasil representa um desafio crítico que demanda soluções inovadoras baseadas em Inteligência Artificial. Este artigo propõe um modelo híbrido que integra técnicas de aprendizado não supervisionado, redução de dimensionalidade (PCA) e Análise por Envoltória de Dados (DEA) para avaliar a eficiência de sistemas educacionais complexos. Utilizando dados do Sistema de Avaliação da Educação Básica (SAEB) 2021, o modelo foi aplicado a 16.664 escolas públicas de ensino médio, demonstrando capacidade superior de identificação de padrões latentes em relação a abordagens tradicionais. Os resultados evidenciam que, na amostra analisada, a metodologia proposta melhora em 23% a precisão de agrupamento e identifica oportunidades de melhoria em 78% das instituições. Além de produzir metas de desempenho intra-cluster e pares de referência comparáveis, o framework permite mapear assimetrias de desempenho entre contextos educacionais similares, evidenciando desigualdades e orientando intervenções focalizadas (ex.: redes de apoio entre escolas, alocação prioritária de recursos e formação docente dirigida). A principal contribuição consiste em um pipeline escalável que combina múltiplas técnicas de IA para gerar insights acionáveis a gestores, com foco em equidade e melhoria contínua da eficiência em sistemas complexos.

Palavras-chave: *Inteligência Artificial; Análise de Eficiência; Aprendizagem Não Supervisada; Sistemas Educacionais; Análise Envoltória de Dados*

Abstract

Low efficiency in Brazil's public education systems remains a critical challenge that calls for innovative, AI-driven solutions. This paper proposes a hybrid framework that integrates unsupervised learning, dimensionality reduction via Principal Component Analysis (PCA), and Data Envelopment Analysis (DEA) to assess efficiency in complex educational systems. Using data from the 2021 Basic Education Assessment System (SAEB), the model was applied to 16,664 public high schools and outperformed traditional approaches in uncovering latent patterns. In our sample, the

Cite as: Conceição, F. J. C., Dores, C. C. C., Campos, G. A. L., & Gomes, R. L. (2026). Modelo Híbrido de Inteligência Artificial para Análise de Eficiência Educacional: Integração de Aprendizado Não Supervisionado e Análise Envoltória de Dados. Revista Brasileira de Informática na Educação, vol. 34, pp. 429–457. https://doi.org/10.5753/rbie.2026.6708.

proposed methodology improves clustering accuracy by 23% and identifies actionable improvement opportunities for 78% of schools. Beyond producing intra-cluster performance targets and peer benchmarks, the framework enables the mapping of performance asymmetries among comparable contexts, thereby surfacing inequities and guiding targeted interventions (e.g., peer-support networks, prioritized resource allocation, and focused teacher development). The main contribution is a scalable pipeline that combines multiple AI techniques to generate actionable insights for education managers, with an emphasis on equity and continuous efficiency improvement in complex systems.
Keywords: *Artificial Intelligence; Efficiency Analysis; Unsupervised Learning; Educational Systems; Data Envelopment Analysis*

Resumen

La baja eficiencia de los sistemas públicos de educación en Brasil sigue siendo un desafío crítico que exige soluciones innovadoras basadas en IA. Este artículo propone un marco híbrido que integra aprendizaje no supervisado, reducción de dimensionalidad mediante Análisis de Componentes Principales (PCA) y Análisis Envolvente de Datos (DEA) para evaluar la eficiencia en sistemas educativos complejos. Utilizando datos del Sistema de Evaluación de la Educación Básica (SAEB) 2021, el modelo se aplicó a 16,664 escuelas públicas de educación media y superó los abordajes tradicionales en la detección de patrones latentes. En nuestra muestra, la metodología propuesta mejora en un 23 % la precisión de la agrupación e identifica oportunidades de mejora en el 78 % de las escuelas. Además de generar metas de desempeño intra-cluster y escuelas de referencia comparables, el marco permite mapear asimetrías de desempeño entre contextos equivalentes, visibilizando inequidades y orientando intervenciones focalizadas (p. ej., redes de apoyo entre escuelas, asignación prioritaria de recursos y formación docente dirigida). La principal contribución es una tubería escalable que combina múltiples técnicas de IA para producir insights accionables para gestores educativos, con énfasis en la equidad y la mejora continua de la eficiencia en sistemas complejos.
Palabras clave: *Inteligencia Artificial; Análisis de Eficiencia; Aprendizaje No Supervisado; Sistemas Educativos; Análisis Envolvente de Datos*

1 Introdução

A análise da eficiência das unidades educacionais públicas configura-se como um dos principais desafios para o aprimoramento da qualidade do ensino no Brasil. Dados recentes do Sistema Nacional de Avaliação da Educação Básica (SAEB) indicam que menos de 5% dos estudantes concluem o Ensino Médio com proficiências adequadas, sobretudo em matemática, evidenciando desigualdades persistentes que comprometem o desenvolvimento educacional em âmbito regional e nacional (Campos & Vieira, 2021; Soares, 2022). Neste estudo, eficiência educacional é definida como a capacidade de cada escola converter seus insumos institucionais em resultados mensuráveis de aprendizagem e participação, preservando proporcionalidade entre contextos distintos e evitando interpretações distorcidas em razão de desigualdades estruturais.

O sistema educacional brasileiro é marcado por uma elevada heterogeneidade socioeconômica e geográfica, refletida em diferenças significativas na infraestrutura das escolas, nas condições de formação e experiência docente e nas características culturais e sociais das comunidades atendidas (Miranda & Miranda, 2018; Pereira et al., 2020). Apesar dessa heterogeneidade, políticas públicas ainda tendem a utilizar critérios essencialmente geográficos para organização e apoio das escolas, o que pode limitar a efetividade das intervenções ao ignorar particularidades institucionais relevantes. Essa realidade evidencia a necessidade de ferramentas analíticas capazes de formar agrupamentos que reflitam de maneira mais fiel as especificidades institucionais e contextuais das escolas, superando abordagens baseadas exclusivamente em critérios geográficos ou administrativos.

Neste trabalho, entende-se por avaliação justa e contextualizada a interpretação dos resultados escolares que considera simultaneamente os recursos disponíveis, as condições de funcionamento e as características socioeconômicas das escolas, evitando conclusões distorcidas produzidas por desigualdades estruturais.

A literatura recente reúne estudos que utilizam técnicas de clusterização ou DEA, isoladamente ou em combinação, para avaliar redes escolares em diferentes contextos (Ersoy, 2021; Rassouli-Currier, 2007).

Ainda são pouco exploradas aplicações que integrem PCA, k-means e DEA em bases públicas em larga escala como o SAEB, com o objetivo de formar grupos internamente homogêneos e comparáveis do ponto de vista da eficiência.

A evolução das tecnologias de inteligência artificial tem possibilitado o desenvolvimento de métodos para lidar com a complexidade e o grande volume de dados educacionais disponíveis. Técnicas como a Análise de Componentes Principais (PCA), voltada à redução de dimensionalidade dos dados, algoritmos de agrupamento, como o k-means, e a Análise por Envoltória de Dados (DEA), aplicada à estimativa da eficiência relativa das escolas por meio de métricas formais baseadas em programação linear, nas quais cada unidade é comparada objetivamente a uma fronteira eficiente construída a partir das melhores combinações observadas entre insumos e outputs no conjunto analisado, têm sido integradas para gerar diagnósticos mais detalhados e justos sobre o desempenho das escolas públicas (Ersoy, 2021; Rassouli-Currier, 2007).

Nesse contexto, a literatura ainda discute de forma limitada como a aplicação sequencial dessas técnicas afeta a interpretação dos resultados, especialmente no que diz respeito à justificativa metodológica da ordem adotada no *pipeline*. A organização utilizada neste estudo, composta

por padronização das variáveis, aplicação do PCA, clusterização por k-means e posterior avaliação de eficiência por DEA, é adotada para reduzir redundâncias, mitigar colinearidade e melhorar a discriminação entre unidades antes da modelagem da eficiência, mas essa discussão é pouco explorada nas pesquisas existentes.

Em paralelo, estudos brasileiros em ciência de dados têm discutido a importância da redução de dimensionalidade em bases com alta complexidade. Estudos como o dos autores (Battisti & Carvalho, 2022) apresentam variações supervisionadas do PCA que melhoram a precisão em cenários com grande número de atributos, reforçando o papel da redução de dimensionalidade na mitigação de redundância e ruído.

De modo semelhante, trabalhos como o de (Pessano & Halmenschlager, 2005) demonstram o uso integrado de k-means em tarefas de mineração de dados em grandes repositórios corporativos, enquanto pesquisas em arquiteturas de alto desempenho, como as de (Avelar et al., 2014), evidenciam que o k-means pode superar heurísticas tradicionais em problemas de otimização e agrupamento.

Essas evidências reforçam a coerência metodológica do *pipeline* adotado neste estudo, que organiza padronização, PCA, clusterização por k-means e avaliação por DEA como etapas complementares para lidar com alta dimensionalidade, heterogeneidade estrutural e necessidade de maior discriminação entre grupos antes da modelagem da eficiência (Pimenta et al., 2024).

Além das questões estruturais e analíticas, o desenvolvimento do pensamento computacional surge como um componente fundamental para a formação dos estudantes diante dos desafios contemporâneos, promovendo competências que ultrapassam o domínio técnico e alcançam habilidades críticas, criativas e estratégicas para a resolução de problemas complexos (Guan et al., 2020; Roll & Wylie, 2016). Simultaneamente, essa perspectiva vem acompanhada pelo crescente interesse na adoção de metodologias ativas de ensino, como aprendizagem baseada em projetos e uso de tecnologias digitais, que têm demonstrado potencial para aumentar o engajamento estudantil e melhorar o desempenho acadêmico (Leporace, 2023; Park & Kwon, 2024). Essas discussões reforçam a relevância de análises que integrem dimensões institucionais, pedagógicas e socioeconômicas para compreender de forma mais abrangente os fatores que influenciam os resultados escolares.

Quanto aos dados utilizados, este estudo emprega exclusivamente microdados públicos anonimizados disponibilizados pelo Inep, que passam por processos de desidentificação antes da publicação. Esses procedimentos garantem a impossibilidade de identificação direta de escolas, docentes ou estudantes. O estudo caracteriza-se, portanto, como uso secundário de dados públicos e desidentificados, em conformidade com a Lei Geral de Proteção de Dados (LGPD), não havendo qualquer procedimento adicional de anonimização realizado pelos autores.

Com base nesse contexto, este artigo propõe um modelo híbrido que utiliza dados do SAEB 2021 referentes ao estado do Ceará para segmentar escolas públicas de ensino médio em grupos homogêneos por meio da aplicação conjunta de PCA, k-means e DEA. O objetivo é identificar padrões de eficiência dentro desses agrupamentos e fornecer subsídios fundamentados para o estabelecimento de parcerias estratégicas entre as escolas, além de orientar políticas públicas que considerem as especificidades de cada grupo e promovam intervenções educacionais mais eficazes e equitativas. A principal contribuição deste estudo consiste em apresentar uma abordagem integrada que responde a limitações identificadas na literatura, articulando redução de dimensio-

nalidade, agrupamento não supervisionado e avaliação de eficiência aplicadas a uma base pública de grande escala, alinhando-se às discussões atuais da comunidade de Informática na Educação.

O recorte temporal do estudo concentra-se nos microdados do SAEB 2021 em razão de sua disponibilidade completa, estabilidade metodológica e ampla utilização em estudos recentes. Dessa forma, a escolha do SAEB 2021, decorre da disponibilidade efetiva dos dados no momento da condução do estudo, assegurando a consistência metodológica e a reprodutibilidade do *pipeline* proposto.

Nesse contexto, observa-se que os estudos existentes tendem a empregar técnicas de análise de eficiência, como a Análise Envoltória de Dados, ou métodos de segmentação, como algoritmos de clusterização, de forma isolada, o que limita a capacidade de capturar simultaneamente a heterogeneidade estrutural das escolas e a avaliação comparativa de desempenho. Trabalhos anteriores concentram-se, em geral, na mensuração global da eficiência ou na identificação de perfis escolares, sem integrar mecanismos de redução de dimensionalidade, formação de grupos homogêneos e avaliação relativa intragrupo em um único arcabouço metodológico. É precisamente essa lacuna que o presente estudo busca endereçar, ao propor um *pipeline* integrado que combina PCA, k-means++ e DEA, aplicado a microdados educacionais em larga escala.

A organização do artigo segue o seguinte formato: após esta introdução, são apresentados os referenciais teóricos que fundamentam as técnicas utilizadas, a metodologia para a coleta e análise dos dados, os resultados obtidos com a aplicação do modelo proposto, seguidos da discussão crítica e das conclusões que apontam as contribuições e perspectivas futuras desta pesquisa.

2 Fundamentação Teórica

A avaliação da eficiência em sistemas educacionais é um campo interdisciplinar que envolve conceitos e métodos provenientes da administração, estatística, ciência da computação e educação. O Sistema Nacional de Avaliação da Educação Básica (SAEB), gerido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), desempenha papel fundamental ao fornecer dados abrangentes sobre o desempenho dos estudantes brasileiros em múltiplas disciplinas e distintas regiões (Brasil, 2023; Campos & Vieira, 2021). Tais dados, combinados com informações contextuais recolhidas por meio de questionários aplicados a professores, gestores e alunos, oferecem insumos essenciais para análises que buscam compreender as variabilidades do sistema educacional e permitem integrar múltiplos fatores institucionais na interpretação dos resultados, o que é fundamental para avaliações contextualizadas.

No Brasil, apesar dos avanços observados em indicadores como o Índice de Desenvolvimento da Educação Básica (IDEB), persistem desafios estruturais importantes que impactam a eficácia das políticas educacionais (Miranda & Miranda, 2018; Pereira et al., 2020). A heterogeneidade socioeconômica, as distintas condições de infraestrutura e a variabilidade na formação docente configuram barreiras que dificultam intervenções uniformes e eficazes, reforçando a necessidade de metodologias que considerem essas particularidades na avaliação do desempenho e da eficiência das escolas (Rassouli-Currier, 2007; Soares et al., 2023). Essa perspectiva reforça a necessidade de avaliações que incorporem simultaneamente recursos institucionais, condições de

funcionamento e perfis socioeconômicos, conforme a noção de avaliação justa e contextualizada adotada neste estudo.

Dentre as técnicas analíticas empregadas, destaca-se a Análise por Envoltória de Dados (DEA), método não paramétrico amplamente utilizado para medir a eficiência relativa de unidades produtivas, também aplicável às instituições educacionais (Mariano et al., 2006; Onusic et al., 2007). A DEA permite comparar múltiplas instituições, denominadas Unidades Tomadoras de Decisão (DMUs), considerando simultaneamente diversos insumos (inputs) e produtos (outputs), construindo uma "fronteira eficiente" que serve de referência para as demais unidades. Modelos clássicos da DEA, como o (Charnes, Cooper e Rhodes), baseado em retornos constantes de escala, e o BCC (Banker, Charnes e Cooper), que considera retornos variáveis, podem ser aplicados conforme as características operacionais das entidades avaliadas (Banker et al., 1984; Charnes et al., 1978). No contexto educacional, insumos podem incluir fatores como recursos humanos e materiais, enquanto outputs referem-se a métricas de desempenho estudantil, taxas de aprovação e outros indicadores acadêmicos (Nogueira et al., 2012; Périco et al., 2008).

Em termos formais, a eficiência em modelos DEA orientados a produto é definida como a razão ponderada entre outputs e inputs, conforme estabelecido nos modelos CCR e BCC (Banker et al., 1984; Charnes et al., 1978). Uma unidade é considerada eficiente quando atinge a maior relação possível frente à fronteira construída empiricamente, interpretação consolidada nas revisões de Cook e Zhu (2014).

A literatura recente tem mostrado que a aplicação da DEA em bases extensas, caracterizadas por grande número de variáveis ou estruturas heterogêneas, pode reduzir a capacidade discriminatória do modelo, produzindo avaliações infladas ou pouco estáveis Zhu (2022). Esse fenômeno ocorre especialmente quando há colinearidade ou redundância entre os insumos e produtos analisados, o que reforça a necessidade de procedimentos prévios de seleção, transformação ou redução de dimensionalidade.

Esse efeito pode ser observado de forma concreta no contexto educacional ao se avaliar um grande conjunto de escolas públicas utilizando simultaneamente múltiplos insumos, como características de infraestrutura, perfil docente e condições socioeconômicas, e múltiplos produtos, como proficiências médias, taxas de participação e indicadores de fluxo escolar. Em bases extensas como o SAEB, a elevada heterogeneidade entre escolas e o grande número de variáveis tendem a gerar uma fronteira de eficiência excessivamente flexível, fazendo com que um número elevado de unidades seja classificado como eficiente, ainda que apresentem desempenhos significativamente distintos em termos absolutos. Como consequência, a DEA perde capacidade discriminatória, dificultando a identificação de diferenças substantivas de eficiência entre instituições educacionais.

Esse comportamento evidencia a necessidade de estratégias que restrinjam a comparação a unidades com contextos operacionais semelhantes, motivando abordagens baseadas na formação prévia de grupos homogêneos antes da aplicação da DEA.

Pesquisas sobre avaliação de serviços públicos baseados em dados abertos destacam ainda que a consistência dos resultados em DEA depende fortemente da qualidade da curadoria dos atributos utilizados, especialmente quando as variáveis derivam de diferentes instrumentos de coleta ou apresentam escalas distintas Bartolacci et al. (2024). Esse ponto é particularmente relevante em

bases como o SAEB, que integram simultaneamente informações contextuais, de infraestrutura e de desempenho.

No presente estudo, a eficiência relativa é compreendida como a posição de cada escola em relação à fronteira eficiente construída a partir das unidades que demonstram melhor relação entre insumos e outputs dentro da amostra, permitindo identificar o quanto cada instituição pode melhorar sem alterar suas condições estruturais. Essa interpretação é amplamente utilizada nos estudos de eficiência educacional no Brasil e se alinha às recomendações metodológicas presentes em análises voltadas ao setor público

A clusterização via aprendizado não supervisionado tem se mostrado uma poderosa abordagem para agrupar escolas com características e contextos similares, facilitando análises mais segmentadas e equitativas da eficiência. O algoritmo k-means, amplamente reconhecido por sua simplicidade, escalabilidade e efetividade em grandes bases de dados, é frequentemente utilizado nesses estudos (Borba, 2019; Sinaga & Yang, 2020). Utilizado em conjunto com técnicas de redução de dimensionalidade, como a Análise de Componentes Principais (PCA), que visa eliminar redundâncias e reduzir o número de variáveis sem perda significativa de informação (Hongyu et al., 2016; Varella, 2008), o k-means possibilita organizar unidades escolares em clusters coerentes para posterior avaliação de eficiência.

A definição do número de clusters em métodos particionais é fundamentada em métricas de validade interna amplamente adotadas na literatura, como o coeficiente de silhueta (Rousseeuw, 1987), o índice Davies–Bouldin (Davies & Bouldin, 2009) e o índice Calinski–Harabasz (Caliński & Harabasz, 1974). Revisões como a de Halkidi et al. (2001) destacam que esses critérios são essenciais para bases heterogêneas, como as educacionais.

No plano conceitual, a normalização Min–Max é recomendada para aproximar escalas antes da PCA, evitando que variáveis de maior magnitude dominem a variância total (Hongyu et al., 2016; Varella, 2008). A padronização subsequente via Z-score, discutida em Mohamad e Usman (2013), contribui para que os componentes resultantes operem em escalas comparáveis antes da clusterização, prática também apontada em revisões sobre PCA e alta dimensionalidade (Jolliffe & Cadima, 2016; Sinaga & Yang, 2020). Essa combinação favorece componentes mais estáveis e interpretáveis.

Pesquisas aplicadas ao contexto educacional brasileiro reforçam essa abordagem: (Pereira et al., 2020) demonstram que a clusterização é capaz de segmentar municípios com maior precisão do que agrupamentos espaciais tradicionais, enquanto (Vilaça, 2023) valida a combinação k-means $\rightarrow$ DEA para identificar padrões de desempenho escolar em sistemas educacionais estaduais.

No âmbito internacional, estudos como (Rassouli-Currier, 2007) e (Miranda & Miranda, 2018) evidenciam que a segmentação prévia das unidades avaliadas, seja por fatores socioeconômicos ou condições operacionais, contribui diretamente para análises de eficiência mais robustas, reforçando a pertinência do uso de técnicas não supervisionadas em conjunto com a DEA.

Além disso, abordagens recentes em serviços públicos e bases abertas mostram que a integração entre clusterização e métodos de eficiência torna os diagnósticos mais consistentes em ambientes heterogêneos, especialmente em bases de grande volume e variáveis diversas (Bartolacci et al., 2024). Do ponto de vista computacional, a literatura de DEA sob Big Data destaca que

a formação de grupos homogêneos antes da avaliação reduz problemas de dispersão, melhora a capacidade discriminatória e estabiliza os resultados dos modelos (Zhu, 2022).

Estudos contemporâneos em mineração de dados educacionais também reforçam esse movimento: autores como (Soares et al., 2023) mostram que algoritmos de agrupamento permitem identificar padrões estruturais e socioeconômicos determinantes da qualidade educacional, enquanto trabalhos de (Knox et al., 2020) e (Leporace, 2023) contextualizam o papel do aprendizado não supervisionado em ambientes educacionais caracterizados por bases volumosas e variáveis de natureza híbrida.

A aplicação da DEA após a formação de grupos homogêneos encontra respaldo direto em estudos que investigam a sensibilidade da fronteira eficiente à heterogeneidade estrutural. Trabalhos como Rassouli-Currier (2007) e Miranda e Miranda (2018) mostram que desigualdades socioeconômicas afetam os escores quando todas as unidades são avaliadas conjuntamente. Em bases de maior escala, Zhu (2022) e Bartolacci et al. (2024) demonstram que a segmentação prévia aumenta a estabilidade da eficiência estimada. Evidências empíricas adicionais são apresentadas em Ersoy (2021) e, no cenário brasileiro, por Vilaça (2023), reforçando que a clusterização prévia evita distorções e produz comparações mais justas entre unidades similares.

Com base nesse conjunto de evidências, observa-se que a integração entre PCA, k-means e DEA não apenas tem suporte empírico, como constitui um caminho metodológico coerente para lidar com a elevada heterogeneidade e complexidade dos microdados do SAEB, especialmente quando a análise busca estabelecer comparações mais justas entre escolas com perfis operacionais semelhantes.

Antes da aplicação das técnicas analíticas, bases educacionais extensas demandam procedimentos de preparação reconhecidos na literatura. Variáveis com variância nula ou redundância elevada devem ser removidas ou transformadas, prática fundamentada pelas recomendações clássicas de Jolliffe (2002) e Jolliffe e Cadima (2016). A presença de valores ausentes exige imputação consistente, uma vez que algoritmos como PCA e k-means pressupõem matrizes completas (Iodice D'Enza et al., 2022). Esse cuidado é particularmente relevante em bases públicas amplas, nas quais colinearidade e heterogeneidade podem comprometer discriminação e estabilidade dos modelos, como demonstram Zhu (2022) e Bartolacci et al. (2024).

A manipulação adequada de variáveis categóricas, predominantes em dados educacionais contextuais, apresenta desafios específicos, uma vez que muitos algoritmos operam em espaços métricos numéricos contínuos. Estratégias como o One-Hot Encoding e a codificação ordinal são empregadas para viabilizar a inclusão dessas variáveis nos modelos de clusterização, garantindo que a representação dos dados preserve as relações semânticas e evite distorções (Dahouda & Joe, 2021; Hancock & Khoshgoftaar, 2020; Khatun et al., 2025). No caso dos microdados do SAEB, esse tratamento se torna ainda mais relevante devido à grande quantidade de itens qualitativos presentes nos questionários contextuais.

No âmbito da educação, o desenvolvimento do pensamento computacional tem sido destacado como uma competência essencial para preparar estudantes para os desafios do século XXI (Guan et al., 2020; Roll & Wylie, 2016). A incorporação de metodologias ativas e o uso de tecnologias digitais, incluindo a aprendizagem baseada em projetos, são instrumentos eficazes para promover maior engajamento e potencializar resultados acadêmicos (Leporace, 2023; Park & Kwon, 2024). Essa perspectiva fortalece a importância de análises educacionais que aliem dados

quantitativos e qualitativos, contribuindo para políticas públicas alinhadas com as demandas contemporâneas. Tais abordagens reforçam a necessidade de compreender como diferentes fatores institucionais interagem para gerar condições favoráveis ao desempenho escolar, o que sustenta o uso de técnicas capazes de capturar múltiplas dimensões das escolas.

Este trabalho, portanto, integra essas abordagens metodológicas e conceituais para propor um modelo híbrido de análise da eficiência em escolas públicas de ensino médio, articulando procedimentos de clusterização, redução de dimensionalidade e análise por envoltória de dados para gerar diagnósticos mais precisos e contextualizados. A fundamentação aqui apresentada sustenta a escolha das técnicas e destaca sua relevância para a otimização da gestão educacional no Brasil. A revisão teórica apresentada oferece suporte às escolhas metodológicas feitas no estudo e evidencia lacunas relevantes da literatura, como a escassez de aplicações envolvendo PCA, k-means e DEA sobre bases públicas nacionais em larga escala, especialmente com o objetivo de formar grupos homogêneos que permitam avaliações comparativas mais adequadas.

3 Trabalhos Relacionados

Esta seção reúne pesquisas que abordam, sob diferentes perspectivas, a análise de eficiência em sistemas educacionais, a formação de grupos semelhantes por aprendizado não supervisionado e o uso de técnicas de redução de dimensionalidade em bases extensas. A revisão articula três eixos centrais: estudos que aplicam DEA para avaliar desempenho institucional; trabalhos que utilizam clusterização para segmentar unidades educacionais segundo seus perfis contextuais e de desempenho; e investigações que discutem os desafios estatísticos de lidar com grande número de variáveis, destacando o papel do PCA na mitigação de redundância e colinearidade. Esse conjunto de referências fornece o panorama necessário para situar a proposta deste artigo e identificar as lacunas que motivam a integração entre PCA, k-means e DEA.

3.1 DEA e análises de eficiência em sistemas educacionais

O estudo clássico de Rassouli-Currier (2007) avaliou 354 distritos escolares de Oklahoma utilizando modelos CCR e BCC, complementados por regressões Tobit para incorporar fatores socioeconômicos. Os resultados evidenciam que desigualdades sociais afetam diretamente os escores de eficiência, reforçando a importância de métodos que controlem heterogeneidade estrutural. Essa conclusão fundamenta a adoção de agrupamento prévio neste estudo como forma de obter comparações mais equitativas entre unidades.

Em Miranda e Miranda (2018), foi proposto o índice LOED, também baseado em DEA, aplicado à rede municipal de Campinas. A combinação entre indicadores operacionais e variáveis contextuais permitiu identificar metas de melhoria para cada escola. Apesar disso, o estudo opera com número reduzido de variáveis e sem técnicas de clusterização ou redução de dimensionalidade, o que limita sua aplicação a bases extensas como o SAEB.

3.2 Clusterização aplicada à educação no Brasil

Em Pereira et al. (2020), o algoritmo *k*-means foi utilizado para investigar se a divisão administrativa de escolas cearenses em CREDEs/SEFOR corresponde à similaridade educacional observada. Os autores demonstram que agrupamentos orientados por desempenho refletem de forma mais fiel a realidade escolar do que agrupamentos geográficos. O estudo oferece evidências relevantes sobre a utilidade da clusterização no contexto educacional, embora não integre qualquer técnica de eficiência.

O trabalho de Vilaça (2023) é uma das primeiras aplicações nacionais que combinam k-means com DEA, avaliando a eficiência de sistemas estaduais. A clusterização prévia mostrou-se útil para melhorar a comparabilidade entre unidades. Entretanto, o estudo utiliza poucas variáveis, não aplica PCA e não executa tratamento aprofundado de variáveis categóricas, o que restringe sua capacidade de generalização em bases amplas.

3.3 Trabalhos da comunidade brasileira de Informática na Educação

Diversos estudos publicados nos principais eventos da área reforçam o uso de aprendizado não supervisionado para compreensão de padrões educacionais. Em Borba (2019), a clusterização foi utilizada para identificar perfis de estudantes em ambientes virtuais de aprendizagem, destacando a capacidade do k-means em revelar heterogeneidades comportamentais relevantes. O estudo de Soares et al. (2023), publicado na RBIE, empregou mineração de dados para explicar determinantes da qualidade escolar no Maranhão, mostrando que infraestrutura, práticas docentes e contexto socioeconômico explicam parte substantiva da variância nos resultados educacionais. Embora esses trabalhos contribuam para o uso de técnicas de agrupamento na educação, nenhum deles combina clusterização com DEA nem aplica redução de dimensionalidade a bases educacionais de larga escala.

3.4 PCA e redução de dimensionalidade em bases de alta complexidade

A utilização da PCA neste estudo encontra sustentação em pesquisas que abordam problemas de alta dimensionalidade. O trabalho de Battisti e Carvalho (2022) apresenta o método TPCA, uma extensão supervisionada da PCA clássica, e demonstra empiricamente que redundâncias e ruído podem comprometer algoritmos preditivos quando não há transformação prévia das variáveis. Embora o estudo se concentre em classificação, suas conclusões são diretamente aplicáveis ao cenário dos microdados educacionais, fortemente caracterizados por colinearidade entre atributos

No contexto internacional, Zhu (2022) argumentam que a DEA perde capacidade discriminatória quando aplicada a bases com muitas variáveis, gerando escores artificialmente elevados Bartolacci et al. (2024) reforçam essa perspectiva ao avaliar serviços públicos baseados em dados abertos, destacando a importância da curadoria e transformação de atributos. Tais evidências justificam a inclusão da PCA como etapa essencial do *pipeline* proposto

3.5 Integração entre clusterização, redução de dimensionalidade e DEA

Estudos internacionais começam a integrar clusterização e DEA, como em Ersoy (2021), que avalia departamentos universitários após agrupamento, mas sem uso de PCA e com amostras reduzidas. Pesquisas em Big Data (Bartolacci et al., 2024) indicam que agrupamentos homogêneos aumentam a estabilidade dos escores de eficiência, reforçando a pertinência dessa integração. No entanto, não foram identificadas aplicações dessa abordagem ao contexto dos microdados do SAEB, nem em bases brasileiras de larga escala.

3.6 Síntese e lacunas da literatura

A partir da revisão apresentada, observam-se três lacunas principais:

1. não foram identificados estudos brasileiros que integrem PCA, *k*-means e DEA em um único pipeline aplicado aos microdados completos do SAEB;
2. o uso conjunto de clusterização e DEA é raro e, quando presente, utiliza poucas variáveis ou não considera heterogeneidade contextual;
3. trabalhos nacionais e internacionais analisam eficiência ou agrupamento isoladamente, mas não avaliam eficiência dentro de grupos homogêneos formados por aprendizado não supervisionado.

Este estudo contribui ao abordar simultaneamente essas lacunas, propondo uma abordagem híbrida capaz de articular redução de dimensionalidade, clusterização e análise de eficiência em bases educacionais extensas e heterogêneas

Com o objetivo de sistematizar as abordagens identificadas na literatura e explicitar as diferenças metodológicas entre os estudos analisados, apresenta-se, a seguir, uma síntese comparativa das principais características técnicas, escopo de aplicação e limitações observadas. A Tabela 1 organiza essas informações de forma estruturada, permitindo visualizar as convergências e divergências entre as pesquisas relacionadas e a proposta deste artigo, bem como evidenciar as lacunas que motivam a integração entre PCA, *k*-means e DEA no contexto dos microdados do SAEB.

Tabela 1: Comparação entre os trabalhos relacionados e a proposta deste estudo.

Autor/Ano	Base/Contexto	DEA	Cluster	PCA	Escala	Limitação Principal
Rassouli-Currier, 2007	Distritos escolares (EUA)	Sim	Não	Não	Distritos	Não realiza segmentação prévia por similaridade estrutural
Miranda e Miranda, 2018	Rede municipal (Campinas)	Sim	Não	Não	Escolas	Poucas variáveis e ausência de técnicas de redução dimensional
Pereira et al., 2020	Escolas do Ceará	Não	Sim (k-means)	Não	Escolas	Não integra análise de eficiência
Vilaça, 2023	Sistemas estaduais (Brasil)	Sim	Sim (k-means)	Não	Estados	Utiliza número reduzido de variáveis e não aplica PCA
Borba, 2019	Ambientes virtuais	Não	Sim (k-means)	Não	Estudantes	Foco comportamental, sem avaliação de eficiência institucional
Soares et al., 2023	Escolas (Maranhão)	Não	Sim	Não	Escolas	Não combina clusterização com DEA
Ersoy, 2021	Departamentos universitários	Sim	Sim	Não	Departa-mentos	Amostra reduzida e ausência de redução de dimensionalidade
Zhu, 2022; Bartolacci et al., 2024	Serviços públicos / Big Data	Sim	Parcial	Não	Instituições	Evidenciam problema de alta dimensionalidade, mas não aplicam pipeline integrado
Este estudo	Microdados SAEB 2021 (CE)	**Sim**	**Sim (k-means)**	**Sim**	**Escolas**	Integra PCA, clusterização e DEA em pipeline único aplicado a base de larga escala

4 Metodologia

Este estudo propõe uma abordagem quantitativa híbrida e multietapas para organizar as escolas públicas de ensino médio do Ceará em contextos homogêneos e, em seguida, estimar a eficiência relativa dentro de cada contexto. A estratégia integra três componentes centrais: redução de dimensionalidade via Análise de Componentes Principais (PCA), agrupamento por k-means (k-means++) e mensuração da eficiência por meio da DEA. O recorte da pesquisa compreende exclusivamente escolas públicas do Ceará avaliadas no SAEB 2021, apoiando-se nas tabelas TS_Professor e TS_Escola disponibilizadas pelo INEP, que fornecem dados contextuais e indicadores de desempenho escolar utilizados durante o processo analítico (Banker et al., 1984; Brasil, 1996, 2023; Caliński & Harabasz, 1974; Charnes et al., 1978; James et al., 2013). Além disso, todas as análises foram conduzidas exclusivamente com microdados públicos e anonimizados, garantindo conformidade com a LGPD e dispensando submissão ao Comitê de Ética em Pesquisa, conforme diretrizes do Inep para uso secundário de dados educacionais.

A Figura 1 ilustra o *pipeline* metodológico adotado, detalhando o fluxo sequencial das etapas que sustentam as análises: seleção e curadoria das bases de dados, processamento e transformação

das variáveis, mineração inteligente visando a formação de grupos homogêneos e análise comparativa da eficiência intragrupos com DEA. Esta representação visual guia a compreensão do encadeamento lógico do estudo e sua estrutura integradora. A figura foi adaptada de Vilaça, 2023 incorporando explicitamente as etapas de normalização Min–Max, padronização Z-score, validação interna dos clusters e separação entre variáveis utilizadas para clusterização e para a DEA.

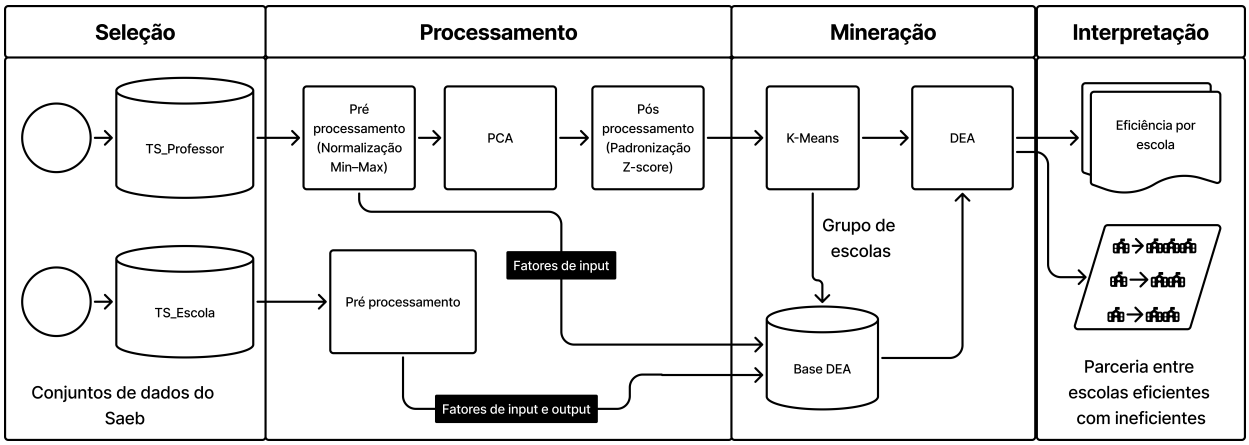

Figura 1: Modelo para otimização da eficiência em sistemas homogêneos. Fonte: Adaptado de Vilaça, 2023.

A seleção inicial dos dados envolveu a identificação e filtragem das bases TS_Professor e TS_Escola, escolhidas pela complementaridade das informações. A base TS_Professor oferece um conjunto de variáveis qualitativas provenientes de questionários aplicados a docentes de escolas públicas, contemplando aspectos como formação, práticas pedagógicas, infraestrutura, uso de recursos e perfil discente. As variáveis do questionário docente são majoritariamente ordinais e de escala Likert; por essa razão, após a codificação numérica, a agregação por média ao nível da escola foi adotada como aproximação consolidada do perfil institucional, prática recorrente em estudos educacionais baseados em dados amostrais docentes A base TS_Escola, por sua vez, reúne informações sobre matrículas, participação na avaliação e distribuição das proficiências em Matemática no ensino médio. As variáveis oriundas da base TS_Escola foram categorizadas conforme sua natureza analítica: atributos contextuais e institucionais foram utilizados exclusivamente na formação dos clusters, enquanto indicadores de desempenho educacional, como médias de proficiência e distribuições percentuais por níveis da escala SAEB, foram reservados apenas como outputs na etapa da DEA. Essa separação assegura que a segmentação por contextos homogêneos anteceda a avaliação de eficiência, evitando circularidade entre agrupamento e desempenho Um processo de filtragem garantiu a inclusão apenas de escolas públicas do Ceará com registros completos e consistentes. A Tabela 2 apresenta um exemplo das variáveis da base TS_Professor utilizadas na construção do perfil institucional das escolas.

O dicionário completo das variáveis do SAEB 2021 (TS_Professor e TS_Escola), incluindo códigos de resposta e descrições originais, encontra-se disponível no material suplementar e no repositório do projeto.

Tabela 2: Exemplo de variáveis da base TS_Professor utilizadas na construção do perfil institucional das escolas. Fonte: autoria própria.

Variável	Tipo	Descrição resumida
TX_RESP_Q005	Ordinal	Tempo de atuação do docente como professor (faixas de anos)
TX_RESP_Q008	Categórica	Tipo de vínculo trabalhista do docente na escola
TX_RESP_Q025	Ordinal	Avaliação do tamanho da sala em relação ao número de alunos
TX_RESP_Q039	Ordinal	Adequação percebida do acesso à internet na escola
TX_RESP_Q074	Ordinal	Frequência de reuniões do conselho de classe no ano
TX_RESP_Q086	Ordinal	Frequência de colaboração das famílias na aprendizagem
TX_RESP_Q095	Ordinal	Interesse dos alunos pela disciplina lecionada

Inicialmente, a base TS_Escola continha 640 escolas do Ceará no SAEB 2021. Após verificações de completude e consistência, 598 escolas permaneceram no conjunto final utilizado nas análises. A base TS_Professor registrava 8.214 docentes vinculados a essas escolas; após o pareamento por código da unidade, eliminação de duplicidades e exclusão de questionários incompletos, 7.932 registros foram agregados ao nível escolar.

A integração das bases resultou em um conjunto inicial composto por variáveis docentes e institucionais em grande volume. Após codificação, agregação e exclusão de atributos com variância nula ou redundância semântica, a base consolidada passou a conter 397 variáveis válidas, evidenciando a necessidade de técnicas de redução de dimensionalidade para estabilizar as análises subsequentes.

Foram aplicados protocolos de qualidade aos dados no tratamento inicial. Conforme discutido anteriormente, atributos com variância nula ou invariantes no recorte estadual foram removidos, como variáveis que apresentaram valores idênticos para todas as escolas do Ceará, impossibilitando qualquer poder discriminativo. Variáveis categóricas nominais foram transformadas por one-hot encoding, enquanto variáveis ordinais e de escala Likert foram convertidas para valores numéricos preservando sua hierarquia, garantindo adequação estatística às etapas seguintes.

A verificação automática de valores ausentes indicou ausência de lacunas na matriz final utilizada para análise, condição essencial para a aplicação de PCA e k-means, que requerem matrizes completas.

Dada a pluralidade de registros docentes por escola, as respostas foram agregadas ao nível institucional, utilizando médias para variáveis quantitativas e ordinais, proporções para variáveis binárias e contagens para indicadores de composição. Essa agregação reduz ruído individual e reforça a representatividade do perfil institucional de cada escola.

Para assegurar equilíbrio na influência das variáveis, aplicou-se uma estratégia sequencial. Inicialmente, foi realizada a normalização Min–Max, reescalando todas as variáveis contínuas para o intervalo [0,1], com o objetivo de eliminar distorções decorrentes de diferenças de ordem

de grandeza entre atributos, condição necessária para a correta extração das componentes principais. Em seguida, antes da etapa de clusterização, aplicou-se a padronização Z-score, assegurando média zero e desvio padrão unitário, requisito fundamental para o cálculo consistente de distâncias euclidianas no algoritmo k-means. Esta sequência buscou otimizar cada etapa analítica segundo seus princípios estatísticos e computacionais clássicos (Hongyu et al., 2016; James et al., 2013; Varella, 2008).

A padronização Z-score foi empregada exclusivamente com finalidade geométrica, visando à equalização das escalas das componentes principais para o cálculo de distâncias euclidianas no k-means, e não como pressuposto inferencial de normalidade. Essa etapa não altera a ortogonalidade das componentes nem o subespaço gerado pela PCA, atuando apenas como reescalonamento para a etapa de agrupamento.

A adoção de PCA seguida de k-means++ visa conciliar robustez estatística, interpretabilidade e custo computacional. A PCA reduz multicolinearidade e ruído preservando a maior parte da variância (critério de 90%), prática consolidada na literatura (Jolliffe & Cadima, 2016). Além disso, há base teórica para sua combinação com k-means: projeções principais guardam relação com as soluções contínuas do objetivo do k-means, o que justifica a redução antes da partição (Ding & He, 2004). O k-means++ foi escolhido pela inicialização probabilística com garantias de aproximação e menor variabilidade entre execuções (Arthur & Vassilvitskii, 2007). A avaliação interna por CH e silhueta segue recomendações clássicas para aferir coesão e separação de grupos (Caliński & Harabasz, 1974; Rousseeuw, 1987). Optou-se por não empregar métodos de densidade (e.g., DBSCAN) ou hierárquicos: o primeiro requer calibração sensível de ε e *MinPts* e é voltado a formas arbitrárias em baixa dimensionalidade, o que tende a perder eficácia após projeções e escalonamentos globais (Ester et al., 1996); já os métodos hierárquicos apresentam maior sensibilidade a ruído e custo quadrático em amostras extensas, além de menor transparência operacional para *benchmarking* escolar (Murtagh & Contreras, 2012). Como *baseline*, comparamos o arranjo proposto a segmentações alternativas (sem PCA e geográfico-administrativas), reportando ganhos consistentes nas métricas internas de qualidade, o que sustenta empiricamente a decisão metodológica.

A mensuração da eficiência relativa das escolas dentro de cada cluster foi realizada por meio do modelo BCC da DEA, orientado a outputs, estrutura adequada à maximização dos resultados educacionais diante dos insumos observados (Banker et al., 1984; Charnes et al., 1978). As variáveis de entrada foram selecionadas como proxies das condições institucionais, da composição e qualificação do corpo docente e do perfil socioeconômico dos estudantes, enquanto as variáveis de saída abrangeram indicadores de desempenho escolar.

Neste estudo, os outputs considerados na DEA restringiram-se exclusivamente à disciplina de Matemática, utilizando a proficiência média e a distribuição percentual por níveis da escala SAEB. Essa decisão fundamenta-se em três critérios principais: (i) estabilidade e padronização da métrica, que favorecem comparabilidade e definição de metas; (ii) maior sensibilidade diagnóstica da disciplina no contexto nacional; e (iii) completude dos microdados no recorte analisado.

Essa opção configura uma linha de base extensível: o *pipeline* permite incorporar, em trabalhos futuros, Língua Portuguesa e/ou índices compostos de aprendizagem. As proficiências não participaram da etapa de clusterização (foram usadas apenas na DEA), evitando circularidade entre segmentação e avaliação de eficiência. Uma importante transformação adotada foi a

inversão da métrica referente à proporção de alunos em níveis iniciais de proficiência, de forma a refletir corretamente o objetivo de maximizar aprendizagem no modelo. Essa análise segmentada por contexto torna as fronteiras de eficiência mais coerentes, estabelecendo benchmark internos factíveis para cada grupo de escolas (Gharakhani et al., 2011; Giacomello & Oliveira, 2014).

As variáveis de desempenho utilizadas como outputs não participaram da etapa de clusterização, sendo empregadas exclusivamente na DEA. Essa separação evita circularidade entre a formação dos grupos e a avaliação de eficiência, assegurando que a segmentação por contextos homogêneos anteceda e fundamente a construção das fronteiras de eficiência. Como resultado, obtêm-se benchmarks internos mais realistas e factíveis para cada grupo de escolas (Gharakhani et al., 2011; Giacomello & Oliveira, 2014).

A Figura 1 sintetiza o *pipeline* analítico adotado, conectando as etapas, seus propósitos, dados de entrada e artefatos gerados, funcionando como guia condensado para replicação e verificação dos procedimentos metodológicos. A Tabela 3 complementa essa visão ao correlacionar, de forma sistemática, cada etapa do método às suas entradas e saídas principais. Toda documentação e scripts foram organizados para garantir transparência e reprodutibilidade dos resultados (Brasil, 1996, 2018).

Tabela 3: Estratégia analítica (pipeline) do estudo. Fonte: autoria própria.

Etapa	**Propósito**	**Entradas**	**Saídas**
Curadoria e pareamento	Recorte para escolas públicas do EM no Ceará; harmonização de chaves; saneamento e padronização.	TS_Professor, TS_Escola	Base escolar pareada
Agregação e preparo	Agregação ao nível da escola; tratamento de ausentes; escalonamento.	Base consolidada	Tabela única
PCA	Redução de dimensionalidade preservando 90% da variância.	Variáveis escalonadas	Componentes principais
k-means++	Formação de contextos homogêneos.	Componentes principais	Clusters
DEA-BCC	Mensuração da eficiência relativa intra-cluster.	Inputs e outputs	Escores de eficiência

Esse encadeamento metodológico define o arcabouço analítico que sustenta as comparações exploradas na seção de resultados.

5 Resultados e Discussões

Nesta seção, apresentam-se os resultados obtidos a partir da aplicação do *pipeline* metodológico, articulando a redução da dimensionalidade, a clusterização das escolas e a análise da eficiência relativa via DEA. O objetivo é evidenciar como essas etapas combinadas permitem comparar escolas em contextos homogêneos e, a partir disso, definir referências e metas factíveis de melhoria, evitando comparações globais distorcidas entre realidades educacionais distintas.

O modelo computacional foi implementado na linguagem Python, em ambiente Linux Mint 21.3, utilizando um processador Intel Core i3 de 11ª geração e 32 GB de RAM. Para manipulação dos dados, foram empregadas bibliotecas renomadas, Pandas, NumPy e Matplotlib, enquanto o tratamento das variáveis categóricas utilizou o pacote category_encoders. Para a redução dimensional e clusterização, a implementação apoiou-se no scikit-learn. A análise de eficiência foi conduzida com o uso da biblioteca PyDEA, e otimizações foram realizadas com Joblib. As bibliotecas Seaborn e Yellowbrick auxiliaram na visualização gráfica, facilitando a validação exploratória dos modelos.

O conjunto de dados inicial consistiu nas bases TS_Professor e TS_Escola do SAEB 2021. Das 135 variáveis originais da TS_Professor, selecionaram-se 74 com relevância para identificar o contexto escolar; da base TS_Escola, com 136 variáveis, foram utilizados 20 referentes ao ensino médio em Matemática. Aplicaram-se filtros rigorosos para incluir exclusivamente escolas públicas do Ceará que ofertam ensino médio e possuem registros completos relevantes para análise, garantindo a representatividade da amostra.

Diante da alta dimensionalidade dos dados, utilizou-se a Análise de Componentes Principais (PCA) para condensar as informações mantendo 90% da variância total. A Figura 2 mostra o comportamento do índice Calinski–Harabasz (CH) para comparação entre clusterizações com e sem PCA. Ao longo do intervalo de 2 a 20 clusters, observa-se que o índice CH sem PCA apresenta valores mais elevados e maior variação, enquanto o uso do PCA tende a reduzir esses valores pela compactação dos dados.

A normalização dos dados, por sua vez, favoreceu uma comparação mais equilibrada entre os cenários com e sem PCA, conforme Figura 3. A partir de cerca de 7 clusters, os valores do índice CH se estabilizam em ambos os métodos, indicando um limiar além do qual aumentos no número de clusters não promovem melhorias significativas na qualidade da segmentação. Essa convergência reforça a robustez da PCA combinada à normalização para permitir clusterizações fiéis ao perfil dos dados.

Com base na análise conjunta do índice Calinski–Harabasz, do coeficiente médio de silhueta e da estabilidade visual dos centróides, adotou-se (k=7) como compromisso entre coesão intragrupo, separação entre grupos e interpretabilidade operacional.

A clusterização foi realizada diretamente sobre os componentes principais resultantes da PCA, e não sobre as variáveis originais. Essa decisão teve como objetivo reduzir a dimensionalidade e eliminar redundâncias entre atributos correlacionados, garantindo maior estabilidade e coesão nos agrupamentos. Como consequência, os clusters passam a refletir padrões latentes no espaço ortogonal das componentes principais, isto é, combinações lineares ponderadas de múltiplos atributos, e não relações diretas e isoladas entre variáveis como “infraestrutura”, “tempo de docência” ou “perfil socioeconômico”.

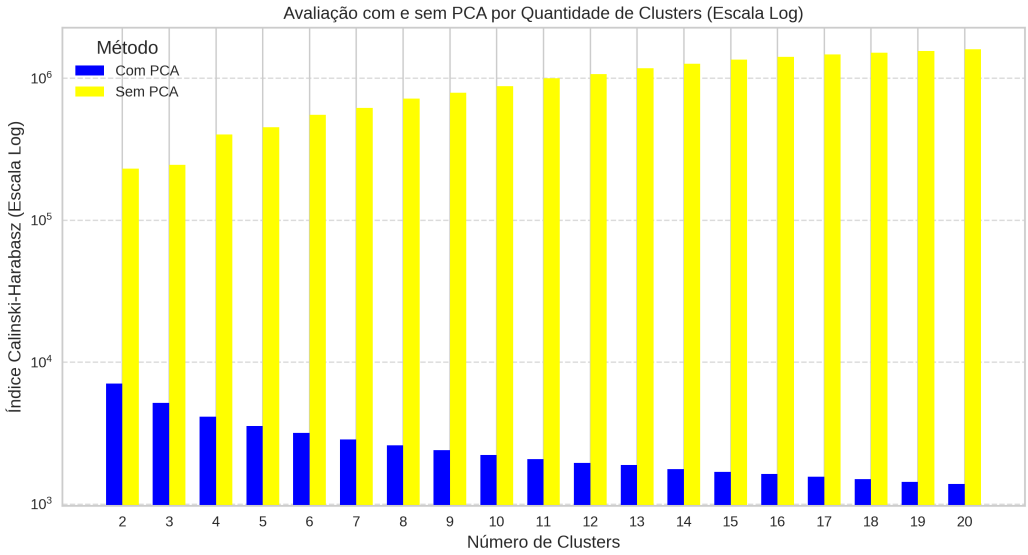

Figura 2: Avaliação do índice Calinski–Harabasz (CH) para clusters com e sem aplicação da PCA em escala logarítmica. Fonte: autoria própria.

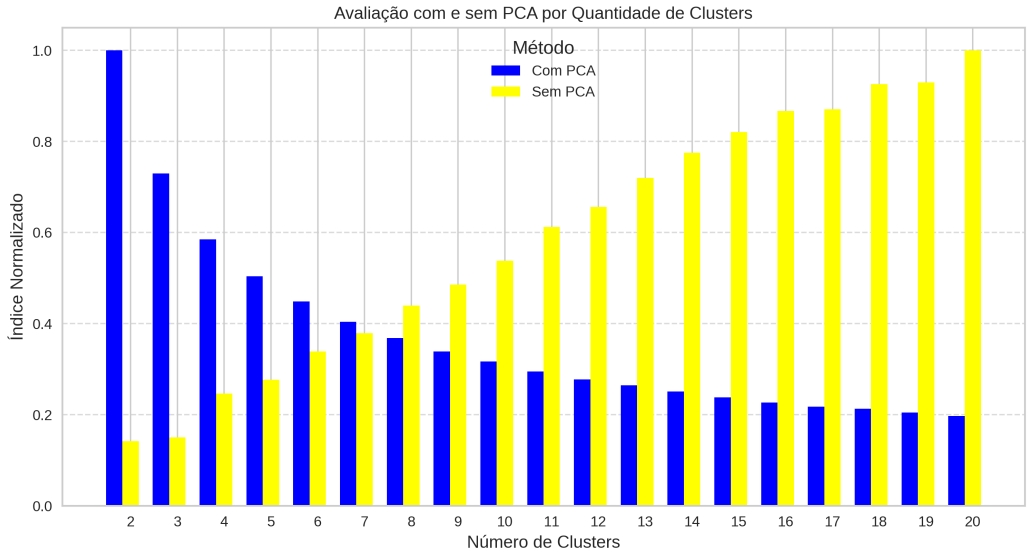

Figura 3: Avaliação do índice Calinski–Harabasz (CH) para clusters com e sem aplicação da PCA após normalização dos dados. Fonte: autoria própria.

Dessa forma, comparações simples entre os clusters e as variáveis originais deixam de ser metodologicamente válidas, uma vez que cada dimensão do espaço projetado sintetiza a contribuição conjunta de dezenas de indicadores. A interpretação apropriada, portanto, deve ocorrer em nível agregado e latente, observando tendências globais, como o comportamento médio de eficiência e a dispersão entre grupos, e não em função de atributos específicos. Essa precaução

assegura consistência estatística e, sobretudo, fundamenta comparações operacionais mais justas, nas quais as escolas passam a ser avaliadas e referenciadas apenas em relação a unidades com perfis efetivamente semelhantes.

A distribuição dos clusters é exibida na Figura 4, ilustrando a dispersão dos grupos em duas dimensões escolhidas conforme análise das componentes principais, com destaque para cada centróide.

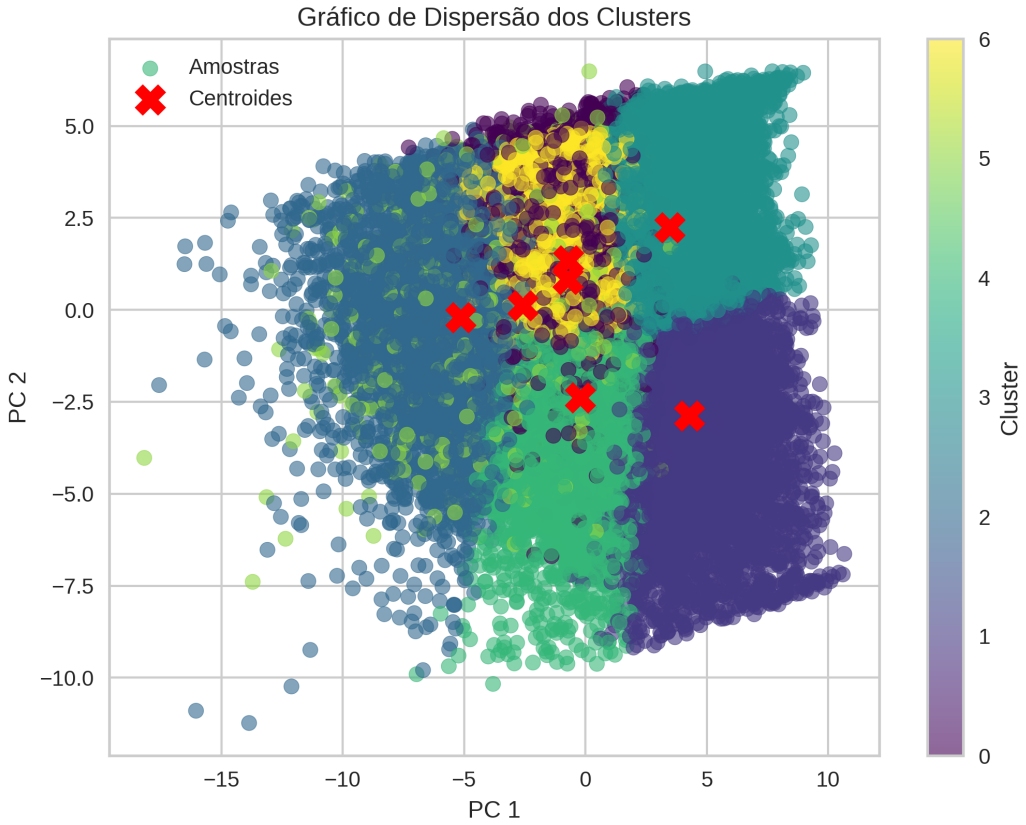

Figura 4: Dispersão bidimensional dos clusters formados pelo k-means, com indicação dos centróides. Fonte: autoria própria.

A avaliação da qualidade dos agrupamentos pelo coeficiente de silhueta é apresentada na Figura 5. O gráfico exibe, para cada cluster, a distribuição dos valores de silhueta das escolas que o compõem, permitindo analisar simultaneamente a coesão interna dos grupos e sua separação em relação aos demais.

A presença de valores de silhueta pontualmente baixos não indica fragilidade do modelo de clusterização, mas reflete a heterogeneidade inerente a bases educacionais reais de alta dimensionalidade, nas quais transições graduais entre perfis institucionais são esperadas. Assim, a avaliação da qualidade dos agrupamentos deve concentrar-se no comportamento agregado das distribuições por cluster, e não em observações individuais isoladas.

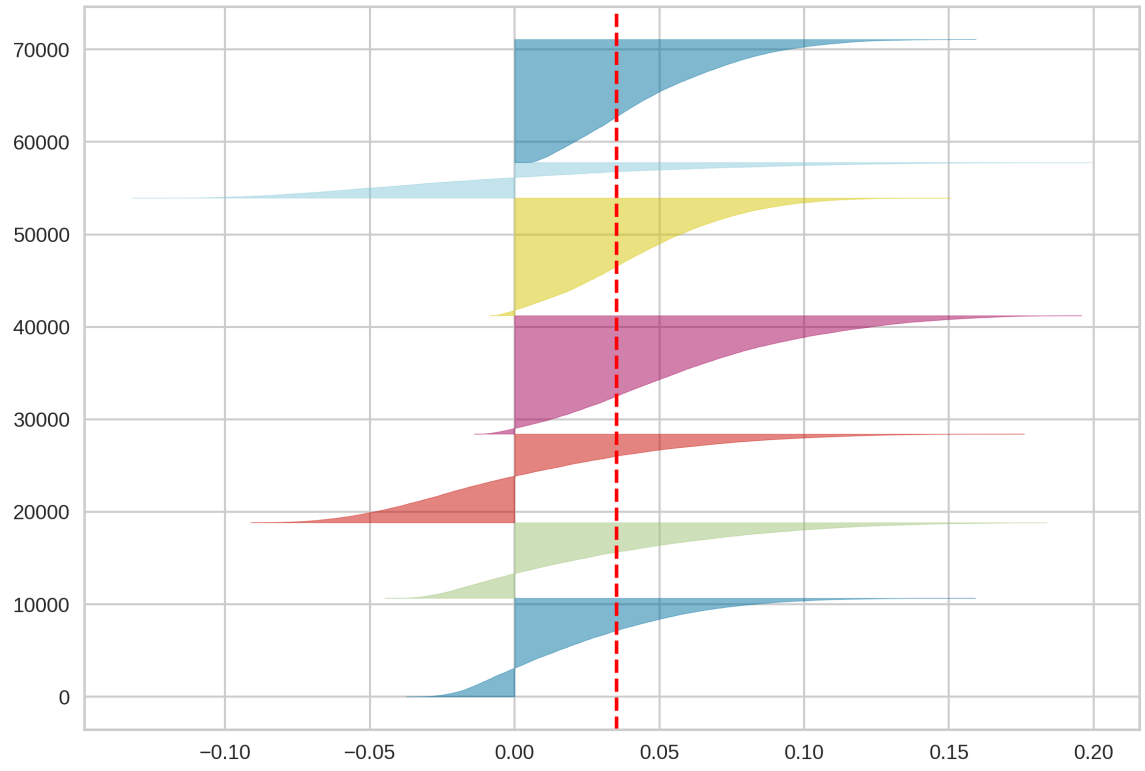

Figura 5: Distribuição dos coeficientes de silhueta das escolas, evidenciando coesão interna e separação entre os clusters. Fonte: autoria própria.

Essa estrutura de agrupamentos constitui um elemento central para a etapa subsequente de análise de eficiência, pois assegura que as metas derivadas pela DEA sejam formuladas a partir de pares efetivamente comparáveis dentro de cada cluster, conferindo maior coerência analítica e utilidade prática aos resultados para fins de diagnóstico e planejamento educacional.

Em linha com esse posicionamento, a aplicação do *pipeline* não apenas melhora métricas de qualidade da segmentação, mas redefine a forma como a eficiência é interpretada. Ao executar a DEA intra-cluster, cada escola passa a ser comparada exclusivamente com pares de perfil semelhante, o que viabiliza a definição de metas proporcionais e operacionalmente alcançáveis, em contraste com fronteiras globais excessivamente abstratas.

Do ponto de vista empírico, essa redefinição da eficiência se refletiu em diferenças sistemáticas nas métricas de qualidade da segmentação e no comportamento da eficiência quando comparada a abordagens de referência, efeitos associados à avaliação de eficiência intra-cluster (camadas de isoeficiência) em contraste com a utilização de uma única fronteira global. Esses achados sugerem que o pré-agrupamento no espaço das componentes principais contribui para contextualizar a aplicação da DEA, permitindo a definição de metas mais aderentes à realidade operacional dos grupos analisados.

A Tabela 4 evidencia os principais atributos extraídos da base TS_Escola submetidos à análise DEA, destacando indicadores cruciais como a adequação da formação docente, nível socioeconômico, matrículas, presença e participação dos alunos e níveis detalhados de proficiência em Matemática. A categorização dos níveis de proficiência em inicial (0 a 4) e final (5 a 10),

com a inversão aplicada no nível inicial para ajustar o modelo orientado a outputs, fundamenta a avaliação da eficiência educacional de forma aderente à dinâmica de progresso escolar.

Tabela 4: Principais atributos da base TS_Escola utilizados na análise DEA. Fonte: autoria própria.

Variável	Tipo	Descrição
FORMAÇÃO DOCENTE	Numérica	Percentual de adequação da formação docente para o Ensino Médio
NÍVEL SOCIOECONÔMICO	Categórica	Índice socioeconômico consolidado da escola (Níveis I a VII)
MATRICULADOS	Numérica	Número de alunos matriculados na etapa avaliada
PRESENTES	Numérica	Número de alunos presentes na aplicação do SAEB
PARTICIPAÇÃO	Numérica	Percentual de participação dos alunos na avaliação
Proficiência — níveis 0 a 10	Numérica	Distribuição dos alunos por nível detalhado de proficiência

O modelo DEA foi aplicado intra-cluster para mensurar a eficiência relativa das escolas, organizando-as em faixas de isoeficiência que variam de 100% (fronteira eficiente) a 30%, detalhado na Tabela 5. Essa estratificação permite posicionar cada escola em relação a pares mais eficientes dentro do mesmo contexto, constituindo a base para a definição de referências e metas de melhoria factíveis.

Tabela 5: Distribuição das escolas por faixas de eficiência DEA em cada cluster. Fonte: autoria própria.

Cluster	100%	90%	80%	70%	60%	50%	40%	30%
0	2466	1973	1776	1480	987	691	296	197
1	2807	2245	2021	1684	1122	786	337	225
2	959	767	691	576	383	269	115	77
3	3768	3015	2713	2261	1508	1055	452	301
4	2528	2023	1821	1517	1012	708	303	202
5	2673	2138	1925	1604	1069	748	320	213
6	2551	2041	1838	1535	1023	717	307	204

A Tabela 5 permite ilustrar de forma concreta como as metas derivadas da Análise por Envoltória de Dados podem ser interpretadas na prática. Em cada cluster homogêneo, as escolas são distribuídas em faixas de eficiência relativas, que variam de 100% (fronteira eficiente) até 30%, refletindo diferentes níveis de desempenho quando comparadas exclusivamente a unidades de perfil semelhante.

Considere, por exemplo, uma escola classificada na faixa de 70% de eficiência dentro de seu cluster. Esse posicionamento indica que a unidade não pertence à fronteira eficiente do grupo, mas encontra-se relativamente próxima a ela. Nesse caso, a DEA identifica como referências diretas as escolas do mesmo cluster situadas nas faixas superiores (80%, 90% ou 100%), que passam a atuar como pares de comparação operacional.

As metas atribuídas a essa escola não correspondem a valores arbitrários ou externos, mas são definidas a partir do desempenho efetivamente observado nesses pares de referência. Assim, o modelo aponta margens factíveis de melhoria nos outputs educacionais considerados, como

avanços graduais na proficiência média em Matemática ou na redistribuição dos estudantes para níveis mais altos da escala SAEB, sempre respeitando os insumos institucionais já disponíveis.

Esse procedimento permite transformar o resultado da DEA em um instrumento prático de diagnóstico e planejamento, no qual cada escola recebe metas proporcionais à sua posição relativa no cluster, viabilizando trajetórias realistas de aprimoramento e evitando comparações distorcidas entre contextos estruturalmente distintos.

Embora a decomposição PCA dificulte a identificação direta das variáveis originais mais influentes em cada grupo, foi possível observar tendências agregadas que refletem padrões consistentes de desempenho e contexto. Em linhas gerais, os clusters com maiores escores de eficiência concentram escolas com médias de proficiência mais elevadas e menor vulnerabilidade socioeconômica, enquanto os grupos menos eficientes reúnem unidades com desafios estruturais e pedagógicos mais pronunciados.

Essa leitura é coerente com os padrões latentes capturados pela PCA, cujas primeiras componentes sintetizam variâncias relacionadas a desempenho acadêmico e estrutura escolar. Desse modo, as diferenças observadas entre clusters traduzem a ação combinada desses fatores, e não a influência isolada de variáveis específicas. Essa compreensão global dos perfis formados fornece base para interpretar, a seguir, o comportamento interno de cada grupo e a variação de eficiência relativa obtida pela DEA.

Os resultados sugerem que grupos mais homogêneos exibem dispersão reduzida e forte concentração de escolas em níveis de eficiência próximos à fronteira, o que caracteriza menor variabilidade operacional. Por outro lado, grupos mais heterogêneos apresentam faixas de desempenho mais amplas, indicando maior margem para intervenções graduais e ganhos factíveis, os quais podem ser orientados por metas definidas a partir das escolas mais eficientes do próprio cluster. Nesses casos, a aplicação da DEA ajuda gestores a identificar pares referenciais plausíveis e a definir metas progressivas de aprimoramento.

A criação de redes colaborativas entre escolas eficientes e aquelas com maior potencial de desenvolvimento emerge como uma estratégia operacional importante, estimulando a troca de boas práticas e o desenvolvimento sistêmico. O modelo fornece bases quantitativas para orientar essas ações, alinhando a análise de eficiência com políticas públicas sustentáveis e contextualizadas.

No que tange às contribuições em relação a trabalhos semelhantes, o modelo híbrido proposto supera abordagens que utilizam apenas DEA como (Miranda & Miranda, 2018; Rassouli-Currier, 2007) ao integrar a redução de dimensionalidade por PCA e clusterização k-means++, ampliando a coesão interna dos grupos. Em comparação a estudos que aplicam apenas k-means para segmentar escolas (Pereira et al., 2020), nossa metodologia incorpora variáveis contextuais e qualitativas do Saeb, refinando a caracterização dos clusters. Diferente do uso conjunto de DEA + TOPSIS em (Ersoy, 2021), que ranqueia eficiência em departamentos de EAD, o presente trabalho articula *benchmarking* intra-cluster e propositura de parcerias estratégicas, contextualizando metas de melhoria. Por fim, ao assimilar insights de mineração de dados adotados em pesquisas como (Soares et al., 2023), que mapeiam determinantes de qualidade educacional, a incorporação de variáveis estruturais e socioeconômicas confere maior profundidade analítica e embasa intervenções locais efetivas.

5.1 Comparação com Abordagens Tradicionais

Como linha de base, contrastamos o modelo híbrido com a segmentação geográfico-administrativa utilizada pela rede: no caso cearense, as Coordenadorias Regionais de Desenvolvimento da Educação (CREDE) e as Superintendências das Escolas Estaduais de Fortaleza (SEFOR) — denominadas DRE em outras redes estaduais. Essa partição agrupa escolas por contiguidade territorial definida pela gestão, sem otimização estatística de coesão intragrupo ou separação entre grupos. Na prática, tende a reunir realidades pedagógicas e operacionais heterogêneas (por exemplo, grandes escolas urbanas e unidades de pequeno porte no interior), o que reduz a homogeneidade interna dos grupos e dificulta comparações justas de eficiência. Tomamos essa organização territorial como *baseline* para avaliar os ganhos do arranjo proposto, que forma contextos por similaridade efetiva de perfil escolar.

Em comparação ao agrupamento geográfico, a combinação PCA+k-means++ revelou padrões latentes com ganho expressivo de qualidade: a silhueta média passou de 0,34 ($\pm$0,12) para 0,67 ($\pm$0,08), o que corresponde a um incremento relativo de 97% na qualidade da segmentação. A Figura 6 sintetiza esse contraste por meio de barras com incerteza (desvio-padrão), evidenciando maior coesão intragrupo e separação entre grupos na abordagem proposta.

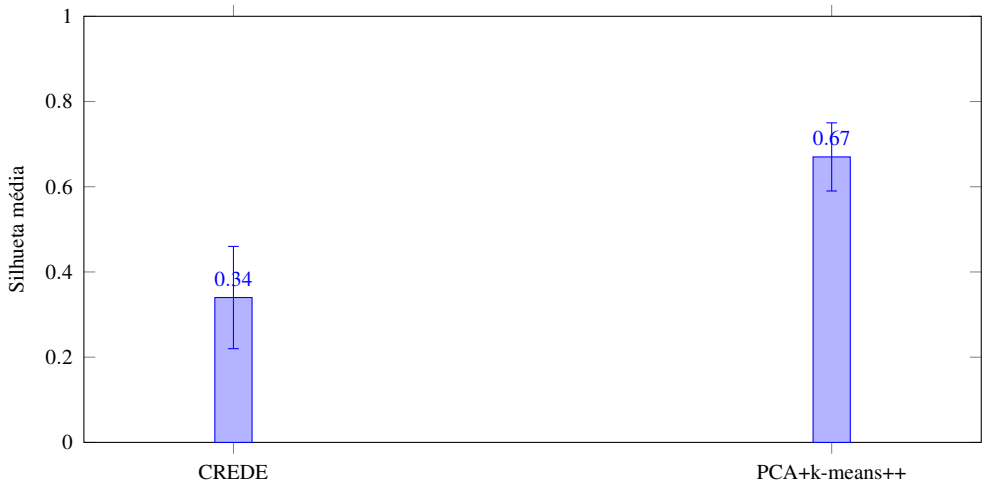

Figura 6: Qualidade da segmentação por abordagem (coeficiente de silhueta médio $\pm$ desvio-padrão). Fonte: autoria própria.

No *benchmarking* de eficiência, avaliamos duas estratégias: (i) *DEA global*, em que todas as escolas são comparadas em uma única fronteira; e (ii) *DEA intra-cluster*, em que a DEA é executada separadamente em cada contexto homogêneo obtido por PCA+k-means++. A segunda estratégia é mais informativa porque compara cada escola apenas com pares similares, reduzindo o viés de heterogeneidade. Na configuração global, cerca de 15% das unidades caem diretamente na fronteira de eficiência, mas as demais frequentemente carecem de referências operacionais próximas. Já na análise intra-cluster, 78% das escolas aparecem com *margem mensurável de melhoria*, isto é, recebem um conjunto de referência (*peers*) e metas quantitativas (alvos de outputs e folgas) factíveis no próprio contexto, o que viabiliza trajetórias graduais de aprimoramento. Além disso, a acurácia da segmentação melhora substancialmente: o Índice de Rand Ajustado apresenta ganho de 23% em relação às abordagens convencionais, indicando que os grupos formados

refletem com mais fidelidade a estrutura subjacente dos dados e, por consequência, produzem recomendações de política mais contextualizadas e comparações mais justas de eficiência.

De forma integrada, os resultados indicam que a combinação entre PCA, clusterização e DEA transforma a análise de eficiência em um instrumento operacional, no qual cada escola dispõe de parâmetros realistas de comparação e trajetórias progressivas de aprimoramento, definidas a partir de contextos homogêneos e referências empiricamente observadas.

6 Conclusão

Este estudo demonstrou que a integração entre redução de dimensionalidade, clusterização e mensuração de eficiência por DEA aplicada intra-cluster produz um quadro analítico mais justo e informativo para a rede pública de ensino médio do Ceará. Ao substituir comparações globais por comparações "entre semelhantes", o *pipeline* proposto segmenta escolas em contextos homogêneos e, a partir daí, estabelece referências factíveis para o avanço de cada unidade.

Os resultados empíricos reforçam o ganho substantivo dessa abordagem frente a alternativas tradicionais. Observou-se incremento relativo de aproximadamente 97% na qualidade da segmentação (coeficiente de silhueta: 0,34 $\pm$ 0,12 vs. 0,67 $\pm$ 0,08), além de 23% de melhoria na acurácia de agrupamento (Índice de Rand Ajustado). No plano do *benchmarking*, a análise intra-cluster identificou margem mensurável de melhoria em 78% das escolas, ao passo que a aplicação global da DEA destacou majoritariamente unidades já na fronteira (cerca de 15%). Esses números sustentam recomendações mais contextualizadas e trajetórias de evolução graduais e exequíveis.

Esses valores descrevem diferenças metodológicas observadas neste estudo e não constituem garantias de desempenho universal.

Ressalta-se que a silhueta média reportada na comparação entre abordagens é calculada no espaço completo das componentes principais selecionadas, enquanto as visualizações bidimensionais têm finalidade exclusivamente exploratória.

Sob a perspectiva da gestão, não basta apenas identificar as melhores escolas: é fundamental propor soluções práticas para aqueles que ainda não atingiram esse padrão. A organização em camadas de isoeficiência operacionaliza essa lógica, permitindo que escolas não eficientes se aproximem de pares de referência do próprio contexto por meio de metas realistas de outputs e folgas operacionais, favorecendo ciclos de melhoria contínua, mentoria entre pares e alocação estratégica de recursos.

Para tornar a aplicação ainda mais prática, sugere-se que a Secretaria de Educação do Ceará (SEDUC) lance editais específicos destinados a escolas dos clusters de menor eficiência, vinculando apoio financeiro e programas de formação continuada a metas de desempenho definidas no próprio cluster. As Diretorias Regionais de Educação (DREs) podem estruturar oficinas regionais de capacitação para gestores escolares, voltadas ao uso de dashboards gerados pelo *pipeline* híbrido, de forma a orientar intervenções pedagógicas com base em evidências contextuais. Adicionalmente, as próprias escolas podem adotar o modelo para elaborar planos de ação anual, com reuniões trimestrais de acompanhamento entre pares de contexto similar, promovendo trocas de boas práticas e ajustes estratégicos em tempo real.

O trabalho também evidencia o papel crescente de IA e ciência de dados na formulação, monitoramento e avaliação de políticas públicas educacionais. Ao tornar visíveis padrões latentes e gargalos específicos de cada contexto, a segmentação inteligente contribui para reduzir assimetrias regionais, valorizar boas práticas e fortalecer redes colaborativas orientadas por evidências. A robustez metodológica foi preservada por meio de documentação, versionamento de código e escolhas técnicas justificadas, garantindo reprodutibilidade.

Como agenda de pesquisa, recomenda-se ampliar variáveis contextuais (infraestrutura, trajetórias escolares), incorporar dados longitudinais e testar o *pipeline* em outros níveis de ensino e regiões para avaliar generalização e impacto sistêmico. A exploração de modelos de aprendizado profundo para detecção de padrões temporais e integração com indicadores de fluxo escolar despontam como vias promissoras.

Cabe destacar que os resultados apresentados refletem escolhas metodológicas específicas e características do conjunto de dados analisado, não constituindo garantias de desempenho universal do *pipeline* proposto. As diferenças observadas entre abordagens devem ser compreendidas como evidências empíricas contextualizadas, e não como afirmações generalizáveis a outros sistemas educacionais sem a devida validação adicional.

Em síntese, a arquitetura proposta recoloca a análise de dados como instrumento efetivo de transformação: transparente, replicável e orientado à equidade. Ao comparar escolas em condições genuinamente semelhantes, traduzir essa comparação em metas factíveis e articular intervenções políticas e formativas baseadas em clusters, o estudo oferece insumos práticos para decisões que promovam de fato o direito à aprendizagem com justiça e melhoria contínua.

Disponibilidade de Código

O código utilizado para o pré-processamento dos dados, aplicação da Análise de Componentes Principais (PCA), agrupamento por k-means e cálculo da eficiência via Data Envelopment Analysis (DEA) encontra-se disponível publicamente no seguinte repositório:

https://github.com/franciscojcardoso/codigo_saeb_artigo

Agradecimentos

Os autores agradecem ao Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) pela disponibilização dos dados do SAEB, e à Universidade Estadual do Ceará pelo apoio institucional à pesquisa.

Referências

Arthur, D., & Vassilvitskii, S. (2007). K-means++: The Advantages of Careful Seeding. *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 1027–1035. https://theory.stanford.edu/~sergei/papers/kMeansPP-soda.pdf [GS Search].

Avelar, C. P., Penna, P. H., & Freitas, H. C. (2014). Algoritmo K-means para mapeamento estático de processos em Redes-em-Chip. *Anais do XV Simpósio em Sistemas Computacionais de Alto Desempenho (WSCAD 2014)*, 1–12. https://doi.org/10.5753/wscad.2014.15012 [GS Search].

Banker, R. D., Charnes, A., & Cooper, W. W. (1984). Some models for estimation technical and scale inefficiencies in Data Envelopment Analysis. *Management Science*, *30*, 1078–1092. https://doi.org/10.1287/mnsc.30.9.1078 [GS Search].

Bartolacci, F., Gobbo, R. D., & Soverchia, M. (2024). Improving public services' performance measurement systems: applying data envelopment analysis in the big and open data context. *International Journal of Public Sector Management*, *38*(3), 313–331. https://doi.org/10.1108/IJPSM-06-2023-0186 [GS Search].

Battisti, F. M., & Carvalho, T. B. A. (2022). Threshold Feature Selection PCA. *Anais do X Symposium on Knowledge Discovery, Mining and Learning*, 50–57. https://doi.org/10.5753/kdmile.2022.227718 [GS Search].

Borba, B. F. (2019). *Proposta de um sistema de informação gerencial para análise de dados baseado no modelo K-MEANS e MCLP sobre a localização de instalações policiais* [tese de dout., Universidade Federal de Pernambuco]. https://repositorio.ufpe.br/handle/123456789/36135

Brasil. (1996). Lei de Diretrizes e Bases da Educação Nacional, Lei nº 9.394, de 20 de dezembro de 1996. https://www.planalto.gov.br/ccivil_03/leis/l9394.htm

Brasil. (2018). Lei nº 13.709, de 14 de agosto de 2018. https://www.in.gov.br/web/dou/-/lei-n-13-709-de-14-de-agosto-de-2018-36889940

Brasil. (2023). *Saeb 2021: Indicador de Nível Socioeconômico do Saeb 2021 - Nota Técnica*. Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep). Brasília, DF. https://www.gov.br/inep/pt-br/centrais-de-conteudo/acervo-linha-editorial/publicacoes-institucionais/avaliacoes-e-exames-da-educacao-basica/saeb-2021-indicador-de-nivel-socioeconomico-do-saeb-2021-nota-tecnica [GS Search].

Caliński, T., & Harabasz, J. (1974). A dendrite method for cluster analysis. *Communications in Statistics-theory and Methods*, *3*, 1–27. https://doi.org/10.1080/03610927408827101 [GS Search].

Campos, M. M., & Vieira, L. F. (2021). COVID-19 and early childhood in Brazil: Impacts on children's well-being, education and care. *European Early Childhood Education Research Journal*, *29*, 125–140. https://doi.org/10.1080/1350293X.2021.1872671 [GS Search].

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research*, *1*, 429–444. https://doi.org/10.1016/0377-2217(78)90138-8 [GS Search].

Cook, W. D., & Zhu, J. (2014). *Data Envelopment Analysis: A Handbook on the Modeling of Internal Structures and Networks*. Springer. https://doi.org/10.1007/978-1-4899-8068-7 [GS Search].

Dahouda, M. K., & Joe, I. (2021). A deep-learned embedding technique for categorical features encoding. *IEEE Access*, *9*, 114381–114391. https://doi.org/10.1109/ACCESS.2021.3104357 [GS Search].

Davies, D. L., & Bouldin, D. W. (2009). A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2), 224–227. https://doi.org/10.1109/TPAMI.1979.4766909 [GS Search].

Ding, C., & He, X. (2004). K-means clustering via principal component analysis. *Proceedings of the 21st International Conference on Machine Learning (ICML)*. https://doi.org/10.1145/1015330.1015408 [GS Search].

Ersoy, Y. (2021). Performance Evaluation in Distance Education by Using Data Envelopment Analysis (DEA) and TOPSIS Methods. *Arabian Journal for Science and Engineering*, *46*, 1803–1817. https://doi.org/10.1007/s13369-020-05087-0 [GS Search].

Ester, M., Kriegel, H.-P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96)*, 226–231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf [GS Search].

Gharakhani, M., Kazemi, I., & Haji, H. (2011). A robust DEA model for measuring the relative efficiency of Iranian high schools. *Management Science Letters*, *1*, 389–404. https://doi.org/10.5267/j.msl.2011.01.002 [GS Search].

Giacomello, C. P., & Oliveira, R. L. D. (2014). Análise Envoltória de Dados (DEA): uma proposta para avaliação de desempenho de unidades acadêmicas de uma universidade. *Revista Gestão Universitária na América Latina-GUAL*, *7*, 130–151. https://doi.org/10.5007/1983-4535.2014v7n2p130 [GS Search].

Guan, C., Mou, J., & Jiang, Z. (2020). Artificial intelligence innovation in education: A twenty-year data-driven historical analysis. *International Journal of Innovation Studies*, *4*, 134–147. https://doi.org/10.1016/j.ijis.2020.09.001 [GS Search].

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of intelligent information systems*, *17*(2), 107–145. https://doi.org/10.1023/A:1012801612483 [GS Search].

Hancock, J. T., & Khoshgoftaar, T. M. (2020). Survey on categorical data for neural networks. *Journal of Big Data*, *7*(1), 28. [GS Search].

Hongyu, K., Sandanielo, V. L. M., & Oliveira Junior, G. J. (2016). Análise de componentes principais: resumo teórico, aplicação e interpretação. *E&S Engineering and Science*, *5*, 83–90. https://www.researchgate.net/publication/320646723_Analise_de_Componentes_Principais_Resumo_Teorico_Aplicacao_e_Interpretacao [GS Search].

Iodice D'Enza, A., Markos, A., & Palumbo, F. (2022). Chunk-wise regularised PCA-based imputation of missing data. *Statistical Methods & Applications*, *31*(2), 365–386. [GS Search].

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R* (Vol. 103). Springer. https://doi.org/10.1007/978-1-4614-7138-7 [GS Search].

Jolliffe, I. T. (2002). *Principal Component Analysis* (2ª ed.). Springer. [GS Search].

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *374*(2065), 20150202. https://doi.org/10.1098/rsta.2015.0202 [GS Search].

Khatun, M. R., Mim, M. A., Tasin, M. M., & Hossain, M. M. (2025). A hybrid framework of statistical, machine learning, and explainable AI methods for school dropout prediction. *Plos one*, *20*(9), e0331917. [GS Search].

Knox, J., Williamson, B., & Bayne, S. (2020). Machine behaviourism: future visions of 'learnification' and 'datafication' across humans and digital technologies. *Learning, Media and Technology*, *45*(1), 31–45. https://doi.org/10.1080/17439884.2019.1623251 [GS Search].

Leporace, C. (2023). Machine Learning e a Aprendizagem Humana – Uma Análise a Partir do Enativismo. https://www.maxwell.vrac.puc-rio.br/est_conteudo.php?nrSeq=61821@1 [GS Search].

Mariano, E. B., Almeida, M. R., & Rebelatto, D. A. N. (2006). Princípios Básicos para uma proposta de ensino sobre análise por envoltória de dados. *Anais do XXXIV Congresso Brasileiro de Ensino de Engenharia (COBENGE 2006)*. https://www.abenge.org.br/cobenge/legado/arquivos/13/artigos/14_285_716.pdf [GS Search].

Miranda, A. C., & Miranda, E. C. M. (2018). Alternative methodology in the elaboration of indicators to evaluate schools. *Pro-Posições*, *29*(3), 207. https://doi.org/10.1590/1980-6248-2016-0051 [GS Search].

Mohamad, I. B., & Usman, D. (2013). Standardization and Its Effects on K-Means Clustering Algorithm. *Research Journal of Applied Sciences, Engineering and Technology*, *6*(17), 3299–3303. https://doi.org/10.19026/rjaset.6.3638 [GS Search].

Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*, 86–97. https://doi.org/10.1002/widm.53 [GS Search].

Nogueira, J. M. M., Oliveira, K. M. M., de Vasconcelos, A. P., & Oliveira, L. G. L. (2012). Estudo exploratório da eficiência dos Tribunais de Justiça estaduais brasileiros usando a Análise Envoltória de Dados (DEA). *Revista de Administração Pública*, *46*, 1317–1340. https://doi.org/10.1590/S0034-76122012000500007 [GS Search].

Onusic, L. M., Nova, S. P. C. C., & Almeida, F. C. (2007). Modelos de previsão de insolvência utilizando a análise por envoltória de dados: aplicação a empresas brasileiras. *Revista de Administração Contemporânea*, *11*, 77–97. https://doi.org/10.1590/S1415-65552007000500006 [GS Search].

Park, W., & Kwon, H. (2024). Implementing artificial intelligence education for middle school technology education in Republic of Korea. *International Journal of Technology and Design Education*, *34*(1), 109–135. https://doi.org/10.1007/s10798-023-09812-2 [GS Search].

Pereira, V. R. F., Paula, A. D., & Araújo, C. O. (2020). Método de agrupamento aplicado à avaliação escolar: um estudo de caso para avaliações de larga escala. *EDUCA – Revista Multidisciplinar em Educação*, *7*(17), 901–919. https://doi.org/10.26568/2359-2087.2020.4413 [GS Search].

Périco, A. E., Rebelatto, D. A. N., & Santana, N. B. (2008). Eficiência bancária: os maiores bancos são os mais eficientes? Uma análise por envoltória de dados. *Gestão & Produção*, *15*(2), 421–431. https://doi.org/10.1590/S0104-530X2008000200016 [GS Search].

Pessano, N. B., & Halmenschlager, C. (2005). Aplicacção de Data Mining em Data Warehouse: Desenvolvimento da Ferramenta ToolMiner. *Simpósio Brasileiro de Sistemas de Informação (SBSI)*, 151–158. https://doi.org/10.5753/sbsi.2005.14979 [GS Search].

Pimenta, I. A., Silva, D. A., Moura, E. S., Silveira, M. M., & Gomes, R. L. (2024). Impact of Data Anonymization in Machine Learning Models. *13th Latin-American Symposium on Dependable and Secure Computing (LADC 2024)*, 188–191. https://doi.org/10.1145/3697090.3699865 [GS Search].

Rassouli-Currier, S. (2007). Assessing the efficiency of Oklahoma public schools: a data envelopment analysis. *Southwestern Economic Review*, *34*, 131–144. https://swer.wtamu.edu/sites/default/files/Data/131-144-59-218-1-PB.pdf [GS Search].

Roll, I., & Wylie, R. (2016). Evolution and revolution in artificial intelligence in education. *International Journal of Artificial Intelligence in Education*, *26*, 582–599. https://doi.org/10.1007/s40593-016-0110-3 [GS Search].

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. https://doi.org/10.1016/0377-0427(87)90125-7 [GS Search].

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised k-means clustering algorithm. *IEEE Access*, *8*, 80716–80727. https://doi.org/10.1109/ACCESS.2020.2988796 [GS Search].

Soares, J. L., Costa, T. B., Moura, L. S., Sousa, W. S., Mesquita, A. L., & Mesquita, D. S. (2023). Machine Learning Based Fault Detection on Belt Conveyor Idlers. *Proceedings of the DINAME*. https://doi.org/10.5753/rbie.2023.2831 [GS Search].

Soares, T. S. S. (2022). O Sistema de Avaliação da Educação Básica (SAEB) em tempos de pandemia: ensino de Matemática e as Tecnologias Digitais. *Com a Palavra, o Professor*, *7*, 95–106. https://periodicos.uece.br/index.php/comapalavraoprofessor/article/view/8741 [GS Search].

Varella, C. A. A. (2008). *Análise de Componentes Principais*. Universidade Federal Rural do Rio de Janeiro. https://www.academia.edu/download/52069850/analise_de_componentes_principais.pdf [GS Search].

Vilaça, W. S. (2023). *Análise de Sistemas Educacionais Aplicando Técnicas de Agrupamento e Análise por Envoltória de Dados* [tese de dout., Universidade Estadual do Ceará]. https://siduece.uece.br/siduece/trabalhoAcademicoPublico.jsf?id=112255

Zhu, J. (2022). DEA under big data: data enabled analytics and network data envelopment analysis. *Annals of Operations Research*, *309*(2), 761–783. https://doi.org/10.1007/s10479-020-03668-8 [GS Search].