Desigualdades Educacionais no Brasil: Uma Análise por Clusterização de Indicadores Educacionais e de Desempenho Escolar

Matheus L. de Melo Silva[1], Lívia Almada Cruz[1], Regis Pires Magalhães[1], Tatieures Gomes Pires[2], José Antonio Macedo[3], Rossana Maria de Castro Andrade[3]

[1]Campus Quixadá - Universidade Federal do Ceará - CE – Brasil

[2]Centro de Referência em Inteligência Artificial (CRIA)
Universidade Federal do Ceará - CE – Brasil

[3]Departamento de Computação - Universidade Federal do Ceará - CE – Brasil

matheusleandro@alu.ufc.br,
{livia.almada,regismagalhaes,tatieures}@ufc.br,
{jose.macedo,rossana}@dc.ufc.br

Abstract. *The Brazilian education system faces structural and socioeconomic challenges, reflected in unequal access to education and low academic performance rates, especially in vulnerable regions. Analyzing educational indicators helps identify structural changes in education, assess the effectiveness of implemented policies, and monitor the evolution of educational quality. This work employs the clustering of educational and school performance indicators to identify factors for educational inequalities in Brazil. Based on data from several educational indicators from 2015, 2019, and 2021 provided by INEP, it was possible to identify municipalities with more similar profiles. In addition, the temporal analysis of the clusters allowed us to understand the evolution of inequalities over the years, providing information that can be useful for formulating more effective public policies and strategically allocating resources.*

Resumo. *O sistema educacional brasileiro enfrenta desafios estruturais e socioeconômicos, refletidos no acesso desigual à educação e nos baixos índices de desempenho acadêmico, especialmente em regiões vulneráveis. A análise de indicadores educacionais auxilia na identificação de mudanças estruturais no ensino, avaliação da efetividade de políticas implementadas e monitoramento da evolução da qualidade educacional. Este trabalho visa identificar fatores relacionados às desigualdades educacionais no Brasil e compreender a evolução das desigualdades ao longo dos anos, oferecendo informações úteis para a formulação de políticas públicas mais eficazes e a alocação estratégica de recursos. Utilizou-se clusterização de indicadores educacionais e de desempenho escolar, a partir de dados de diversos indicadores educacionais dos anos de 2015, 2019 e 2021 fornecidos pelo INEP. Foi possível identificar grupos de municípios com perfis mais semelhantes e indicadores que melhor discriminam tais grupos. Além disso, uma análise de evolução de clusters permitiu uma avaliação temporal da qualidade do ensino.*

1. Introdução

A análise de dados, impulsionada pelo crescente volume de informações disponíveis [Dhar 2013, Janiesch et al. 2021, Zhang and Oles 2000], tornou-se essencial para com-

preender e solucionar questões complexas em diversas áreas nos últimos anos. Essa abordagem permite identificar padrões e correlações que auxiliam na tomada de decisões, especialmente no âmbito das políticas públicas, onde a complexidade dos desafios sociais e econômicos exige estratégias mais eficazes [Jain 2010, Shinde and Shah 2018]. Entre as técnicas de análise de dados, a clusterização se destaca por agrupar dados semelhantes e revelar inter-relações, oferecendo percepções sobre diferentes fenômenos [Xu and Wunsch 2008, Kriegel et al. 2011].

No Brasil, o sistema educacional enfrenta desafios estruturais e socioeconômicos, refletidos no acesso desigual à educação e nos baixos índices de desempenho acadêmico, especialmente em regiões vulneráveis. O Programa Internacional de Avaliação de Estudantes aponta que menos da metade dos estudantes brasileiros de 15 anos atinge o aprendizado mínimo em matemática e ciências [CNN 2023]. Estudos anteriores [Cutler and Lleras-Muney 2012] reforçam como as desigualdades regionais impactam os resultados escolares, tornando essencial compreender a evolução dos padrões educacionais para embasar políticas públicas mais eficazes.

A clusterização tem sido amplamente aplicada no contexto educacional para analisar desempenho estudantil, identificar fatores de evasão e orientar políticas públicas [Nikita Sachdeva 2023]. Como a clusterização permite segmentar dados não rotulados [Tan et al. 2016], sua aplicação possibilita identificar padrões e tendências que podem fundamentar intervenções mais eficazes [MacQueen et al. 1967][Mohamed Nafuri et al. 2022]. Além disso, a análise de clusters pode auxiliar na identificação de mudanças estruturais no ensino, avaliação da efetividade de políticas implementadas e monitoramento da evolução da qualidade educacional. A constante transformação dos dados torna a clusterização uma abordagem promissora para a análise educacional, pois dispensa categorização prévia [Li et al. 2021].

Este trabalho analisa as disparidades e a evolução da qualidade do ensino nas escolas públicas brasileiras por meio da clusterização de indicadores educacionais. Esses indicadores fornecem uma visão abrangente da educação em diferentes regiões, permitindo identificar desigualdades, avanços e áreas que necessitam de melhorias [Gonçalves et al. 2017]. Neste trabalho, usamos os indicadores educacionais do INEP das escolas públicas do ensino fundamental. Os indicadores são agrupados a nível municipal para análise de clusterização. Após a clusterização, as variáveis que melhor discriminam os clusters são identificadas usando aprendizado de máquina supervisionado e análise de importância de características. Propomos também uma metodologia para atribuição da qualidade educacional dos clusters com base nos indicadores. Por fim, analisamos a evolução dos clusters no período analisado, onde é possível identificar padrões de evolução de clusters, tais como migração, divisão e nascimento de novos clusters. Utilizando técnicas de clusterização, pode-se compreender disparidades estruturais e necessidades específicas de grupos de municípios, contribuindo para uma alocação mais eficiente e equitativa de recursos.

O restante do artigo está organizado da seguinte forma: a Seção 2 discute os trabalhos relacionados. A Seção 3 apresenta o conjunto de dados e a metodologia utilizada. A Seção 4 discute os resultados obtidos. Por fim, a Seção 5 traz as considerações finais.

2. Trabalhos relacionados

Esta seção apresenta pesquisas que aplicam técnicas de aprendizado não supervisionado para compreensão de fenômenos no contexto educacional ou governamental, destacando abordagens e metodologias comparáveis.

[Mohamed Nafuri et al. 2022] emprega aprendizado não supervisionado para auxiliar o governo da Malásia na redução da evasão universitária, identificando padrões nos dados educacionais. O algoritmo *k-means* foi utilizado para clusterização e o coeficiente de silhueta [Rousseeuw 1987] para validação dos clusters. De maneira similar, [Valles-Coral et al. 2022] propõe um modelo preditivo para evasão universitária com o algoritmo *HDBSCAN*, validado pelo índice *Calinski-Harabasz*. Os dados analisados foram coletados de chatbots interagindo com estudantes, permitindo uma compreensão mais detalhada do comportamento acadêmico. Diferente desses estudos focados na evasão no ensino superior, [Fernández et al. 2023] desenvolve um framework para identificar escolas que necessitam de investimentos prioritários. Utilizando um algoritmo personalizado, dados estruturais e financeiros das instituições são analisados avaliando critérios como funcionalidade e segurança. [Quintero et al. 2022] investiga o impacto econômico da COVID-19 por meio de clusterização com o algoritmo K-medoides, acompanhando a evolução dos grupos formados ao longo dos anos.

Este trabalho se diferencia ao aplicar aprendizado de máquina para analisar diferentes perfis de municípios brasileiros a partir de indicadores educacionais e governamentais. Enquanto [Mohamed Nafuri et al. 2022] e [Valles-Coral et al. 2022] exploram a evasão universitária e [Fernández et al. 2023] avaliam a infraestrutura escolar, investigamos a relação entre fatores educacionais, proporcionando uma visão mais abrangente do cenário educacional. Além disso, ao contrário de [Quintero et al. 2022], que foca na evolução de clusters para entendimento do impacto econômico da pandemia da COVID-19, analisamos a evolução de clusters com relação aos indicadores educacionais ao longo do tempo, permitindo avaliar diferentes padrões de evolução.

Na Tabela 1, são resumidos os trabalhos relacionados, destacando-se os aspectos mais relevantes de cada estudo, como os métodos de clusterização utilizados (**Algoritmo**), as métricas de validação empregadas (**Métricas Principais**), a origem dos dados analisados (**Fonte dos Dados**), o domínio de aplicação (**Domínio**) e os objetivos principais de cada trabalho (**Objetivo**). Além disso, a tabela evidencia as diferenças entre este trabalho e os demais, ressaltando as particularidades da abordagem proposta. Diferentemente dos trabalhos relacionados apresentados na Tabela 1, o presente estudo distingue-se por concentrar-se especificamente na análise de indicadores educacionais oficiais definidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Enquanto outras abordagens utilizaram dados de natureza diversa, como interações em chatbots, informações de estudantes universitários ou indicadores regionais, este trabalho emprega dados educacionais estruturados e padronizados, amplamente reconhecidos e utilizados na formulação de políticas públicas no Brasil. Além disso, este estudo propõe uma análise longitudinal, avaliando a evolução dos clusters formados ao longo do tempo. Essa abordagem possibilita não apenas a identificação de padrões estáticos, mas também a compreensão das dinâmicas de migração entre clusters, contribuindo para detectar tendências, avaliar o impacto de políticas públicas e subsidiar decisões estratégicas no âmbito educacional.

Tabela 1. Comparativo dos trabalhos relacionados

Artigo	Algoritmo	Métrica	Fonte de Dados	Domínio
[Mohamed Nafuri et al. 2022]	*k-means*	Coeficiente de Silhueta	Estudantes universitários	Universitário
[Valles-Coral et al. 2022]	HDBSCAN	*Calinski–Harabasz*	*Online chatbot* com estudantes universitários	Universitário
[Fernández et al. 2023]	Customizado	*Building quality Index*	Dados estruturais e financeiros de escolas	Escolar e Governamental
[Quintero et al. 2022]	*k-medoids*	*Davies–Boulding*	Indicadores regionais	Saúde e Governamental
Este trabalho	*k-means*	Coeficiente de Silhueta	Indicadores educacionais	Escolar e Governamental

3. Dados e métodos

Esta seção apresenta o conjunto de dados utilizado, o pré-processamento aplicado para garantir a qualidade e a consistência dos dados, a metodologia para aplicação dos algoritmos de clusterização e análise de evolução dos clusters.

3.1. Conjunto de dados

Este estudo utilizou dados de indicadores educacionais disponibilizados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)[1]. Foram selecionados dados referentes aos anos de 2015, 2019 e 2021. Primeiramente, definiu-se como ano de partida o ano de 2015, por coincidir com o fim do Plano de Ações Articuladas (PAR 3), programa do Ministério da Educação (MEC) voltado para a melhoria da qualidade da educação básica. Posteriormente, a partir de 2015, entre os anos dos biênios em que são divulgadas as notas do IDEB foram selecionados o ano de 2019, equivalente ao período anterior à pandemia da COVID-19 e o ano de 2021 – último ano com dados disponíveis no momento do estudo.

Tabela 2. Indicadores educacionais do INEP utilizados

Indicador	Descrição	Monotonicidade
Adequação da Formação Docente (AFD)	Responsável por classificar a formação dos professores em relação a suas disciplinas	Crescente
Esforço Docente (IED)	Mensura o esforço empreendido pelos docentes da educação básica no exercício de sua profissão	Decrescente
Índice de Desenvolvimento da Educação Básica (IDEB)	Usado pelo governo brasileiro para medir a qualidade da educação básica	Crescente
Média de Alunos por Turma (ATU)	Estima a quantidade média de alunos por turma em determinada escola	Decrescente
Média de Horas-aula diária (HAD)	A quantidade média de horas-aula diárias em uma escola	Crescente
Percentual de Docentes com Curso Superior (DSU)	Proporção de professores da educação básica que possuem formação superior completa	Crescente
Taxas de Distorção Idade-série (TDI)	Percentual de alunos com idade acima da recomendada para a série que estão cursando.	Decrescente

[1]Os dados e as respectivas notas técnicas explicativas foram obtidos no Portal de Dados Abertos do INEP: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/indicadores-educacionais/

Os indicadores do INEP são segmentados entre ensino fundamental inicial, ensino fundamental final e ensino médio. A análise concentrou-se nos anos iniciais do ensino fundamental, que compreendem os primeiros cinco anos de escolaridade (do 1º ao 5º ano) e a primeira faixa de ensino avaliada pelos indicadores educacionais após a educação infantil. Além disso, os dados foram coletados dos registros individuais por escola, visto que nem todas as informações estavam disponíveis ao nível municipal. Essa abordagem permitiu agregar os dados das escolas de acordo com seus respectivos municípios, possibilitando a realização de análises e clusterização ao nível municipal. Por fim, foram coletados os indicadores da Tabela 2, em mais de 130 mil escolas, totalizando 5.571 municípios brasileiros.

Os indicadores usados na análise são apresentados na Tabela 2, que informa também que os indicadores possuem diferentes monotonicidades em relação à qualidade da educação. A monotonicidade crescente indica que um aumento do valor do indicador implica em uma melhoria na qualidade e a monotonicidade decrescente indica que um aumento do valor do indicador implica em uma piora na qualidade. A Tabela 3 apresenta medidas de estatística descritiva do IDEB, o qual é o principal indicador utilizado para medir a qualidade da educação nas escolas públicas brasileiras. A média passou de 5,24 em 2015 para 5,59 em 2019, mantendo-se em 5,51 em 2021, enquanto a mediana seguiu um padrão semelhante, partindo de 5,3 em 2015, atingindo 5,7 em 2019 e 5,6 em 2021, refletindo um certo avanço no desempenho educacional. Além disso, a redução do desvio padrão, de 1,13 para 0,98 entre 2019 e 2021, indica menor dispersão dos resultados e maior consistência. Da mesma forma, a queda do coeficiente de variação, de 0,21 para 0,18, reforça a homogeneidade dos dados.

Tabela 3. Estatísticas do índice de desenvolvimento da educação básica

Estatísticas	Dados		
	2015	2019	2021
Média	5,24	5,59	5,51
Mediana	5,3	5,7	5,6
Desvio padrão	1,12	1,13	0,98
Coeficiente de variação	0,21	0,20	0,18
Valor máximo	9,8	9,8	9,9
Valor mínimo	0,8	1,4	0,6

Quanto aos valores extremos, o mínimo variou de 0,8 em 2015 para 1,4 em 2019, mas caiu para 0,6 em 2021, refletindo possivelmente o impacto da pandemia de COVID-19. Em contrapartida, o valor máximo permaneceu estável, oscilando entre 9,8 e 9,9. Essas variações indicam uma tendência geral de melhoria e estabilização do IDEB, com redução da variabilidade dos dados. Além disso, há maior concentração de escolas em torno dos valores centrais.

3.2. Metodologia

Pré-processamento dos dados: Esta etapa destina-se à limpeza e tratamento dos dados, incluindo a identificação e o tratamento de valores ausentes, por meio da remoção de escolas com dados faltantes e inconsistências, como a presença de diferentes tipos de dados em uma mesma coluna. Para cada indicador, o INEP disponibiliza um arquivo específico por ano. Inicialmente, realizou-se um processo de integração dos dados de

todos os indicadores educacionais para cada ano. Após isso, procedeu-se à limpeza e tratamento dos dados.

Tabela 4. Informações sobre o conjunto de dados de indicadores de cada ano

Descrição	Valores		
	2015	**2019**	**2021**
Total de escolas	142.034	134.714	132.963
Total de municípios	5571	5571	5571
Número de colunas	80	88	88

A Tabela 4 apresenta as estatísticas dos conjuntos de dados para os anos de 2015, 2019 e 2021, após o tratamento de valores faltantes e a integração dos indicadores. Nela, observa-se que o número de colunas aumentou de 80 para 88 entre 2015 e 2019, mantendo-se constante em 2021, refletindo variações na composição dos indicadores para os anos investigados. O número de linhas, por sua vez, diminuiu ao longo dos anos, passando de 142.033 em 2015 para 132.962 em 2021. A porcentagem de células em branco também apresentou uma redução, de 40,2% em 2015 para 37,1% em 2021, indicando uma melhoria na completude dos dados.

Transformação dos dados para clusterização: Inicialmente, os dados, que estavam organizados ao nível de escola, foram reagrupados para o nível municipal. Para isso, aplicou-se uma média ponderada para a obtenção dos valores de indicadores por município, utilizando como peso o número total de matrículas dos anos iniciais, em cada escola. Assim, assegurou-se que escolas com maior número de alunos tivessem uma representatividade proporcionalmente adequada. Em seguida, procedeu-se à normalização dos dados utilizando a técnica de normalização Min-Max, para garantir que todas as variáveis possuíssem a mesma escala. Por fim, foi realizada a inversão da monotonicidade dos indicadores que originalmente apresentavam monotonicidade decrescente, conforme Tabela 2, para garantir uma interpretação coerente dos dados.

Clusterização: A clusterização foi aplicada individualmente para cada ano da análise. As variáveis que compõem os indicadores foram usadas para agrupar municípios com características educacionais semelhantes. Para a clusterização, foi utilizado o algoritmo *k-means*. Para determinar o número ideal de clusters (k), foram empregados o método do cotovelo [Thorndike 1953], utilizando a métrica de distorção, e o coeficiente de silhueta [Rousseeuw 1987]. No método do cotovelo, o algoritmo *k-means* foi executado com valores de k variando de 2 a 9, calculando-se a distorção para cada caso. O coeficiente de silhueta, por sua vez, avaliou a qualidade da separação entre os grupos, considerando o valor de k que maximizou essa métrica como indicador adicional do número ideal de clusters.

Análise dos perfis dos *clusters*: Para identificar as variáveis mais relevantes para a discriminação dos clusters obtidos nos anos escolhidos, adotou-se uma abordagem baseada em aprendizado de máquina supervisionado, utilizando o framework AutoGluon[2], que automatiza o processo de treinamento, seleção e avaliação de modelos. A tarefa consistiu em classificar o cluster ao qual cada instância pertence, sendo a variável alvo a coluna que identifica o cluster. As demais variáveis foram empregadas como atributos para o treinamento dos modelos. Além disso, o AutoGluon foi configurado para otimizar a métrica

[2]https://auto.gluon.ai/

F1 ponderada, adequada para problemas de classificação multiclasse, especialmente em cenários com possível desbalanceamento entre as classes. A seleção do modelo baseou-se na configuração de *best_quality*, que prioriza o desempenho preditivo através do treinamento de múltiplos algoritmos, seguido de uma avaliação comparativa por meio de um placar. Por fim, a análise de importância das variáveis foi conduzida utilizando o método *feature_importance*[3], que calcula a pontuação da importância de cada variável com base na técnica de importância por permutação. Nesse método, é quantificado o impacto de uma variável no desempenho do modelo ao embaralhar seus valores e medir a queda resultante na métrica de avaliação.

Com o intuito de caracterizar e diferenciar os *clusters*, propôs-se calcular um índice numérico capaz de mensurar o perfil de cada cluster, refletindo sua qualidade de acordo com os indicadores educacionais. Para isso, foi primeiramente calculada a média dos indicadores de cada uma das entidades. Mais precisamente, o índice do município j é dado por $m(j) = \sum_{i=1}^{n} \frac{e_{i,j}}{n}$, onde $e_{i,j}$ é o valor da variável i para o município j, e n o número total de variáveis considerando todos os indicadores. O índice geral do cluster é a média aritmética dos índices dos municípios pertencentes ao cluster. Esse processo pretende avaliar a posição relativa dos clusters de acordo com os níveis de qualidade dados pelos indicadores.

Em seguida, foi construída uma tabela de pontuação para o intervalo de pontuação dos clusters encontrados. Inicialmente, os clusters foram classificados alfabeticamente, atribuindo-se a letra 'A' ao cluster com melhor desempenho e 'F' ao cluster com pior desempenho, considerando o desempenho dos clusters de todos os anos analisados. Em seguida, estabeleceu-se uma escala comparativa calculando a diferença entre a pontuação máxima do melhor cluster no último ano e a pontuação mínima do pior cluster no primeiro ano. Posteriormente, essa amplitude foi aplicada para normalizar os valores de todos os clusters através da subtração progressiva desta diferença das pontuações originais, partindo do cluster de melhor desempenho (A) até o de pior desempenho (F). Finalmente, os valores resultantes foram organizados em uma tabela contendo todos os clusters identificados no período da análise, permitindo uma comparação do desempenho relativo dos clusters ao longo do tempo.

Análise da evolução dos *clusters*: A análise da evolução dos clusters busca entender como os grupos identificados em uma análise de clusterização se transformam com o tempo. Inicialmente, os dados foram segmentados por ano, e a clusterização foi realizada separadamente para cada período. Em seguida, compararam-se os clusters ao longo dos anos, analisando sua composição e a permanência das mesmas entidades nos grupos. Para cada cluster, foi traçada uma trajetória temporal, identificando padrões de evolução, como fusão, fragmentação ou estabilidade. Esses padrões foram analisados para compreender seus fatores determinantes. Assim, foi possível evidenciar tendências nos indicadores educacionais ao longo do tempo, fornecendo informações para a compreensão de fatores que influenciam a educação e a formulação de políticas públicas.

4. Resultados

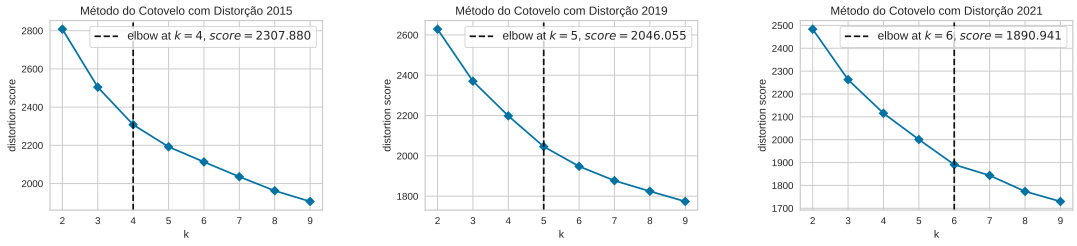
Esta seção apresenta os resultados obtidos durante a clusterização e a análise da migração das entidades.

[3]https://explained.ai/rf-importance/

4.1. Clusterização

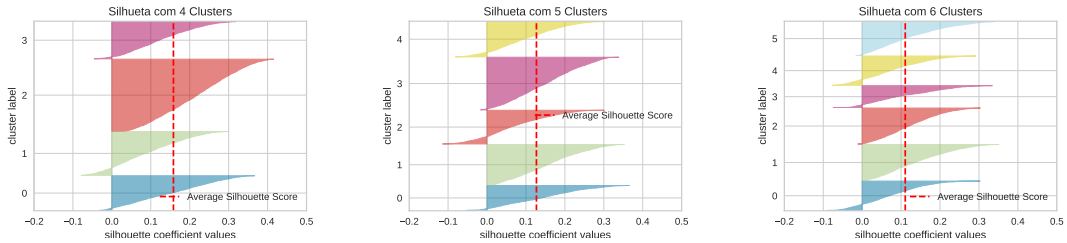
O método do cotovelo (Figura 1) indicou como quantidades ideais de clusters os valores 4, 5 e 6 para os anos de 2015, 2019 e 2021, respectivamente, onde a identificação do ponto ideal do cotovelo foi baseada na técnica Kneedle [Satopaa et al. 2011]. Adicionalmente, foi aplicado o método da silhueta (Figura 2) para os diferentes anos, utilizando valores de k variando de 3 a 7, com o objetivo de validar os resultados obtidos pelo método do cotovelo. Selecionou-se o valor de k que apresentou coeficientes de silhueta mais equilibrados entre os clusters e com menor proporção de valores negativos. Essa análise confirmou os valores ideais de clusters identificados previamente pelo método do cotovelo.

Figura 1. Método do cotovelo aplicado aos dados dos anos 2015, 2019 e 2021.



(a) Clusters 2015 **(b) Clusters 2019** **(c) Clusters 2021**

Figura 2. Método da silhueta aplicado aos dados dos anos 2015, 2019 e 2021.



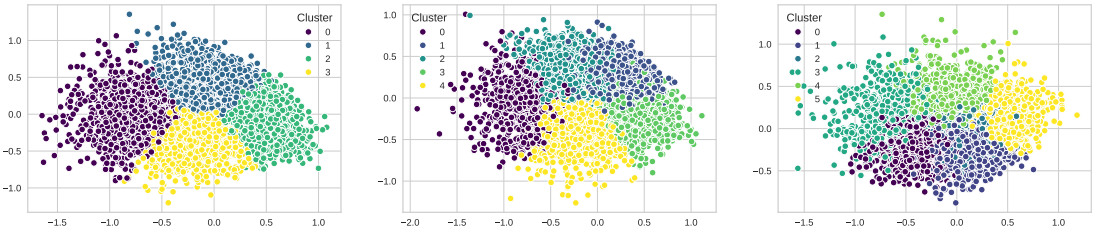
(a) Clusters 2015 **(b) Clusters 2019** **(c) Clusters 2021**

4.2. Análise dos perfis dos clusters

Para cada período selecionado, foi aplicada a clusterização utilizando o algoritmo *k-means*. O método de redução de dimensionalidade *Principal Component Analysis* (PCA) foi usado para visualização gráfica (Figura 3). A análise comparativa dos clusters gerados por PCA em diferentes anos revela dinâmicas temporais significativas na estrutura dos dados. Embora a metodologia mantenha grupos discerníveis com limites relativamente claros, observa-se uma complexidade crescente, onde o número de clusters aumentou de 4 em 2015 para 5 em 2019 e para 6 em 2021.

Também é possível notar um incremento nas sobreposições intercluster, particularmente entre 2019 e 2021, sugerindo maior convergência de características em subgrupos específicos. Paralelamente, é possível observar o surgimento de novos clusters a partir da divisão de clusters preexistentes, como reflexo de mudanças estruturais nos dados ao longo dos anos.

Figura 3. Clusterização com PCA dos diferentes anos



(a) Clusterização 2015 **(b) Clusterização 2019** **(c) Clusterização 2021**

Para identificar as variáveis mais relevantes para a discriminação entre os clusters, foram treinados os modelos classificadores[4] para cada ano analisado. Em cada caso, selecionou-se o modelo com melhor desempenho com base na métrica F1 ponderada. A partir do classificador escolhido, aplicou-se a técnica de análise de importância por permutação, que avalia o impacto de cada variável no desempenho do modelo ao medir a variação da métrica de avaliação quando os valores da variável são embaralhados. Os resultados são apresentados na Tabela 6, a qual reporta as dez variáveis mais relevantes em cada ano, acompanhadas por um número que indica a posição de prioridade atribuída a cada variável no respectivo período. Valores superiores a 10 indicam que a variável não figurou entre as dez primeiras naquele ano específico, mas foi considerada relevante em outros períodos. Como exemplo, observa-se que o indicador IED – Nível 1 ocupou a primeira posição nos anos de 2019 e 2021, enquanto em 2015 figurava apenas na 16ª colocação. Esse comportamento evidencia que determinadas variáveis podem ganhar ou perder relevância conforme alterações nas políticas públicas, nas condições socioeconômicas ou na dinâmica de desempenho dos municípios. Assim, a comparação temporal da importância das características permite não apenas compreender quais fatores estão mais fortemente associados à formação dos clusters em cada período, mas também identificar tendências ou rupturas que podem orientar futuras intervenções no âmbito educacional. Observa-se também que o conjunto de características mais influentes para a determinação dos rótulos dos clusters varia ao longo do tempo, refletindo possíveis transformações no cenário educacional analisado.

O IDEB, indicador criado pelo MEC para medir a qualidade da educação básica, está presente nos três anos analisados, com importâncias atribuídas de 1, 4 e 2, respectivamente. O IED - Nível 1, que representa docentes com até 25 alunos e apenas um turno de trabalho nos anos iniciais do Ensino Fundamental, é o atributo mais importante para os dois últimos anos. Além disso, a Taxa de Distorção Idade-Série (TDI) do 3º ao 5º ano do Ensino Fundamental e a Média de Alunos por Turma (ATU) do 1º ao 2º ano e do 4º ao 5º também se destacam como indicadores que aparecem como relevantes para o perfil dos clusters em diferentes anos.

Além do IDEB, apenas o índice de Adequação da Formação Docente (AFD) - Grupo 1, que representa o percentual de docentes que possuem formação superior na mesma área em que lecionam, e o Percentual de Docentes com Curso Superior (DSU) es-

[4]Modelos: LightGBM, CatBoost, XGBoost, Random Forest, Extra Trees, K-Nearest Neighbors, Regressão Linear, Rede Neural com PyTorch, Rede Neural com FastAI, WeightedEnsemble e StackerEnsembleModel com seus respectivos níveis.

tiveram presentes entre os 10 indicadores mais relevantes em todos os anos considerados no estudo.

Tabela 5. Intervalo das pontuações dos índices de qualidade dos clusters.

Índice	Pontuação	Índice	Pontuação
A	[0.678, 0.716]	D	[0.566, 0.603)
B	[0.641, 0.678)	E	[0.528, 0.566)
C	[0.603, 0.641)	F	[0.491, 0.528)

A distribuição das entidades nos clusters passou por uma reconfiguração significativa entre 2015 e 2021, refletindo tanto mudanças nas características internas dos grupos quanto a migração de entidades entre eles. Entre 2015 e 2019, houve um aumento no número de clusters, indicando uma maior diversidade de cenários. Esse crescimento sugere que as redes de ensino passaram a apresentar configurações mais heterogêneas, demandando estratégias de apoio e planejamento mais específicas e adaptadas às realidades locais. De modo geral, a pontuação atribuída pelos indicadores apresentou uma melhoria gradativa ano a ano (Tabela 7). Em 2015, os índices de qualidade dos clusters foram respectivamente C, D, D e F. Em 2019, apenas um cluster obteve índice F e quatro clusters obtiveram índice C. Em 2021, todos os clusters obtiveram índices superiores a D, sendo um A, um B e três C.

O cluster de melhor desempenho em 2015 apresentou uma redução contínua, diminuindo de 2.132 entidades no Cluster 1 de 2015 para 1.565 no Cluster 1 de 2019 e atingindo 1.014 no Cluster 1 de 2021. Entretanto, é importante destacar que as redes estão em constante transformação e que a redução numérica não indica um retrocesso, mas sim um processo de reconfiguração resultante da elevação dos critérios de desempenho. Cada rede municipal de educação evolui conforme sua realidade local, capacidade de gestão e contexto socioeconômico, sendo natural que cada uma acompanhe a evolução dos indicadores em ritmos distintos. Isso pode justificar a criação de clusters intermediários com uma quantidade reduzida de elementos, como observado no ano de 2021 (Id 2 e 4).

4.3. Análise da evolução dos clusters

A Figura 4 apresenta a evolução dos clusters entre os anos de 2015, 2019 e 2021, mapeando a permanência ou a migração das instâncias de um cluster para outro no período. Entre 2015 e 2019, pode-se observar uma migração de entidades do Cluster 1 de 2015 para o Cluster 4 de 2019. Essas entidades formaram um novo cluster, com pontuação relativamente menor, mas dentro do mesmo índice de qualidade (C). Os demais, que permaneceram no grupo de origem, seguiram uma tendência de melhoria. A pontuação média do cluster aumentou de 0,61 em 2015 para 0,63 em 2019 e alcançou 0,71 em 2021, indicando uma evolução positiva nas características das entidades que o compõem e sugerindo a consolidação de melhorias estruturais. A diminuição da amostra sugere a migração de entidades que não mantiveram o mesmo padrão de desempenho ao longo dos anos para um novo grupo.

O Cluster 2 de 2015 (índice D) migrou majoritariamente para o Cluster 3 de 2019 (índice C), enquanto o Cluster 3 de 2015 (índice D) migrou para o Cluster 2 de 2019 (índice C). Ambos apresentaram uma melhora no nível de qualidade, passando do índice D para o C. No entanto, observa-se uma trajetória distinta entre esses grupos no biênio

Tabela 6. Importância das características por ano.

Característica	Importância		
	2015	2019	2021
IED - Nível 1	16	**1**	**1**
IED - Nível 2	15	19	10
IED - Nível 3	22	7	9
IED - Nível 4	23	9	8
AFD - Grupo 1	4	2	4
AFD - Grupo 5	3	8	6
TDI - 3º ano fundamental	7	17	21
TDI - 4º ano fundamental	5	5	14
TDI - 5º ano fundamental	6	3	11
Valor do IDEB	**1**	4	2
DSU - Ano fundamental inicial	2	6	3
ATU - Ano fundamental inicial	8	16	15
ATU - 1º ano fundamental	9	10	12
ATU - 2º ano fundamental	11	11	7
ATU - 4º ano fundamental	15	13	5
ATU - 5º ano fundamental	10	15	17

Tabela 7. Estatisticas dos clusters

(a) Clusters 2015

Id	Índice	Amostras	Pontuação
1	C	2.132	0,611
2	D	1.304	0,591
3	D	1.105	0,584
4	F	1.029	0,491

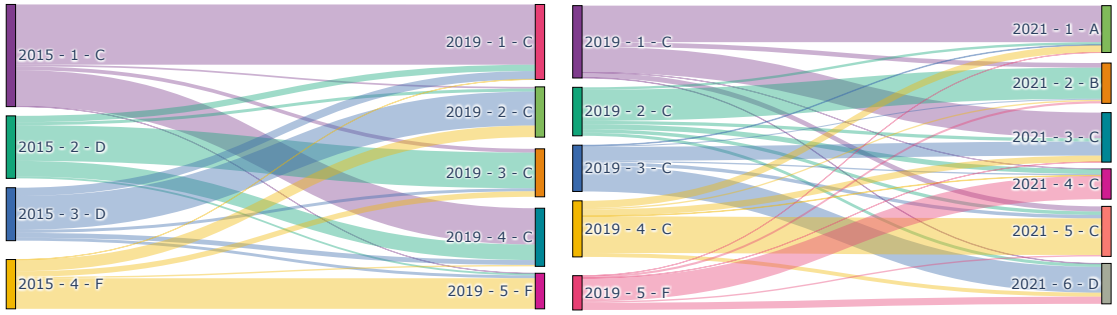
(b) Clusters 2019

Id	Índice	Amostras	Pontuação
1	C	1.565	0,636
2	C	1.049	0,625
3	C	1.001	0,625
4	C	1.213	0,606
5	F	742	0,510

(c) Clusters 2021

Id	Índice	Amostras	Pontuação
1	A	1.014	0,716
2	B	873	0,641
3	C	1.071	0,629
4	C	656	0,624
5	C	1.082	0,619
6	D	874	0,568

Figura 4. Evolução dos clusters.



(a) Migração de 2015-2019 **(b) Migração de 2019-2021**

seguinte: o Cluster 2 de 2019 mantém uma tendência positiva, elevando-se ao nível B em 2021, enquanto o Cluster 3 de 2019 se fragmenta e uma parcela significativa de suas redes de educação regride para o Cluster 6 de 2021, com índice D. Esse movimento de retrocesso levanta questionamentos sobre os fatores que impactaram negativamente esse grupo, sendo necessário um aprofundamento investigativo para compreender as causas dessa reversão de desempenho.

O Cluster 4 de 2015 (índice F) também apresentou uma migração gradual de suas

entidades para outros grupos, passando de 1.029 elementos para 742 em 2019, com somente um leve aumento em seu escore geral. Os elementos que permaneceram foram alocados no Cluster 5 de 2019 (índice F). Esse comportamento pode indicar a existência de barreiras persistentes na implementação de melhorias dentro desse grupo, sugerindo que apenas uma parcela menor das entidades conseguiu lograr êxito em suas políticas educacionais, migrando para clusters com desempenho superior. O que surpreende é que esse mesmo grupo, no biênio subsequente, migra majoritariamente do Cluster 5 de 2019, com índice F, para o Cluster 4 de 2021, com índice C. Esse salto considerável em termos de desempenho pode indicar a adoção de estratégias mais eficazes ou o impacto positivo de políticas públicas implementadas nesse intervalo. Esse caso também requer um estudo adicional no sentido de compreender os fatores que influenciaram as mudanças.

Por fim, vale destacar que, a despeito de algumas entidades terem regredido na análise, todas as pontuações dos grupos apresentaram uma evolução contínua ao longo dos anos. A título de comparação, a menor pontuação do Cluster 6 de 2021, considerado o de pior desempenho naquele ano, supera a menor pontuação registrada no Cluster 4 de 2015, evidenciando um avanço geral nos indicadores de qualidade ao longo do tempo.

5. Conclusões e trabalhos futuros

Este artigo analisou a evolução dos indicadores educacionais nas redes municipais de ensino por meio da clusterização, aplicando o algoritmo *k-means* e técnicas de redução de dimensionalidade. O estudo revelou uma melhoria gradual dos indicadores educacionais no período, embora algumas barreiras persistam. Além disso, foi possível observar uma dinâmica de mobilidade entre clusters, com alguns municípios migrando para grupos de melhor desempenho, enquanto outros apresentaram quedas de desempenho significativas.

Dessa forma, a aplicação de técnicas de clusterização mostrou-se valiosa para a identificação de correlações entre os indicadores educacionais; para o mapeamento de padrões temporais de evolução (melhoria ou piora) dos municípios clusterizados; e para destacar casos críticos que demandam intervenções específicas. Por fim, futuras investigações podem explorar a inclusão de variáveis complementares, como investimentos públicos e fatores demográficos, bem como aprofundar a análise das causas subjacentes à migração entre clusters e suas relações com mudanças socioeconômicas e políticas. Isso é especialmente relevante nos casos em que se observou melhora significativa — como o grupo que migrou do índice F em 2019 para o índice C em 2021 — ou queda acentuada de desempenho, como o cluster que regrediu do índice C em 2019 para o índice D em 2021, ambos demandando investigação mais aprofundada para compreensão dos fatores que impulsionaram essas transformações. Além disso, a validação dos resultados com especialistas e comparações com classificações oficiais também são recomendadas para fortalecer a aplicabilidade prática do estudo. Essa validação ampliará a utilidade prática dos resultados para gestores educacionais, oferecendo subsídios mais robustos para a formulação de políticas públicas setoriais.

Agradecimentos

Este trabalho foi apoiado pelo Fundo Nacional de Desenvolvimento Educacional (FNDE) em parceria com a Universidade Federal do Ceará (UFC) por meio do Projeto “Plataforma Big Data e Inteligência Artificial para Governança dos Programas de Educação Básica do FNDE” [TED 12.222/2023].

Referências

CNN (2023). Brasil tem baixo desempenho e estagna em ranking mundial da educação básica. Dados publicados na CNN Brasil.

Cutler, D. M. and Lleras-Muney, A. (2012). Education and health: insights from international comparisons. *Encyclopedia of Health Economics*.

Dhar, V. (2013). Data science and prediction. *Communications of the ACM*, 56(12):64–73.

Fernández, R., Correal, J. F., D'Ayala, D., and Medaglia, A. L. (2023). A decision-making framework for school infrastructure improvement programs. *Structure and Infrastructure Engineering*, pages 1–20.

Gonçalves, T. G. G. L., do Santo, S. C., and dos Santos, N. G. (2017). Indicadores educacionais brasileiros: limites e perspectivas. *Educação Em Perspectiva*, 8(3):444–461.

Jain, A. K. (2010). Data clustering: 50 years beyond k-means. *Pattern recognition letters*, 31(8):651–666.

Janiesch, C., Zschech, P., and Heinrich, K. (2021). Machine learning and deep learning. *Electronic Markets*, 31(3):685–695.

Kriegel, H.-P., Kröger, P., Sander, J., and Zimek, A. (2011). Density-based clustering. *Wiley interdisciplinary reviews: data mining and knowledge discovery*, 1(3):231–240.

Li, X., Zhang, Y., Cheng, H., Zhou, F., and Yin, B. (2021). An unsupervised ensemble clustering approach for the analysis of student behavioral patterns. *Ieee Access*, 9:7076–7091.

MacQueen, J. et al. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA.

Mohamed Nafuri, A. F., Sani, N. S., Zainudin, N. F. A., Rahman, A. H. A., and Aliff, M. (2022). Clustering analysis for classifying student academic performance in higher education. *Applied Sciences*, 12(19):9467.

Nikita Sachdeva (2023). Top 12 clustering algorithms in machine learning. Dados publicados no daffodil – Os 12 algorimtos mais populares de clusterização.

Quintero, Y., Ardila, D., Aguilar, J., and Cortes, S. (2022). Analysis of the socioeconomic impact due to covid-19 using a deep clustering approach. *Applied Soft Computing*, 129:109606.

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.

Satopaa, V., Albrecht, J., Irwin, D., and Raghavan, B. (2011). Finding a "kneedle"in a haystack: Detecting knee points in system behavior. In *2011 31st International Conference on Distributed Computing Systems Workshops*, pages 166–171.

Shinde, P. P. and Shah, S. (2018). A review of machine learning and deep learning applications. In *2018 Fourth international conference on computing communication control and automation (ICCUBEA)*, pages 1–6. IEEE.

Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining*. Pearson Education India.

Thorndike, R. L. (1953). Who belongs in the family? *Psychometrika*, 18(4):267–276.

Valles-Coral, M. A., Salazar-Ramírez, L., Injante, R., Hernandez-Torres, E. A., Juárez-Díaz, J., Navarro-Cabrera, J. R., Pinedo, L., and Vidaurre-Rojas, P. (2022). Density-based unsupervised learning algorithm to categorize college students into dropout risk levels. *Data*, 7(11):165.

Xu, R. and Wunsch, D. (2008). *Clustering*. John Wiley & Sons.

Zhang, T. and Oles, F. (2000). The value of unlabeled data for classification problems. In *Proceedings of the Seventeenth International Conference on Machine Learning,(Langley, P., ed.)*, volume 20. Citeseer.