



II CIGELUBRA

Congresso Internacional de Gestão e Engenharia Luso-Brasileiro

03 a 05 de dezembro de 2025

As Engenharias na Transformação Digital das PMEs

Limitações do IDEB do Ensino Médio Brasileiro: Uma análise a partir de Técnicas de Mineração de Dados

Giliene da Silva Sousa

Engenharia de Produção – Universidade do Estado do Pará

Luciana Nina Silva Soares

Engenharia de Produção – Universidade do Estado do Pará

Marielly de Sousa Santos

Engenharia de Produção – Universidade do Estado do Pará

Daniel Meireles de Amorim

Engenharia de Produção – Universidade do Estado do Pará

Resumo: A educação básica no Brasil é um pilar fundamental para o desenvolvimento social e econômico, impactando a formação cidadã, a inserção no mercado de trabalho e a redução das desigualdades. Entretanto, desafios como desigualdades regionais e discrepâncias de desempenho entre redes e estados ainda persistem. Este estudo analisou indicadores do Ensino Médio nos estados brasileiros por meio de técnicas de mineração de dados e agrupamento hierárquico, utilizando o *software Orange Data Mining*. Foram consideradas variáveis como notas do Sistema de Avaliação da Educação Básica (SAEB), taxa de distorção idade-série, taxa de aprovação, carga horária diária e número de alunos por turma. Os resultados identificaram quatro grupos de desempenho, sendo o SAEB e a distorção idade-série as variáveis mais relevantes. De modo geral, estados das regiões Sul e Sudeste apresentaram melhores resultados, enquanto muitos do Norte e Nordeste mostraram maior vulnerabilidade, embora algumas exceções tenham sido observadas. A taxa de aprovação apresentou baixo poder explicativo, evidenciando limitações do IDEB quando utilizado isoladamente, pois ele combina apenas taxas de aprovação e exames padronizados, permitindo que escolas obtenham boas notas sem garantir aprendizado efetivo. O estudo ressalta que a utilização de múltiplos indicadores proporciona uma avaliação mais completa da qualidade educacional, subsidiando diagnósticos realistas e políticas públicas voltadas à equidade e melhoria do ensino.

Palavras-chave: Orange, SAEB, Taxa de Aprovação, Desigualdades Regionais.

Limitations of IDEB in Brazilian High School Education: An Analysis Based on Data Mining Techniques

Abstract: Basic education in Brazil is a fundamental pillar for social and economic development, influencing civic formation, labor market integration, and the reduction of inequalities. However, challenges such as regional disparities and performance discrepancies between networks and states still persist. This study analyzed high school indicators across Brazilian states using data mining techniques and hierarchical clustering, employing the *Orange Data Mining* software. Variables considered included SAEB scores, age-grade distortion rates, approval rates, daily class hours, and the number of students per class. The results identified four performance groups, with SAEB scores and age-grade distortion being the most relevant variables. Overall, states in the South and Southeast regions showed better results, while many in the North and Northeast exhibited greater vulnerability, although some exceptions were observed. The approval rate showed low

explanatory power, highlighting the limitations of IDEB when used in isolation, as it combines only approval rates and standardized test results, allowing schools to achieve high scores without ensuring effective learning. The study emphasizes that using multiple indicators provides a more comprehensive assessment of educational quality, supporting realistic diagnostics and public policies aimed at equity and the improvement of teaching.

Keywords: Orange, SAEB, Approval Rate, Regional Inequalities.

1. Introdução

A educação básica no Brasil é considerada um dos principais fatores do desenvolvimento social e econômico, visto que influencia diretamente a formação cidadã, a inserção no mercado de trabalho e a redução das desigualdades sociais. No entanto, apesar dos avanços obtidos nas últimas décadas, o país ainda enfrenta grandes desafios relacionados à qualidade e à igualdade de oportunidades educacionais, tais como desigualdades regionais, infraestrutura escolar insuficiente e disparidades no desempenho entre redes de ensino (Chirinéa; Barreiro, 2018).

Para monitorar e avaliar o desempenho educacional, foram criados diversos indicadores, entre os quais se destaca o Índice de Desenvolvimento da Educação Básica (IDEB), estabelecido pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) em 2007. O IDEB combina duas dimensões principais: o fluxo escolar, medido pela taxa de aprovação, e o desempenho em avaliações, mensurado pelo Sistema de Avaliação da Educação Básica (SAEB), que é uma avaliação nacional aplicada a alunos do ensino fundamental e médio, com o objetivo de aferir competências em Língua Portuguesa e Matemática (Chirinéa, 2017). Embora essa integração forneça um panorama sobre a progressão escolar e o aprendizado médio dos estudantes, sua capacidade de mensurar a qualidade real da educação é limitada. Passos (2024), aponta que a ênfase excessiva nos resultados do IDEB pode gerar distorções, como a adoção de estratégias voltadas apenas para a melhoria dos indicadores, em detrimento do aprendizado efetivo dos alunos.

Nesse sentido, a utilização de técnicas de mineração de dados surge como uma abordagem inovadora para analisar grandes volumes de informações da educação básica de forma ampla e detalhada, permitindo identificar padrões e grupos de similaridade por meio de algoritmos de agrupamento (*clustering*). Como destacam Souza e Kerbauy (2017), a análise de dados quantitativos na educação possibilita mais do que identificar relações entre variáveis, identificando desigualdades estruturais entre diferentes contextos educacionais.

Diante desse contexto, o presente estudo tem como objetivo aplicar técnicas de mineração de dados para analisar informações educacionais do Ensino Médio das redes públicas estaduais brasileiras, considerando dados consolidados do ano de 2023. Para isso, foram utilizados o software Orange e o método de *cluster* hierárquico, de forma a agrupar os estados segundo semelhanças em seus indicadores, identificar aqueles que apresentam maiores vulnerabilidades educacionais e comparar os resultados obtidos com os valores do IDEB. Ademais, a pesquisa sugere uma reflexão crítica sobre as limitações do IDEB como indicador de avaliação educacional. Apesar de ser amplamente reconhecido como um indicador da qualidade da educação básica, o IDEB é basicamente uma combinação de resultados de exames padronizados e taxas de aprovação, o que pode não capturar completamente a complexidade do processo educacional. Dessa forma, a problematização vai além de identificar diferenças regionais no desempenho escolar; ela também questiona se os estados com melhores índices realmente correspondem aos grupos de maior desempenho educacional ou se esses resultados ainda estão ligados a contextos de fragilidade estrutural. Assim, a crítica ao indicador é integrada desde a formulação do problema de pesquisa, prevenindo que seja vista apenas como um resultado imprevisto das descobertas.

Este artigo está estruturado em quatro seções: introdução, que apresenta o contexto e os objetivos do estudo; referencial teórico, que reúne pesquisas e conceitos sobre desempenho educacional e mineração de dados; metodologia, que descreve a base de dados, as variáveis e os procedimentos adotados; e, por fim, resultados e discussões, seguidos da conclusão, que sintetiza os principais achados e indica perspectivas futuras.

2. Referencial Teórico

2.1. Educação Básica

A educação fundamental no Brasil é considerada um direito social essencial, assegurado pela Constituição de 1988 e regulamentado pela Lei de Diretrizes e Bases da Educação Nacional (LDBEN – Lei nº 9.394/1996). Esta fase educacional abrange a educação infantil, o ensino fundamental e o ensino médio, visando garantir o desenvolvimento completo dos indivíduos, promovendo as habilidades necessárias para a cidadania, inserção no mercado de trabalho e a continuidade dos estudos.

Apesar dos progressos em garantir acesso escolar universal, ainda existem desafios estruturais que afetam a qualidade da aprendizagem, a infraestrutura das escolas, a valorização dos educadores e a equidade na oferta de educação (Chirinéa; Barreiro, 2018). O ensino médio, a última etapa da educação básica, enfrenta dificuldades específicas, como elevadas taxas de evasão, falta de motivação dos alunos e currículos que não se relacionam com as realidades dos jovens.

Em nível internacional, destaca-se o PISA (*Programme for International Student Assessment*), que avalia alunos de 15 anos em leitura, matemática e ciências. Os resultados do Brasil nas edições mais recentes demonstram um desempenho inferior à média dos países desenvolvidos, evidenciando a necessidade de mudanças estruturais e investimentos significativos na educação básica.

Nesse cenário, as avaliações educacionais em larga escala tornam-se ferramentas essenciais para diagnosticar o desempenho dos estudantes e orientar políticas públicas. No Brasil, destacam-se o SAEB, a Prova Brasil, o ENEM (Exame Nacional do Ensino Médio), o ENCCEJA (Exame Nacional para Certificação de Competências de Jovens e Adultos) e o IDEB.

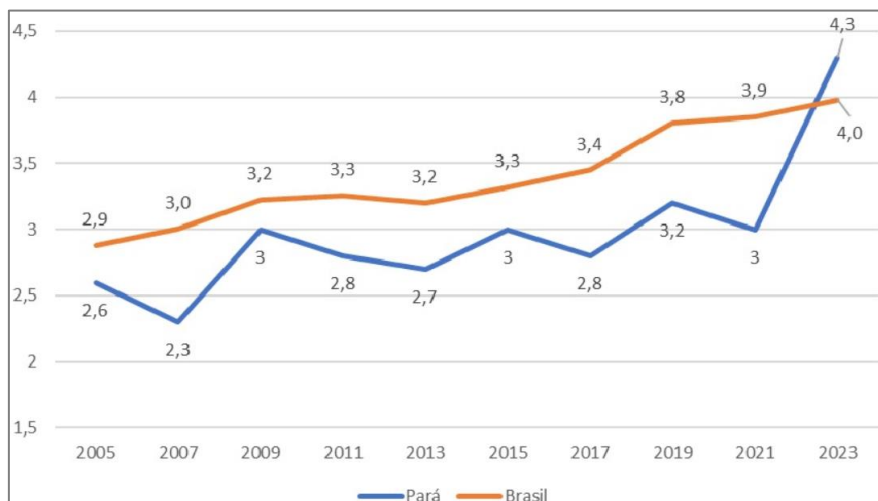
No âmbito nacional, o principal indicador utilizado é o IDEB, criado em 2007 pelo INEP. Este índice é calculado com base na proficiência dos estudantes em Língua Portuguesa e Matemática, conforme medido pelos exames do SAEB, e no fluxo escolar, que leva em consideração as taxas de aprovação, reprovação e evasão. A fórmula oficial usada para essa avaliação é: $\text{IDEB} = \text{Proficiência} \times \text{Fluxo}$.

A proficiência é medida em uma escala de 0 a 10, enquanto o fluxo escolar é representado por uma taxa que varia de 0 a 1. Por exemplo, uma escola que apresenta uma média de proficiência de 6,0 e uma taxa de aprovação de 90% (0,9) terá um IDEB de 5,4 (INEP, 2023).

Apesar de sua relevância como indicador de desempenho educacional, o IDEB apresenta fragilidades que comprometem sua capacidade de refletir com precisão a qualidade do ensino. Conforme exposto por Passos (2024) em reportagem publicada na revista Amazônia Latitude, o caso do estado do Pará ilustra como alterações administrativas no regime de avaliação escolar como a adoção de ciclos de aprendizagem e a redução significativa das reprovações nos primeiros anos do ensino médio podem inflar artificialmente os resultados do índice. Entre 2021 e 2023, o estado subiu para a sexta posição no ensino médio, com aprovação de 99%, a mais alta do país. Contudo, os resultados do ENEM mostram que a aprendizagem permaneceu abaixo da média nacional,

indicando que o avanço aparente oculta desafios estruturais na educação pública. Essa discrepância é evidenciada nas duas figuras a seguir:

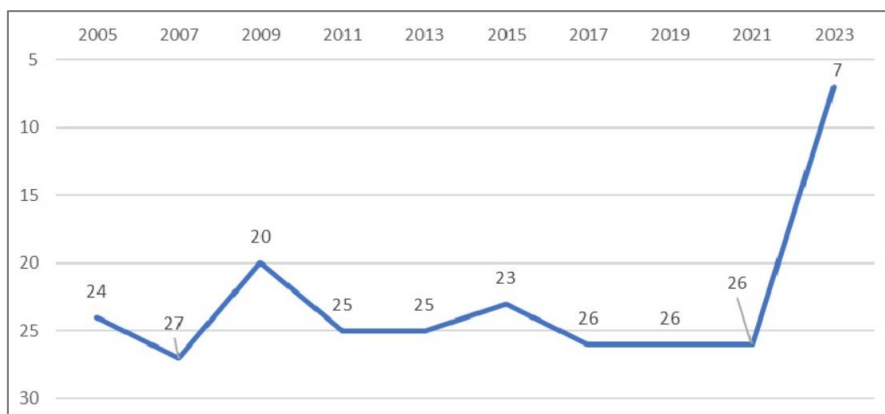
Figura 1 – Série histórica da média do ensino médio público do Pará e do Brasil no IDEB (2005-2023)



Fonte: Passos (2024) Adaptado de INEP (2023)

A figura 1, que apresenta a média do IDEB do ensino médio público do Pará em comparação com a média nacional, mostra que, apesar da melhora na posição do estado, sua proficiência continua abaixo da média brasileira. Em 2023, o Pará registrou média de 4,3, enquanto o Brasil alcançou 4,5, o que reforça a crítica de que o IDEB pode ser manipulado por estratégias administrativas que não refletem a realidade educacional.

Figura 2 – Série histórica da posição do Pará no IDEB do Ensino Médio (2005-2023)



Fonte: Passos (2024) Adaptado de INEP (2023)

Já a figura 2, que mostra a série histórica da posição do Pará no IDEB do ensino médio entre 2005 e 2023, evidencia uma ascensão abrupta: o estado passou da penúltima colocação nacional em 2021 para a sexta posição em 2023. Esse salto não decorre de avanços significativos na aprendizagem, mas sim da adoção de ciclos escolares que reduziram drasticamente as reprovações, elevando a taxa de aprovação para 99%.

Além disso, o IDEB não leva em consideração a proporção de jovens que estão fora da escola, permitindo que estados com altas taxas de evasão escolar ainda alcancem boas classificações. O índice também não avalia fatores como as condições socioeconômicas dos estudantes, a infraestrutura escolar, a formação dos docentes ou a equidade nas políticas educacionais. Essas limitações prejudicam a eficácia do indicador como um meio de planejamento e monitoramento da qualidade da educação.

Diante desse cenário, é vital reconsiderar os métodos de avaliação e os indicadores utilizados, para que eles reflitam com mais precisão a complexidade do contexto

educacional brasileiro. Embora o IDEB seja uma ferramenta útil, é necessário melhorá-lo para incluir aspectos como equidade, condições de aprendizado, infraestrutura das escolas e capacitação de professores, assegurando que os resultados obtidos sejam realmente aplicados para promover uma educação pública de qualidade acessível a todos.

2.2. Mineração de Dados

A mineração de dados (ou *data mining*) é o processo de extrair informações úteis e padrões a partir de grandes volumes de dados. Ela é muito utilizada em áreas como negócios, saúde, educação e ciência de dados para tomada de decisão baseada em evidências. Na mineração de dados, especialmente durante a etapa de modelagem, definem-se as técnicas e algoritmos mais adequados para abordar o problema em questão. A escolha da técnica depende, frequentemente, dos objetivos e das necessidades do projeto, que podem variar desde a identificação de associações até tarefas como classificação, regressão, agrupamento (clusterização) e sumarização (Goldschmidt; Passos, 2005).

2.3. Agrupamento de Dados

O *clustering*, também conhecido como análise de agrupamento, é um conjunto de métodos computacionais destinados a organizar dados em grupos homogêneos, chamados de clusters. Essas técnicas visam identificar padrões e conexões subjacentes entre os elementos, agrupando objetos com características semelhantes e separando-os dos que apresentam diferenças significativas. Além disso, o *clustering* auxilia na visualização e interpretação de grandes volumes de dados, o que apoia a tomada de decisões estratégicas e a descoberta de insights importantes em contextos complexos (Linden, 2009).

Dentro das técnicas de *clustering*, o agrupamento hierárquico se sobressai como uma abordagem de grande importância, pois possibilita a análise da estrutura intrínseca dos dados sem a exigência de determinar antecipadamente o número de grupos. Murtagh (2012) defende que essa técnica permite a formação de uma hierarquia de agrupamentos, revelando conexões detalhadas entre os elementos. Isso facilita a identificação de padrões, a análise da similaridade entre os objetos e, como resultado, a tomada de decisões mais embasadas.

O processo de agrupamento hierárquico começa com a criação de uma matriz de distâncias, normalmente fundamentada na distância Euclidiana, tratando cada elemento como um cluster individual no início. De maneira iterativa, os grupos mais parecidos são combinados de acordo com o critério de ligação utilizado, como *single linkage*, *complete linkage*, média das distâncias ou o método de *Ward*. Este último se destaca por criar agrupamentos mais homogêneos, pois diminui a variância *intracluster*. O dendrograma apresenta o resultado, mostrando a hierarquia das uniões e possibilitando que o pesquisador determine o número de clusters a partir de um ponto de corte apropriado, fornecendo uma análise estruturada, interpretativa e metodologicamente sólida.

3. Metodologia

Este estudo tem uma abordagem quantitativa, baseada na análise de indicadores educacionais do Ensino Médio para identificar padrões de semelhança e desigualdade entre os estados brasileiros. Para isso, foram empregadas duas fontes de dados: a primeira incluía variáveis educacionais adicionais, como taxas de aprovação e reprovação, distorção idade-série, carga horária média diária, número de alunos por professor e por turma, valor repassado e gastos com educação em relação ao PIB; a segunda incluía o Índice de Desenvolvimento da Educação Básica – Ensino Médio (IDEB-EM), ambas as bases fornecidas pelo INEP.

O pré-processamento dos dados deu início ao processo metodológico. Essa etapa incluiu a organização, padronização e limpeza das informações, preparando-as para a análise estatística. Posteriormente, utilizaram-se técnicas de mineração de dados com o *software Orange Data Mining*, que permite a exploração de bases educacionais com o auxílio de algoritmos de aprendizado de máquina.

Na etapa inicial, as variáveis mais significativas foram selecionadas com base no critério de ganho de informação, disponível no ambiente do *Orange*. Esse critério identifica os atributos que mais contribuem para a diferenciação dos estados. Esse método assegurou que a análise fosse realizada com atenção aos indicadores mais relevantes, removendo possíveis interferências.

Em seguida, utilizou-se a técnica de clusterização hierárquica, selecionada por sua habilidade de agrupar os estados com base nas semelhanças entre seus indicadores, apresentando os resultados por meio de dendrogramas. Essa abordagem possibilitou a análise de como os estados se agrupam em categorias de maior ou menor proximidade em termos de desempenho educacional, sem a exigência de estabelecer previamente a quantidade de clusters.

Com base nos agrupamentos obtidos, foi possível comparar os resultados com os valores do Índice de Desenvolvimento da Educação Básica – Ensino Médio (IDEB-EM). Essa análise permitiu verificar se os estados com índices elevados estão entre os grupos de melhor desempenho ou se apresentam características semelhantes aos estados mais vulneráveis. Ademais, essa análise permitiu uma reflexão crítica sobre a formulação do IDEB, que considera apenas o desempenho no Sistema de Avaliação da Educação Básica (SAEB) e a taxa de aprovação, desconsiderando fatores estruturais que impactam diretamente a qualidade da educação. O SAEB é um conjunto de avaliações em larga escala realizadas periodicamente pelo INEP, visando medir o desempenho dos alunos em matérias como Língua Portuguesa e Matemática. Os resultados alcançados servem como referência para acompanhar o aprendizado, direcionar políticas públicas e, no caso do IDEB, integrar a métrica que visa resumir a qualidade da educação básica no país.

Desse modo, a metodologia empregada, baseada em mineração de dados e clusterização hierárquica, permitiu a identificação de grupos de estados com variados níveis de vulnerabilidade e a análise da eficácia do IDEB como única métrica para avaliar o Ensino Médio no Brasil.

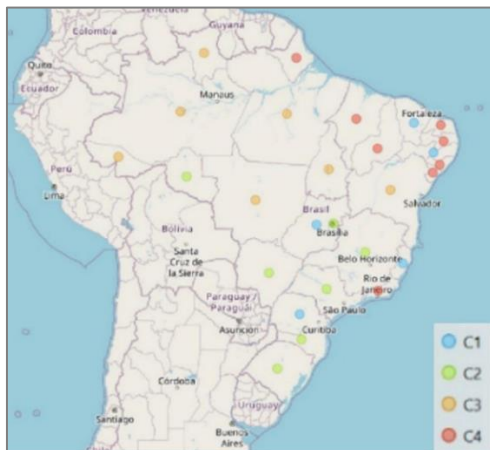
4. Resultados

A análise dos dados educacionais das redes públicas estaduais brasileiras permitiu identificar grupos com perfis semelhantes em rendimento acadêmico e características institucionais. Por meio da clusterização hierárquica, foram classificados quatro grupos de estados, enquanto diagramas de caixa e gráficos de dispersão auxiliaram na visualização das variáveis mais influentes na formação desses clusters. A integração da análise espacial evidenciou padrões regionais e disparidades educacionais, proporcionando uma compreensão mais clara da distribuição geográfica das diferenças de desempenho e vulnerabilidade entre os estados.

A análise espacial, ilustrada pelo mapa geográfico (Figura 3), permitiu observar como os clusters estão distribuídos pelo Brasil. A presença dos estados do *Cluster 4*, que teve o pior desempenho, nas áreas Norte e Nordeste destaca as desigualdades educacionais históricas. Em contrapartida, os estados do *Cluster 1*, que têm o melhor desempenho, estão distribuídos entre diferentes regiões, abrangendo dois estados do Nordeste, um do Sudeste, um do Sul e um do Centro-Oeste. Já o *Cluster 2* concentra-se majoritariamente nas regiões Sul e Sudeste, com Rondônia sendo a única exceção. O *Cluster 3*, por sua vez, é formado predominantemente por estados da Amazônia, excetuando-se a Bahia. Por fim,

o *Cluster* 4 apresenta os piores indicadores, sendo composto principalmente por estados nordestinos, além do Amapá e do Rio de Janeiro, que representam outras regiões.

Figura 3 – Dispersão no país



Fonte: Autores (2025)

Para verificar sua importância na distinção entre os *clusters*, as variáveis educacionais foram analisadas individualmente, complementando a análise espacial.

A primeira análise, referente à nota do SAEB, indicou que essa variável foi a mais relevante para diferenciar os *clusters*, apresentando o maior valor de ANOVA, como pode ser visualizado na tabela 1 a seguir. A Análise de Variância (ANOVA) é um método estatístico empregado para determinar se há diferenças significativas entre as médias de dois ou mais grupos. Neste estudo, o ANOVA possibilita a avaliação do quanto cada variável contribui para a distinção entre os agrupamentos formados, apontando quais fatores têm mais impacto na separação dos estados em relação ao desempenho educacional. Dessa forma, o alto valor de ANOVA relacionado ao SAEB demonstra que essa métrica é fundamental na criação dos grupos identificados.

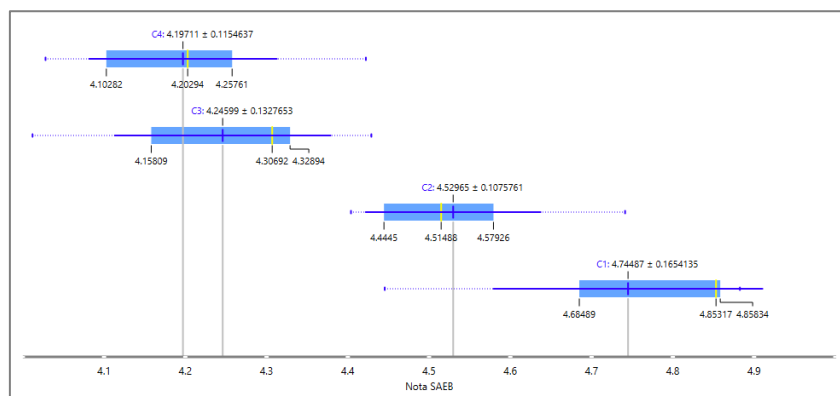
Tabela 1 – Valores ANOVA

Variável	ANOVA
Nota SAEB	20.671
Taxa de Distorção Idade-Série	9.127
Média de Horas-Aulas Diária	8.813
Alunos por Turma	5.360
Taxa de Aprovação	5.345

Fonte: Autores (2025)

A ANOVA apresenta os valores que indicam o poder explicativo de cada variável na distinção entre os clusters formados, conforme indicado na tabela 1. Quanto maior o valor de ANOVA, maior a capacidade da variável de diferenciar os grupos. Os diagramas de caixa, por sua vez, ilustram visualmente essa distinção, mostrando a distribuição das médias e a dispersão dos dados por cluster, conforme dados da Figura 4 seguir a seguir:

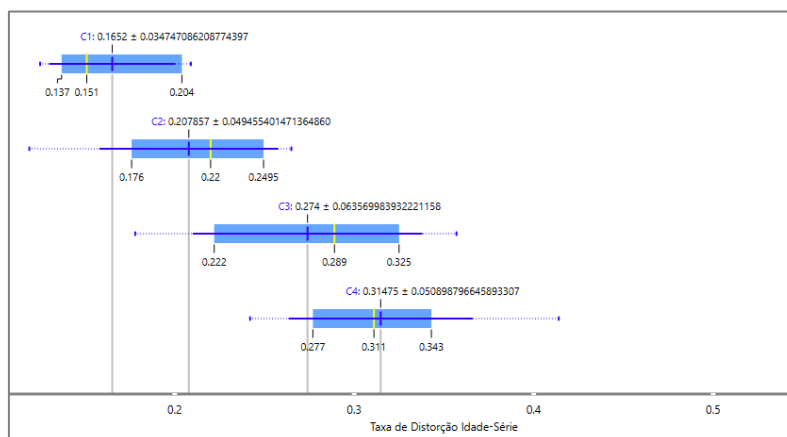
Figura 4 – Notas SAEB por *cluster*



Fonte: Autores (2025)

Na figura 4, o *Cluster 1* abrangeu os estados com as maiores médias, sendo bem definido e com baixa heterogeneidade, ao passo que o *Cluster 2* ocupou uma posição intermediária, também apresentando uma separação adequada. Já os *Clusters 3* e 4 apresentaram médias mais baixas e sobreposição entre si, o que reduziu a clareza na diferenciação. Esses resultados confirmam a centralidade do SAEB como indicador de desempenho acadêmico.

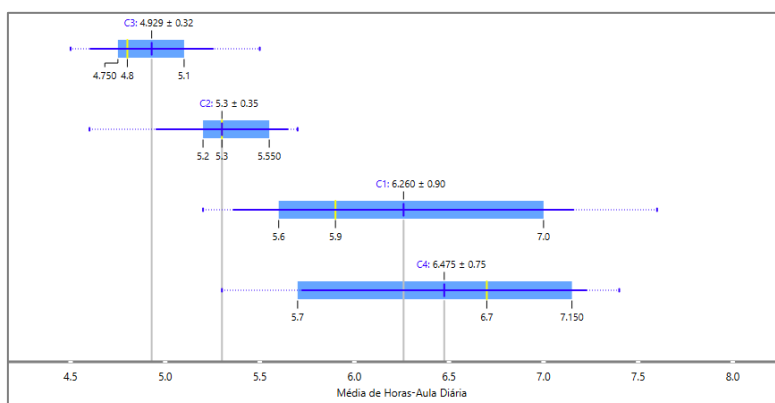
Figura 5 – Taxa de Distorção Idade-Série por cluster



Fonte: Autores (2025)

Ao analisar a taxa de distorção idade-série de acordo com a figura 5, constatou-se que a variável preservou um poder explicativo relevante, embora inferior ao do SAEB. Os resultados mostraram uma hierarquia clara: o *Cluster 1* teve a menor média de distorção, seguido pelos *Clusters 2, 3* e 4, sendo este último o que apresentou os valores mais altos. A distribuição sugere que o atraso escolar está diretamente ligado ao desempenho mais baixo, fortalecendo a conexão inversa entre rendimento acadêmico e distorção idade-série.

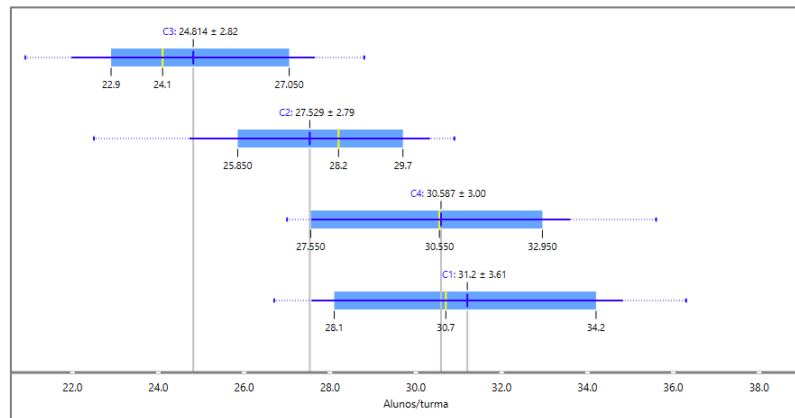
Figura 6 – Média de Horas-Aulas Diárias por cluster



Fonte: Autores (2025)

Já na figura 6, o diagrama de caixa da média de horas-aula diárias mostrou variações moderadas entre os grupos. O *Cluster 1* teve uma das maiores cargas horárias, ficando próximo ao *Cluster 4*, ao passo que os *Clusters 2 e 3* apresentaram médias mais baixas. Apesar de não ser a variável mais determinante, a análise indica que uma carga horária mais longa pode estar ligada a resultados mais positivos, mesmo que essa conexão não seja tão direta quanto nos indicadores anteriores.

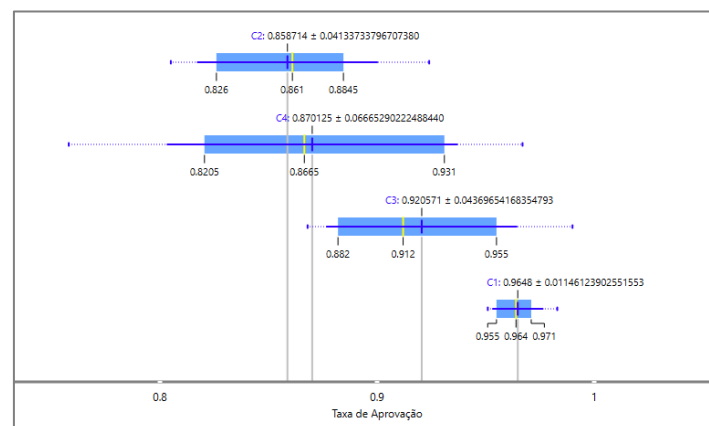
Figura 7 – Alunos por Turma por *cluster*



Fonte: Autores (2025)

A variável quantidade de alunos por turma conforme a figura 7, demonstrou baixa capacidade de distinção entre os grupos, uma vez que as médias apresentaram pouca variação e considerável sobreposição. Ainda assim, observou-se que o *Cluster 3* concentrou as menores médias, enquanto o *Cluster 1* registrou os maiores valores. Esses resultados indicam que o tamanho das turmas, isoladamente, não constitui um fator determinante para explicar as diferenças de desempenho educacional entre os estados analisados.

Figura 8 – Taxa de Aprovação por *cluster*

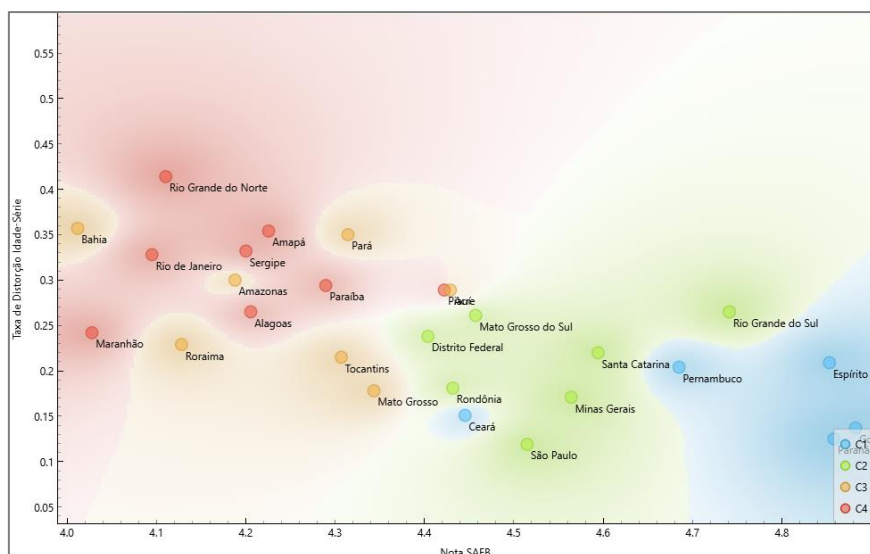


Fonte: Autores (2025)

Na figura 8, a análise da taxa de aprovação mostrou que ela é o indicador com menor capacidade explicativa dos agrupamentos, exibindo uma ampla dispersão interna e um valor de ANOVA baixo. O *Cluster 4*, em particular, apresentou uma variação significativa, com estados posicionados entre os piores e quase os melhores resultados, o que compromete sua confiabilidade como métrica distintiva. Esse comportamento destaca as restrições de utilizar a taxa de aprovação como um indicador isolado da qualidade da educação. A taxa de aprovação tem um peso considerável no cálculo do IDEB, pois, quando combinada com o desempenho no SAEB, estabelece o índice que resume a qualidade da educação básica no país. No entanto, essa métrica é pouco esclarecedora, pois está mais

ligada ao fluxo escolar do que ao nível de aprendizagem dos alunos. Assim, elevadas taxas de aprovação podem ocultar falhas estruturais e pedagógicas, indicando melhorias artificiais nos índices sem que, de fato, haja progresso significativo na qualidade da educação.

Figura 9 – Gráfico de Dispersão dos estados



Fonte: Autores (2025)

Na figura que compara a nota do SAEB com a taxa de distorção idade-série, selecionado por se tratar de uma das variáveis com maior significância na ANOVA. Os grupos aparecem de forma clara, mostrando que quando uma variável aumenta, a outra tende a diminuir. O *Cluster 1* ocupou o quadrante de melhor desempenho, apresentando uma média elevada no SAEB e baixa distorção. Em contrapartida, o *Cluster 4* registrou os piores resultados, vinculado a notas baixas e elevados índices de evasão escolar. *Clusters 2 e 3*, por sua vez, ocuparam áreas intermediárias, com alguma sobreposição, o que sugere condições de transição entre os extremos.

Na análise de clusters, o desempenho no SAEB e as taxas de idade-série apresentaram variáveis mais diferenciadoras entre os grupos. Entretanto, é essencial considerar as limitações metodológicas do IDEB ao interpretar esses resultados. Passos (2024) destaca que as alterações no sistema de reprovação especialmente a escolha de concentrar a reprovação apenas no terceiro ano do ensino médio, aliada a taxas de aprovação de 100% nos anos anteriores aumentaram o índice sem demonstrar avanços reais na aprendizagem dos alunos. Esse exemplo mostra que o IDEB pode ser artificialmente elevado ao combinar apenas o fluxo escolar e o desempenho em provas padronizadas, ignorando aspectos essenciais da educação, como evasão, frequência, inclusão e desigualdades internas. Por outro lado, o estudo realizado aqui procurou expandir essa perspectiva, incluindo fatores como as notas do SAEB, taxa de distorção idade-série, taxa de aprovação, carga horária diária e quantidade de alunos por turma. Esses indicadores possibilitaram a identificação de diferenças mais sutis e consistentes entre os estados, demonstrando que fatores estruturais e pedagógicos têm um papel fundamental no desempenho educacional. Assim, fica claro que uma análise mais completa, baseada em diversas variáveis, proporciona diagnósticos mais precisos e apoia políticas públicas mais eficientes para promover a equidade e a qualidade na educação. No estudo, observou-se que a taxa de aprovação teve pouca influência na formação dos agrupamentos, possivelmente porque as políticas de aprovação ampla reduziram as diferenças entre os estados, diminuindo assim a relevância dessa variável.

Por outro lado, a continuidade de diferenças relevantes no desempenho no SAEB entre os grupos evidencia que a aprendizagem eficaz ainda é o principal fator de distinção entre os

estados. Para superar limitações do IDEB e obter diagnósticos mais completos, seria essencial incluir indicadores como investimentos públicos, status socioeconômico, salário dos professores e taxas de transição escolar, que ainda não estavam atualizados para 2023. Esses resultados indicam que avanços aparentes no índice podem ocultar desigualdades estruturais, reforçando a necessidade de análises mais críticas da educação pública.

5. Conclusão

O uso de técnicas de mineração de dados aplicadas a indicadores educacionais demonstrou ser eficaz para identificar padrões de desempenho e desigualdades entre os estados brasileiros. A análise evidenciou que o desempenho no SAEB e a taxa de distorção idade-série foram as variáveis mais relevantes para distinguir os grupos, reforçando a relação entre aprendizagem eficaz e atraso escolar. Em contraste, a taxa de aprovação e o número de alunos por turma apresentaram baixo poder explicativo, indicando que esses indicadores, isoladamente, não representam adequadamente a qualidade da educação.

Esses achados ressaltam a necessidade de cautela na utilização de indicadores como o IDEB, que combina apenas taxas de aprovação e resultados de exames padronizados, sem considerar fatores estruturais e pedagógicos. Conforme Passos (2024), mudanças administrativas no regime de aprovação podem inflar artificialmente os resultados do IDEB, como ocorreu no estado do Pará, onde a reprovação foi concentrada no terceiro ano do ensino médio, gerando uma taxa de aprovação de 100% nos anos anteriores sem refletir avanços reais na aprendizagem.

A persistência de diferenças significativas no desempenho do SAEB entre os grupos reforça que a aprendizagem continua sendo o principal fator de distinção entre os estados. A inclusão de variáveis adicionais como distorção idade-série, carga horária e indicadores contextuais pode mitigar as limitações do IDEB e oferecer uma visão mais abrangente dos perfis educacionais. No entanto, a análise foi limitada pela ausência de dados atualizados em variáveis relevantes, como investimentos públicos, infraestrutura escolar e formação docente.

Diante disso, o estudo conclui que os avanços aparentes no IDEB podem ocultar desigualdades estruturais persistentes. Recomenda-se que políticas públicas ampliem o conjunto de indicadores utilizados, incorporando aspectos socioeconômicos e metodologias complementares, como a mineração de dados, para diagnósticos mais precisos e estratégias eficazes de enfrentamento das desigualdades educacionais. Para pesquisas futuras, sugere-se a ampliação das variáveis analisadas, a inclusão de outros ciclos da Educação Básica e a realização de estudos para acompanhar a evolução dos indicadores ao longo do tempo.

Referências

BRASIL. Constituição (1988). Constituição da República Federativa do Brasil. Brasília, DF: Senado Federal, 1988. Disponível em: https://www.planalto.gov.br/ccivil_03/constituicao/constituicao.htm.

BRASIL. Lei nº 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Diário Oficial da União: seção 1, Brasília, DF, 23 dez. 1996. Disponível em: https://www.planalto.gov.br/ccivil_03/Leis/L9394.htm.1

BRASIL. Lei nº 13.415, de 16 de fevereiro de 2017. Altera as Leis nº 9.394/1996, nº 11.494/2007 e nº 11.096/2005, e revoga a Lei nº 11.161/2005. Diário Oficial da União: seção 1, Brasília, DF, 17 fev. 2017.

CHIRINÉA, Andréia Melanda; BARREIRO, Iraíde Marques de Freitas. Qualidade da educação: eficiência, eficácia e produtividade escolar. **Revista online de Política e Gestão Educacional**, v. 22, n. 2, p. 646–660, 2018. Disponível em: <<https://periodicos.fclar.unesp.br/rpge/article/view/9254>>.

CHIRINÉA, Andréia Melanda. Mecanismos de regulação para a educação no Brasil: contribuições para o debate. **Revista Ibero-Americana de Estudos em Educação**, v. 12, n. 1, p. 168–189, 2017. Disponível em: <<https://doi.org/10.21723/riaee.v12.n1.9102>>.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining: um guia prático: conceitos, técnicas, ferramentas, orientações e aplicações**. Rio de Janeiro: Elsevier Editora Ltda, 2005. 261 p.

INEP. IDEB – Índice de Desenvolvimento da Educação Básica. Brasília, DF: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023. Disponível em: <https://www.gov.br/inep/pt-br/assuntos/indicadores-educacionais/ideb>.

LINDEN, Ricardo. Técnicas de agrupamento. **Revista de Sistemas de Informação da FSMA**, Macaé, v. 4, n. 4, p. 18–36, 2009.

MURTAGH, F. **Algorithms for hierarchical clustering: an overview**. WIREs Computational Statistics, v. 4, n. 6, p. 485–495, 2012. Disponível em: <<https://doi.org/10.1002/widm.53>>. Acesso em: 25 set. 2025.

PASSOS, Felipe Garcia. **Educação no Pará: para boa nota, aluno não estuda – escancara limitações do IDEB**. Amazônia Latitude, 15 out. 2024. Disponível em: <<https://www.amazonialatitude.com/2024/10/15/ideb-2023-para-boa-nota-aluno-nao-estuda-escancara-limitacoes-indicador/>>.

SOUZA, Luiz Humberto Rodrigues; CARDOSO, Berta Leni Costa; NUNES, Cláudio Pinto (Orgs.). **Análise de dados quantitativos na educação**. 1. ed. Recife: Omnis Scientia, 2024. Disponível em: <https://www.researchgate.net/publication/384571825_ANALISE_DE_DADOS_QUANTITATIVOS_NA_EDUCACAO>.