DESENVOLVIMENTO DE PAINÉIS BI PARA INDICADORES EDUCACIONAIS DO SOMA COM PROGRAMAS GRATUITOS

Cícero Gabriel de Sousa Cardoso[1]

Paulo Fernandes da Silva Junior[2]

INTRODUÇÃO

A gestão da educação contemporânea demanda soluções que possibilitem a análise rápida e precisa de dados, de forma a subsidiar políticas públicas bem-informadas e orientadas por evidências. Painéis de *Business Intelligence* (BI) emergem como instrumentos eficazes para essa finalidade, integrando coleta, processamento e visualização de indicadores-chave de desempenho educacional (ROMERO; VENTURA, 2024). No âmbito do programa SOMA (Sistema de Orientação, Monitoramento e Avaliação da Educação do Estado do Maranhão), essa abordagem é muito relevante para monitorar a rede de ensino do Maranhão, permitindo ajustes estratégicos e acompanhamento contínuo de desempenho escolar.

O BI consiste em práticas analíticas que transformam dados brutos em informação acionável. No setor educacional, *dashboards* facilitam o entendimento de tendências, como fluxos de matrícula, taxas de aprovação e reprovação. Ao combinar múltiplas fontes em visualizações acessíveis, podem-se identificar áreas que demandam intervenção (BROWN et al., 2015, conforme citado em ROMERO; VENTURA, 2024).

Em relação às tecnologias, soluções open de código aberto têm se mostrado altamente vantajosas por sua flexibilidade, custo zero e possibilidade de adaptação local. O uso do *Python* — com bibliotecas como *Pandas* para manipulação de dados (McKINNEY, 2010), *SciPy, scikit-learn* para clusterização (MURTAGH; LEGENDRE, 2011) e *Plotly* para visualizações interativas (PLOTLY, 2025) — oferece um ambiente completo para criar *dashboards* robustos. A biblioteca *Streamlit* viabiliza interfaces web dinâmicas, sem necessidade de desenvolvimento tradicional de *front-end* (STREAMLIT, 2025), enquanto o uso de JSON garante interoperabilidade simples entre *backend* e cliente (BRAY, 2017).

Entretanto, o desenvolvimento de painéis com dados educacionais enfrenta desafios relevantes: integração de diferentes formatos, tratamento e agregação de dados, e validação de agrupamentos, especialmente quando se emprega técnicas como clusterização hierárquica. O método de Ward é reconhecido por sua capacidade de minimizar a variância interna dos clusters, promovendo agrupamentos coerentes e interpretáveis, essenciais para diferenciações entre redes de ensino (WARD, 1963; MURTAGH; LEGENDRE, 2014). A interpretação desses resultados é potencializada por dendrogramas, que representam hierarquias de similaridade de forma visual intuitiva (BOCK, 2025).

Diante desse cenário, este trabalho se justifica pelo uso de ferramentas gratuitas para criar painéis BI aplicados aos indicadores educacionais do SOMA. Essa abordagem oferece uma solução personalizada, transparente e acessível, com capacidade de resposta contínua às necessidades da gestão pública, sem incorrer em custos de licenciamento (CAMPANTE; GONÇALVES; GONÇALVES, 2024).

Na literatura, o desenvolvimento de *dashboards* educacionais costuma privilegiar plataformas proprietárias. Contudo, estudos recentes demonstram a maturidade do ecossistema de código aberto para implementar painéis confiáveis e interativos, com escalabilidade e baixo custo (ROMERO; VENTURA, 2024). Essa constatação reforça a pertinência do presente

1 Discente em Engenharia da Computação, Universidade Estadual do Maranhão. E-mail: cicerodesousacardoso2@gmail.com

2 Doutora em Engenharia Civil, Docente do Departamento de Engenharia Civil da Universidade Estadual do Maranhão. E-mail: mayssasousa@professor.uema.br

estudo, que investiga a convergência entre metodologias analíticas avançadas e ferramentas gratuitas.

Este artigo está organizado da seguinte forma: a Seção 2 apresenta os objetivos gerais e específicos; a Seção 3 aprofunda a fundamentação teórica sobre BI, clusterização e ecossistema Python; a Seção 4 descreve a metodologia de implementação — coleta dos dados do QEdu, conversão para JSON, pipeline de clusterização e desenvolvimento do painel com Streamlit; a Seção 5 apresenta e discute os resultados; e a Seção 6 traz as conclusões e propostas para trabalhos futuros.

METODOLOGIA

O desenvolvimento do trabalho seguiu uma abordagem combinada de pesquisa bibliográfica e implementação prática do sistema, estruturada nas seguintes etapas:

- Levantamento e análise de livros, artigos, portais e sites especializados, com o objetivo de compreender os processos, ferramentas e metodologias utilizados na análise de dados educacionais, particularmente aqueles relacionados aos indicadores do Sistema de Orientação, Monitoramento e Avaliação da Educação (SOMA);
- Processo de coleta e preparação dos dados necessários para prosseguir com o desenvolvimento.
- Construção do código para converter os dados no formato *.XLSX* para *.JSON* que facilita a manipulação dos dados para gerar os *clusters*.
- Desenvolvimento do código que faz a clusterização dos dados em *.JSON* gerando os dendrogramas.
- Produção de uma página interativa utilizando *HTML*, *CSS* e *Javascript* que permite o usuário visualizar os dendrogramas criados e filtrar os anos desejados.

Inicialmente, foi conduzida uma revisão bibliográfica abrangente, contemplando livros, artigos científicos, portais acadêmicos e sites especializados. Essa etapa teve como objetivo estabelecer uma base teórica consistente para compreender, de forma aprofundada, os processos de coleta, preparação, análise e manipulação de dados educacionais referentes ao estado do Maranhão. A partir dessa fundamentação, tornou-se possível identificar e avaliar os indicadores necessários para a construção dos *dashboards* interativos desenvolvidos neste trabalho, os quais utilizaram dados das taxas de rendimento do estado no período de 2018 a 2023.

Com a escolha do site QEdu como fonte de dados, foram baixadas manualmente as planilhas referentes às taxas de rendimento escolar do estado do Maranhão, abrangendo o período de 2018 a 2023. A partir dessas planilhas, foram selecionados especificamente os indicadores de alunos matriculados, reprovados e aprovados, que serviram como base para a aplicação de uma clusterização hierárquica (QEDU, 2025). Essa análise foi realizada considerando a segmentação por tipo de escola (federal, privada, estadual e municipal), permitindo identificar padrões e similaridades no desempenho educacional entre as diferentes redes de ensino e visualizá-las por anos iniciais e finais.

Em seguida, procedeu-se à padronização e organização dos arquivos para análise. Um primeiro script em *Python* leu as planilhas no formato *.xlsx*, tratou linhas inteiramente vazias e converteu cada ano em estruturas JSON (um arquivo consolidado por fase escolar), preservando chaves/valores por indicador. O *pandas* foi empregado para ingestão tabular e transformação (filtros, remoção de vazios e reestruturação), dado seu amplo uso e eficiência na manipulação de dados estruturados (McKINNEY, 2010). A opção pelo *.JSON* visou compatibilidade direta com a camada web (JavaScript) e portabilidade entre componentes do pipeline, uma vez que o formato é um padrão consolidado de intercâmbio de dados (BRAY, 2017).

A análise de agrupamento foi conduzida por clusterização hierárquica aglomerativa, com ênfase no método de Ward, que minimiza o incremento da soma dos quadrados intra-grupo a cada fusão, favorecendo clusters compactos e interpretáveis (WARD, 1963). A implementação

utilizou *SciPy* e *scikit-learn* para cálculo de distâncias, ligação hierárquica e construção da árvore de fusões, seguindo boas práticas e algoritmos consolidados nessas bibliotecas (VIRTANEN et al., 2020; PEDREGOSA et al., 2011). O resultado foi materializado em dendrogramas por fase/ano, recurso que representa graficamente a hierarquia das fusões e auxilia na leitura de semelhanças e na escolha de cortes para formação de grupos (BOCK, 2025; MURTAGH; LEGENDRE, 2014). Os gráficos foram gerados e salvos como imagens estáticas com *matplotlib*, garantindo reprodutibilidade e qualidade de exportação (HUNTER, 2007).

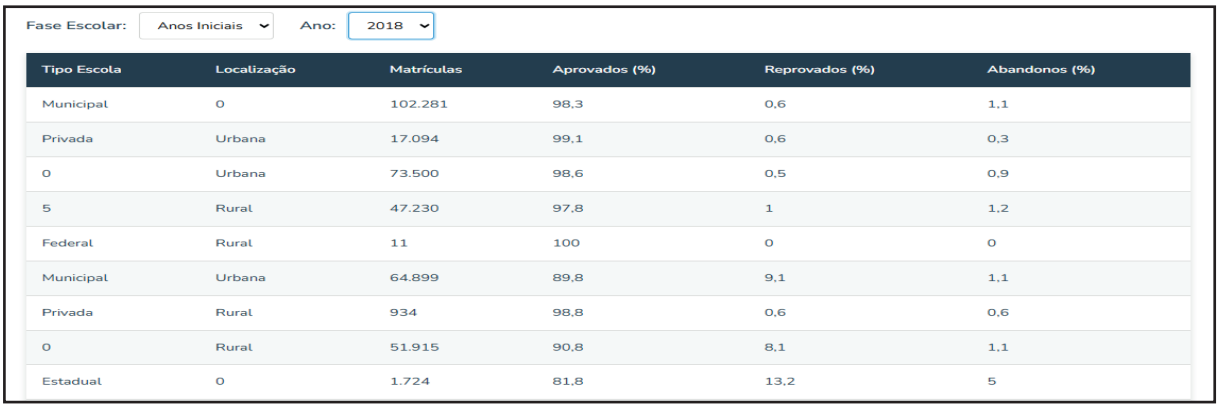
A camada de apresentação foi desenvolvida como uma página web estática (index. html) utilizando HTML5, CSS e JavaScript. Nela, o código cliente realiza o carregamento dinâmico dos arquivos .JSON produzidos no backend, aplica filtros por ano e fase e exibe os dendrogramas correspondentes (imagens geradas previamente), permitindo ao usuário alternar rapidamente entre recortes e comparar padrões entre redes de ensino. Esse arranjo separa responsabilidades: o Python concentra a preparação e a análise, enquanto o *front-end* cuida da navegação, responsividade e interação, mantendo o sistema leve, portável e reprodutível.

RESULTADOS E DISCUSSÃO

CLUSTER HIERÁRQUICO

Como ilustrado na Figura 1, a seção inicial do site apresenta a organização da planilha utilizada como fonte de dados, permitindo ao usuário navegar entre os diferentes anos e etapas escolares, com atualização simultânea da visualização. A estrutura da planilha segue a codificação do Instituto Brasileiro de Geografia e Estatística (IBGE), na qual o código "0" representa o somatório geral dos registros, tanto para o campo referente ao tipo de escola quanto para o campo referente à localização.

Figura 1- Carregamento dos dados no site.



Tipo Escola	Localização	Matrículas	Aprovados (%)	Reprovados (%)	Abandonos (%)
Municipal	0	102.281	98,3	0,6	1,1
Privada	Urbana	17.094	99,1	0,6	0,3
0	Urbana	73.500	98,6	0,5	0,9
5	Rural	47.230	97,8	1	1,2
Federal	Rural	11	100	0	0
Municipal	Urbana	64.899	89,8	9,1	1,1
Privada	Rural	934	98,8	0,6	0,6
0	Rural	51.915	90,8	8,1	1,1
Estadual	0	1.724	81,8	13,2	5

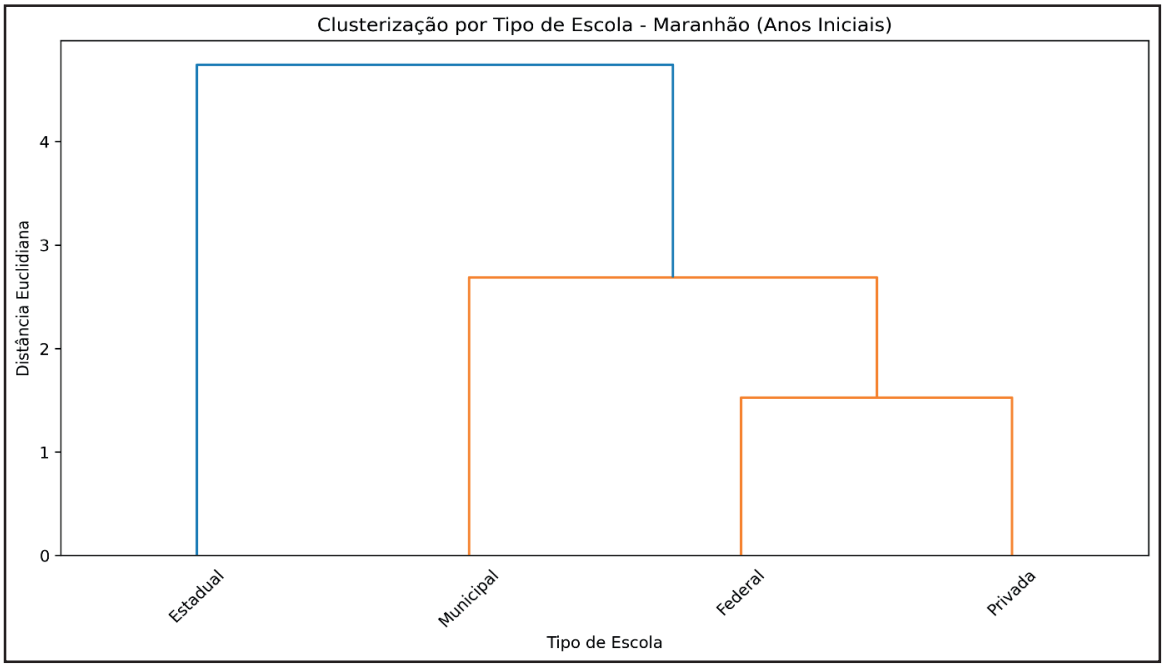
Fonte: Autor, 2025

Os dados foram convertidos para tipos numéricos e filtrados para manter somente os tipos de escolas Federal, Estadual, Municipal e Privada. As taxas foram convertidas de *strings* para *floats*, garantindo o formato adequado para análise:

- **Aprovação:** Taxa de alunos aprovados
- **Reprovação:** Taxa de alunos reprovados
- **Abandono:** Taxa de alunos que abandonaram

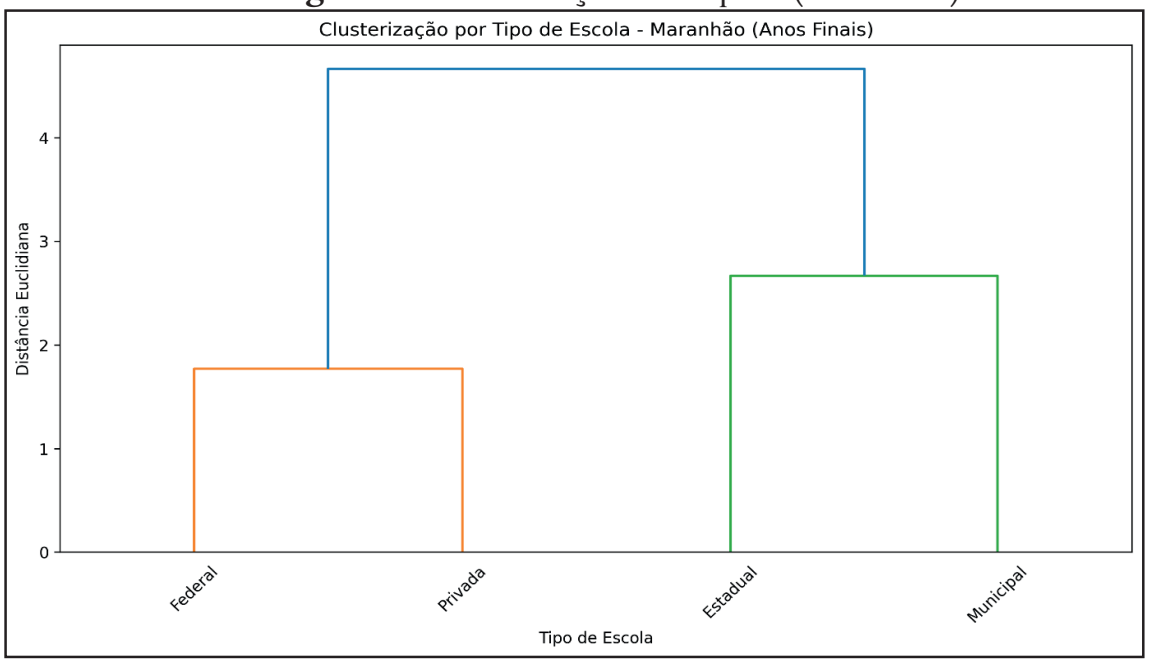
Foram removidos registros com valores ausentes ou inconsistentes para assegurar a qualidade da análise. Os dados foram agregados por tipo de escola, somando o número de matrículas e calculando a média das taxas de aprovação, reprovação e abandono. Para reduzir possíveis distorções decorrentes de diferenças de escala entre os indicadores, aplicou-se a normalização dos dados. Em seguida, realizou-se a clusterização hierárquica utilizando o método Ward, que minimiza a variância total dentro dos grupos, sendo particularmente eficaz na análise educacional ao buscar maximizar a homogeneidade interna de cada *cluster*. Com todos os processos já em andamento, pode-se então visualizar os *clusters* gerados pelo código conforme a Figura 2 e Figura 3:

Figura 2 - Clusterização hierárquica (anos iniciais).



Fonte: Autor, 2025

Figura 3 - Clusterização hierárquica (anos finais).

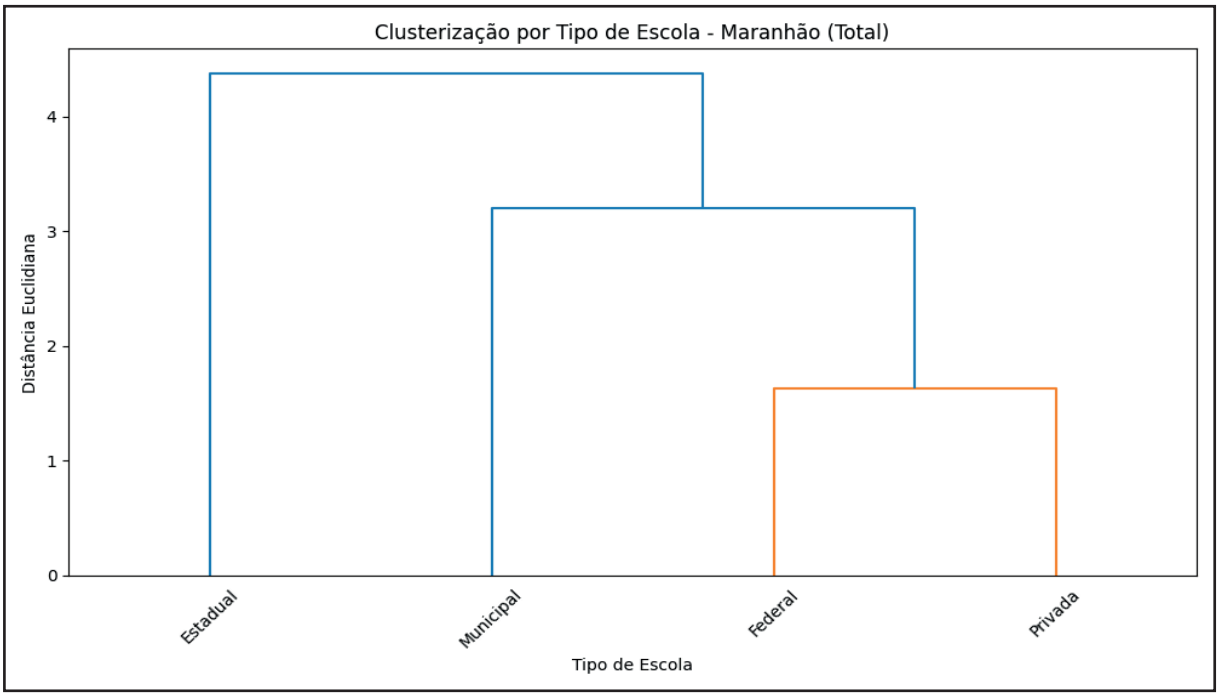


Fonte: Autor, 2025

Nos anos iniciais, observa-se uma padronização no desempenho das escolas federais e privadas, evidenciada pela proximidade entre elas no agrupamento, o que indica elevada similaridade nos indicadores analisados. Esse grupo apresenta, ainda, uma leve semelhança com as escolas municipais, as quais, por sua vez, se diferenciam de forma significativa das escolas estaduais, que constituem o grupo mais distante em relação aos demais. Já nos anos finais, observa-se a formação de dois pares distintos: escolas federais e privadas agrupam-se de maneira muito próxima, evidenciando similaridade significativa nos valores de seus indicadores; enquanto escolas estaduais e municipais formam outro agrupamento. Esses dois grupos, contudo, apresentam grande distância entre si, o que evidencia a disparidade na qualidade do ensino ofertado pela rede pública no Maranhão.

Com isso, a Figura 4 apresenta o somatório total das fases escolares organizadas em um único agrupamento, permitindo uma visão consolidada do cenário educacional no estado. A análise evidencia que, de forma geral, as escolas da rede pública (especialmente as estaduais e municipais) apresentam indicadores significativamente distintos em relação às escolas federais e privadas. Essa diferença se manifesta de maneira consistente em variáveis como taxas de matrícula, aprovação e reprovação, refletindo possíveis desigualdades estruturais, de recursos e de gestão entre os diferentes tipos de administração escolar. Tal resultado reforça a necessidade de políticas públicas voltadas à redução dessas disparidades, buscando aproximar a qualidade do ensino público daquela observada nas instituições federais e privadas.

Figura 4 - Clusterização hierárquica (anos inciais + anos finais).

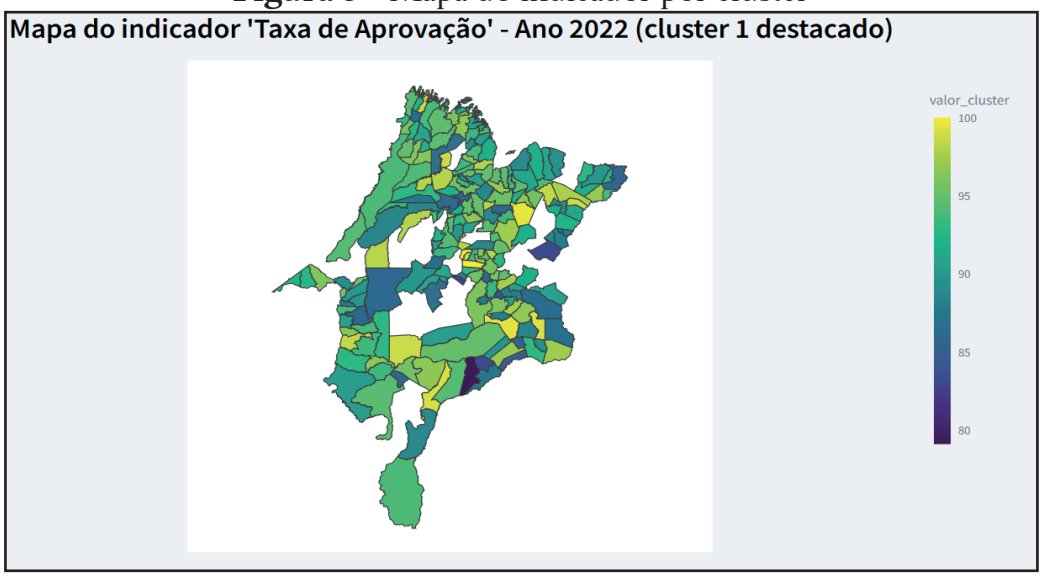


Fonte: Autor, 2025

K-MEANS

O painel de resultados do K-Means apresenta visualizações interativas que permitem compreender de maneira abrangente a estrutura dos agrupamentos obtidos a partir dos indicadores analisados. A Figura 5 ilustra um mapa temático do indicador selecionado, no qual os municípios são coloridos de acordo com seus valores, destacando-se o cluster escolhido no filtro lateral. Essa representação geoespacial possibilita identificar padrões territoriais, permitindo verificar se determinado cluster se concentra em regiões específicas ou se apresenta distribuição mais dispersa pelo estado.

Figura 5 - Mapa do indicador por cluster



Fonte: Autor, 2025

A Figura 6 apresenta a tabela de municípios pertencentes ao cluster em foco, posicionada logo abaixo do painel geoespacial. Além do nome e do código de cada município, a tabela reúne os valores de todos os indicadores utilizados no modelo, constituindo um recurso essencial para relacionar os resultados estatísticos com a realidade local. Essa visualização detalhada permite ao analista compreender como cada município se insere no agrupamento, favorecendo interpretações mais consistentes sobre os padrões identificados.

Figura 6 - Mapa do indicador por cluster

Municípios no cluster 1 - Ano 2022

	cod_ibge	municipio	Taxa de Aprovação	Taxa de Reprovação	Taxa de Abandono	Cobertura Vacinal	Número de Estabelecimentos	Númer
1	2,100,105	AFONSO CUNHA	89	6.7	4.3	66.51	21	
2	2,100,154	ÁGUA DOCE DO MA	92.2	4.8	3	98.57	30	
3	2,100,204	ALCÂNTARA	91.7	6	2.3	51.93	40	
4	2,100,303	ALDEIAS ALTAS	83.2	13.5	3.3	50.45	45	
5	2,100,402	ALTAMIRA DO MARA	92	4.9	3.1	56.87	12	
6	2,100,436	ALTO ALEGRE DO M	94	3.9	2.1	83.49	28	
7	2,100,477	ALTO ALEGRE DO PI	98.1	1.7	0.2	99.06	51	
8	2,100,501	ALTO PARNAÍBA	93.9	4	2.1	64.52	30	
9	2,100,550	AMAPÁ DO MARANH	86.5	9	4.5	67.9	11	
10	2,100,600	AMARANTE DO MAR	85.9	8.8	5.3	66.81	108	

Fonte: Autor, 2025

A Figura 7 apresenta o painel de distribuição do indicador por cluster, estruturado em um gráfico de dispersão no qual cada ponto corresponde a um município, posicionado de acordo com seu valor no indicador selecionado e categorizado pelo agrupamento a que pertence. Essa visualização facilita a comparação entre os clusters, permitindo observar tanto a amplitude quanto a concentração dos valores em cada grupo, o que auxilia na identificação de clusters mais homogêneos e daqueles que apresentam maior variação interna.

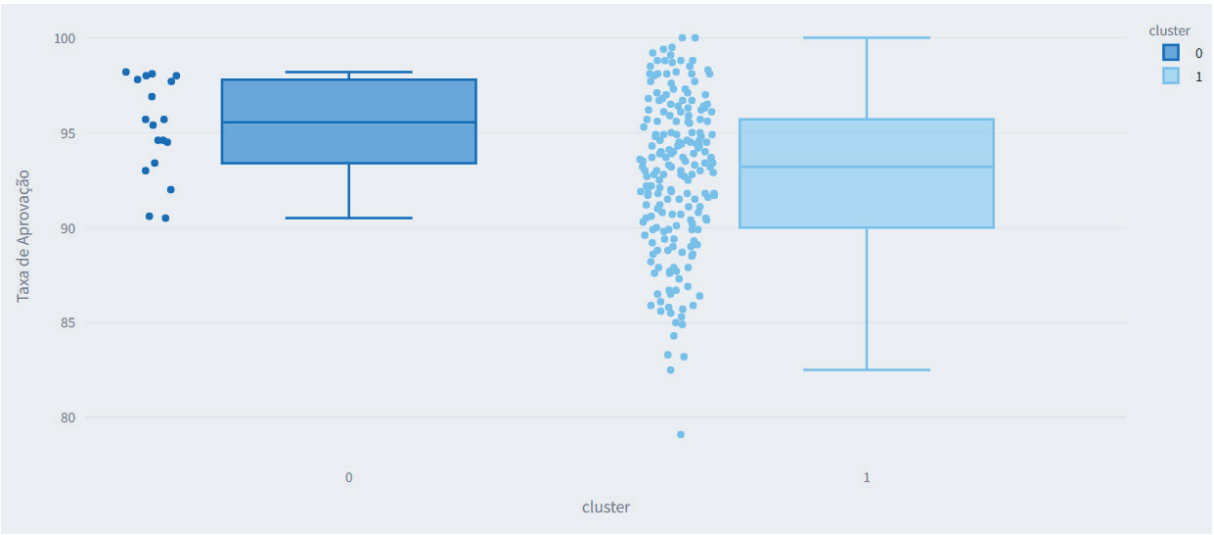
Figura 7 - Distribuição dos indicadores por cluster.



Fonte: Autor, 2025

O boxplot do indicador por cluster, mostrado na Figura 8, complementa a análise estatística ao evidenciar medidas de tendência central, como a mediana, além de informações sobre dispersão (quartis) e a presença de valores atípicos. Esse tipo de visualização torna mais clara a identificação de clusters com desempenho elevado, baixo ou irregular no indicador avaliado, permitindo também uma comparação rápida e objetiva entre grupos.

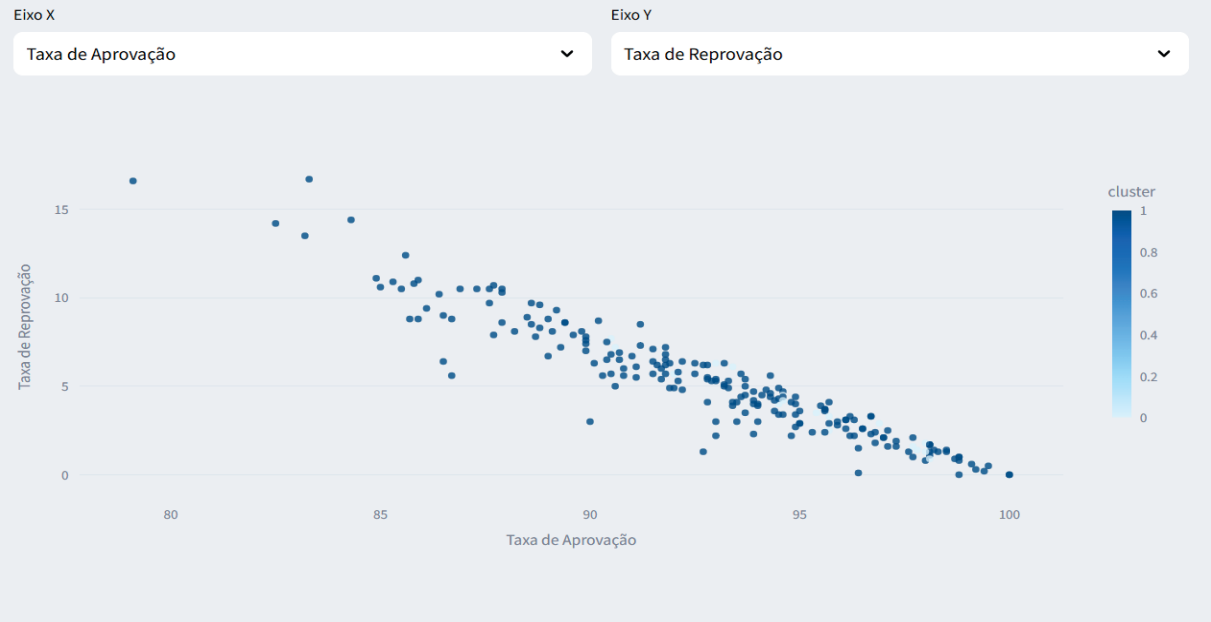
Figura 8 - Boxplot dos indicadores por cluster.



Fonte: Autor, 2025

O painel de relação entre dois indicadores apresenta um gráfico de dispersão bidimensional, mostrado na Figura 9, no qual cada município é posicionado de acordo com seus valores nas duas variáveis escolhidas para os eixos X e Y. A cor dos pontos indica o cluster ao qual cada município pertence, permitindo observar padrões de correlação entre os indicadores e avaliar o posicionamento relativo dos grupos. Essa análise combinada é especialmente útil para identificar relações diretas ou inversas entre variáveis, bem como para destacar subgrupos que se afastam do padrão geral.

Figura 9 - Relacionamento entre os dois indicadores.

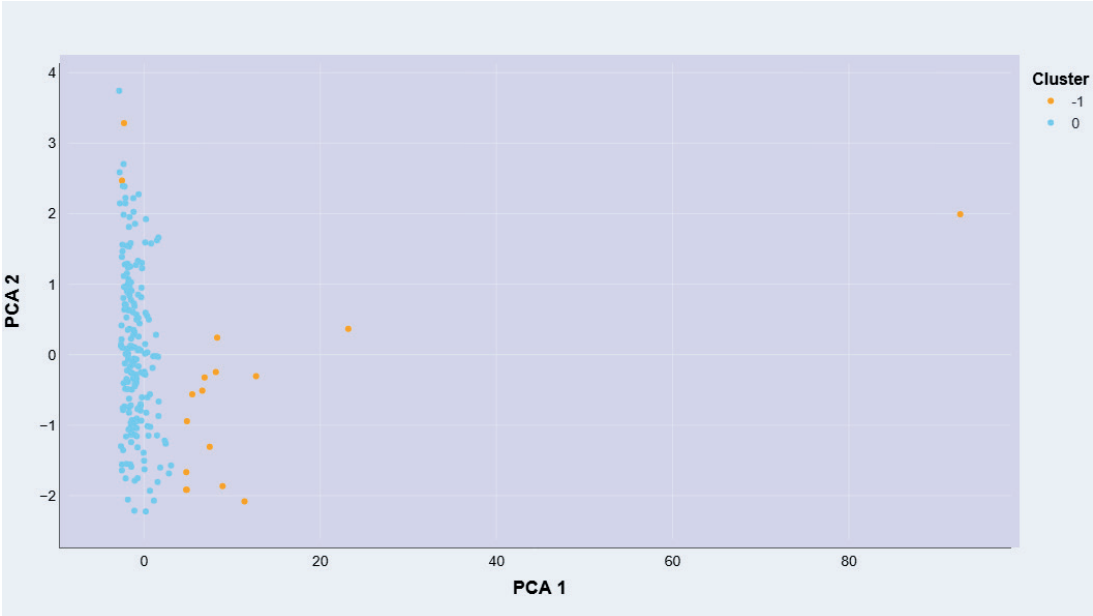


Fonte: Autor, 2025

DBSCAN

O painel de resultados do DBSCAN reúne duas visualizações principais e uma tabela, oferecendo uma interpretação clara dos agrupamentos obtidos. A primeira visualização corresponde à projeção dos municípios em duas dimensões por meio da Análise de Componentes Principais (PCA), ilustrada na Figura 10. Nessa representação, cada ponto corresponde a um município, colorido de acordo com o cluster ao qual pertence, o que facilita identificar a distribuição espacial dos grupos e suas diferenças internas, destacando tanto os agrupamentos mais concentrados quanto os municípios mais afastados do perfil médio.

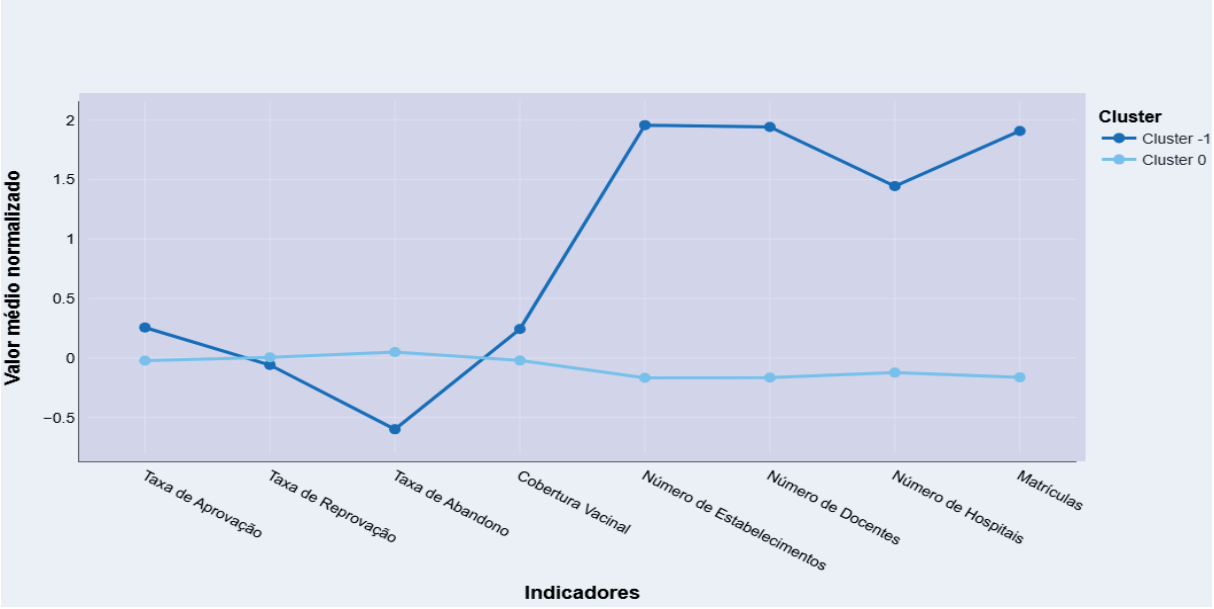
Figura 10 - Projeção PCA dos municípios.



Fonte: Autor, 2025

O segundo painel, mostrado na Figura 11, reúne as médias normalizadas dos indicadores para cada cluster. O gráfico possibilita visualizar, de maneira comparativa, o desempenho relativo dos grupos em todas as variáveis consideradas, destacando aquelas que se encontram acima ou abaixo do valor médio do conjunto. Essa abordagem auxilia na definição do perfil característico de cada cluster, evidenciando diferenças em dimensões como educação, saúde e outros aspectos contemplados na análise.

Figura 11 - Projeção PCA dos municípios.



Fonte: Autor, 2025

Por fim, a Figura 12 apresenta a tabela localizada na parte inferior do painel, onde são listados os municípios pertencentes ao cluster selecionado no filtro lateral. Esse recurso possibilita identificar, de forma detalhada, as localidades que compõem cada agrupamento, favorecendo a análise individualizada dos resultados.

Figura 12 - Municípios de cada cluster.

	Municipio	Codigo
0	AÇAILÂNDIA	2100055
1	BACABAL	2101202
2	BALSAS	2101400
3	BARRA DO CORDA	2101608
4	CAXIAS	2103000
5	CHAPADINHA	2103208
6	CODÓ	2103307
7	GRAJAÚ	2104800
8	IMPERATRIZ	2105302
9	PAÇO DO LUMIAR	2107506

Fonte: Autor, 2025

CONCLUSÃO

Neste estudo, aplicamos técnicas de análise estatística e mineração de dados para compreender padrões nos indicadores educacionais do Maranhão, com foco nos anos iniciais e finais do ensino fundamental. A utilização da clusterização hierárquica pelo método de Ward, aliada à normalização prévia dos dados, mostrou-se altamente eficaz para agrupar escolas com características semelhantes e revelar discrepâncias significativas entre redes de ensino.

A escolha da metodologia foi fundamentada pela capacidade do método Ward em minimizar a variância interna de cada grupo, garantindo que os clusters formados fossem o mais homogêneos possível. Essa abordagem é especialmente relevante no campo educacional, onde compreender a proximidade ou distância entre diferentes tipos de escola pode auxiliar na formulação de políticas públicas mais direcionadas.

Durante o processo, registros com valores ausentes ou inconsistentes foram removidos para preservar a qualidade dos resultados. Em seguida, os dados foram agregados por tipo de escola, somando matrículas e calculando médias das taxas de aprovação, reprovação e abandono. Essa organização permitiu uma visualização clara das diferenças entre escolas federais, privadas, estaduais e municipais, revelando padrões como a proximidade entre redes federal e privada, e o distanciamento da rede estadual em relação às demais.

Os resultados evidenciaram desigualdades persistentes, especialmente entre as redes pública e privada, reforçando a necessidade de estratégias voltadas à melhoria da qualidade do ensino público. A abordagem adotada, por ser metodologicamente robusta e replicável, oferece um caminho sólido para análises futuras, podendo ser aplicada em outros contextos e regiões.

Conclui-se que a aplicação combinada de pré-processamento rigoroso e clusterização hierárquica é uma ferramenta poderosa para diagnosticar e monitorar indicadores educacionais. Essa metodologia não apenas facilita a identificação de padrões e disparidades, mas também fornece subsídios concretos para a tomada de decisão de gestores e formuladores de políticas educacionais, contribuindo para uma gestão mais eficiente e voltada para a equidade no ensino.

REFERÊNCIAS

BOCK, Tim. What is a dendrogram? Displayr, 2025. Disponível em: https://www.displayr.com/what-is-dendrogram/. Acesso em: 13 ago. 2025.

BRAY, Tim. The JavaScript Object Notation (JSON) Data Interchange Format. *RFC 8259*, IETF, 2017. Disponível em: https://www.rfc-editor.org/rfc/rfc8259. Acesso em: 13 ago. 2025.

CAMPANTE, G.; GONÇALVES, R.; GONÇALVES, P. L. Ferramentas *open source* aplicadas à educação. *Revista de Educação e Tecnologias*, v. 5, n. 2, p. 45-60, 2024.

COSTA, Cledjan Torres da. União de dados por clusterização para construção de modelos de predição de evasão. 2024. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Piauí, Centro de Ciências da Natureza, Teresina, 2024.

DUNN, J. C. Well-separated clusters and optimal fuzzy partitions. *Journal of Cybernetics*, v. 4, n. 1, p. 95–104, 1974.

HAN, J.; PEI, J.; KAMBER, M. Data Mining: Concepts and Techniques. 3rd ed. Amsterdam, The Netherlands: Elsevier, 2012. 744 p.

HUNTER, J. D. Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, v. 9, n. 3, p. 90–95, 2007.

McKINNEY, W. Data structures for statistical computing in Python. In: *Proceedings of the 9th Python in Science Conference*. 2010. p. 51–56.

MURTAGH, F.; LEGENDRE, P. Ward's Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm. *arXiv*, 27 nov. 2011. Disponível em: https://arxiv.org/abs/1111.6285. Acesso em: 13 ago. 2025.

MURTAGH, F.; LEGENDRE, P. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification*, v. 31, p. 274–295, 2014.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011.

PLOTLY. Plotly Python Graphing Library. Disponível em: https://plotly.com/python/. Acesso em: 13 ago. 2025.