

Learning Analytics Aplicada ao Desempenho Escolar Brasileiro com Dados da PeNSE e IDEB

Guilherme B. Camboim¹, Dirceu A. M. Junior¹, Tiago T. Primo¹, Ig I. Bittencourt²

¹Programa de Pós-Graduação em Computação – Universidade Federal de Pelotas (UFPEL)
Caixa Postal 354 – 96.010-900 – Pelotas – RS – Brasil

²Universidade Federal de Alagoas (UFAL)
Caixa Postal 61 – 57.072-900 – Maceió – AL – Brasil

{guilherme.camboim,dmaraschin,tiago.primo}@inf.ufpel.edu.br

ig.ibert@gmail.com

Abstract. Introduction: PISA 2022 revealed low performance among Brazilian students in mathematics, reading, and science. **Objective:** To investigate factors associated with academic performance using the K-Means algorithm applied to PeNSE and IDEB data. **Method:** Data from the 2009, 2015, and 2019 PeNSE surveys were analyzed in ecological association with IDEB to identify student profiles. **Results:** Low parental supervision, unhealthy eating habits, physical inactivity, and early alcohol and drug use were associated with lower IDEB performance, whereas healthy habits and socio-emotional support were associated with better outcomes. Socioeconomic inequalities remained key determinants of academic performance.

Keywords Learning Analytics, Educational Data Mining, K-Means, Academic Performance, Machine Learning.

Resumo. Introdução: O PISA 2022 evidenciou baixo desempenho dos estudantes brasileiros em matemática, leitura e ciências. **Objetivo:** Investigar fatores associados ao rendimento escolar por meio do algoritmo K-Means aplicado à PeNSE e ao IDEB. **Metodologia:** Foram analisados dados da PeNSE (2009, 2015 e 2019) em associação ecológica com o IDEB para identificar perfis de estudantes. **Resultados:** Baixa supervisão parental, alimentação inadequada, sedentarismo e uso precoce de álcool e drogas associaram-se a menor desempenho no IDEB, enquanto hábitos saudáveis e apoio socioemocional relacionaram-se a melhores resultados. Desigualdades socioeconômicas permaneceram determinantes do desempenho escolar.

Palavras-Chave Learning Analytics, Mineração de Dados Educacionais, K-Means, Desempenho Escolar, Aprendizagem de Máquina.

1. Introdução

A educação, especialmente no que diz respeito à avaliação e ao aperfeiçoamento dos sistemas de ensino, tem sido foco de intensos debates e pesquisas acadêmicas há décadas. Mesmo assim, os resultados do Programa Internacional de Avaliação de Alunos (PISA) de 2022 escancaram um cenário alarmante: 73% dos estudantes brasileiros não alcançaram o nível mínimo de proficiência em matemática, 50% ficaram abaixo do esperado em

leitura e 55% em ciências. Mais grave ainda, as médias do Brasil permanecem muito aquém das registradas pelos países da Organização para a Cooperação e Desenvolvimento Econômico (OCDE): 379 pontos em matemática contra 472 da OCDE, 410 em leitura frente a 476 e 403 em ciências diante de 485 [OCDE 2022]. Esses números não apenas evidenciam um atraso consistente, como reforçam a urgência de uma revisão profunda nas políticas e práticas educacionais do país.

Ainda no contexto brasileiro, um relatório do Fundo das Nações Unidas para a Infância [UNICEF et al. 2023] evidenciou que aproximadamente 32 milhões de crianças e adolescentes vivem em situação de pobreza multidimensional, uma condição que transcende a noção tradicional de pobreza monetária, caracterizando-se pela sobreposição de privações, exclusões e múltiplas vulnerabilidades. De acordo com o estudo, essas crianças enfrentam restrições significativas relacionadas à renda, à segurança alimentar, ao acesso à informação e à educação, bem como à exposição precoce ao trabalho infantil, à precariedade habitacional e à ausência de infraestrutura básica, como abastecimento de água potável e saneamento adequado. O relatório também evidencia disparidades regionais marcantes: enquanto menos de 40% das crianças residentes no estado de São Paulo e no Distrito Federal enfrentam algum tipo de privação, esse percentual ultrapassa 90% em estados como Rondônia e Amapá.

Para [Gomes 2018], elementos como condições habitacionais e sanitárias inadequadas, fatores afetivo-emocionais, vulnerabilidades socioeconômicas, ambiente familiar desestruturado, comprometimentos na saúde física e mental dos estudantes e limitações na formação e valorização docente constituem um ecossistema adverso à aprendizagem. Estudos recentes também corroboram essa perspectiva multifatorial. [Alves Filho 2024] argumenta que uma alimentação equilibrada e a prática regular de atividades físicas não apenas contribuem para a saúde geral, mas são determinantes diretos no desenvolvimento cognitivo de crianças e adolescentes. Em outra dimensão, [Soares 2018] sustenta que a ausência de supervisão parental, a instabilidade emocional e o desinteresse pelo ambiente escolar não são apenas sintomas periféricos, mas fatores centrais que comprometem a trajetória educacional dos estudantes. Estudos como estes não apenas reforçam a complexidade do problema, mas também desafiam abordagens reducionistas que desconsideram o entrelaçamento de dimensões sociais, emocionais e comportamentais no processo de aprendizagem.

Assumir essa complexidade é determinante para formular análises mais robustas que possam servir de base para o desenvolvimento de políticas públicas mais eficazes e sensíveis à realidade dos estudantes. Nesse contexto, a análise de dados educacionais tornou-se uma ferramenta poderosa para identificar padrões e tendências que possam orientar essas intervenções pedagógicas e administrativas. Diante desse cenário, de múltiplas variáveis que podem se relacionar com a educação, técnicas avançadas de análise de dados têm se consolidado no campo da *learning analytics* como estratégias promissoras, ideais para a identificação de padrões ocultos, com ou sem hipóteses iniciais previamente definidas [Costa et al. 2012]. Entre esses métodos, o algoritmo K-means destaca-se por ser amplamente utilizado e consolidado na tarefa de agrupamento [da Silva et al. 2023], além de não exigir hipóteses prévias — característica particularmente adequada à natureza exploratória dos dados utilizados no estudo.

Com base nisso, o estudo associou diferentes dimensões que, segundo a literatura,

impactam o aprendizado, limitando-se, neste trabalho, à análise de fatores diretamente relacionados aos alunos, conforme apontado por [Gomes 2018]. A pesquisa foi orientada pela seguinte questão: com base em fatores da Pesquisa Nacional de Saúde do Escolar (PeNSE)¹, é possível identificar padrões distintos entre os estados brasileiros que se relacionem a diferentes trajetórias de desempenho no IDEB? Para responder a essa questão, aplicou-se o algoritmo K-Means aos dados da PeNSE coletados em 2009, 2015 e 2019, associando os agrupamentos identificados aos resultados do Índice de Desenvolvimento da Educação Básica (IDEB) e analisando sua variação temporal. Os resultados permitiram evidenciar padrões recorrentes associados à educação brasileira ao longo do tempo, fornecendo subsídios para o aprimoramento das políticas educacionais.

O presente trabalho organiza-se da seguinte forma: o Capítulo 1 apresenta a introdução e os objetivos; o Capítulo 2 aborda a fundamentação teórica; o Capítulo 3 descreve a metodologia; o Capítulo 4 apresenta e discute os resultados; e o Capítulo 5 reúne as conclusões e sugestões para trabalhos futuros.

2. Fundamentação Teórica

2.1. Fatores associados à aprendizagem

A aprendizagem pode ser influenciada por múltiplos fatores que vão desde o número de alunos por sala, a formação e remuneração dos docentes, a infraestrutura escolar, a gestão e a gerência educacional, os investimentos realizados, até aspectos sociais e contextos socioeconômicos, entre outros [de Araújo et al. 2021]. Esses elementos, em conjunto, criam um ambiente que pode influenciar o processo de ensino-aprendizagem. Diante desse cenário, constata-se que a aprendizagem ocorre muito além dos limites da sala de aula.

Nesse sentido, a Pesquisa Nacional de Saúde do Escolar (PeNSE) — um levantamento amostral que coleta dados sobre saúde, comportamentos de risco, contexto social e condições socioeducacionais de crianças e adolescentes — fornece informações valiosas para a compreensão dessa variável tão complexa: a educação [IBGE 2009]. Dada a riqueza e a multidimensionalidade de seu escopo, a PeNSE vem sendo amplamente utilizada em estudos de diversas áreas, como: a prática de bullying entre escolares brasileiros e fatores associados [Mello et al. 2017], a tendência temporal da atividade física em adolescentes [Soares et al. 2023], e os desafios para as políticas públicas voltadas à adolescência e juventude [Reis et al. 2018]. Apesar de sua relevância, é fundamental reconhecer algumas limitações metodológicas, como a natureza autorrelatada das informações que pode introduzir vieses de memória ou de desejabilidade social além das variações amostrais significativas entre os estados, que afetam a comparabilidade e a generalização dos resultados.

Considerando as limitações e potencialidades da PeNSE, a integração com indicadores educacionais como o IDEB permite uma análise associativa exploratória mais robusta. Essa métrica possibilita o monitoramento integrado da eficácia dos sistemas de ensino, articulando os resultados de aprendizagem com a progressão adequada dos estudantes [INEP 2025]. Esse indicador apresenta resultados de esferas públicas e privadas em nível nacional, estadual, municipal e por escola.

¹PeNSE: <https://www.ibge.gov.br/estatisticas/sociais/populacao/9134-pesquisa-nacional-de-saude-do-escolar.html?edicao=31442&t=o-que-e>

Vale ressaltar que, embora amplamente utilizado em políticas públicas e estudos educacionais, esse indicador da educação brasileira possui limitações que devem ser consideradas. Quanto maior a abrangência do índice, maior a tendência de mascaramento de desigualdades internas relevantes, especialmente entre redes privadas e públicas, ou entre zonas urbanas e rurais. Além disso, críticas têm sido direcionadas aos efeitos de distorções provocadas por políticas de aprovação automática, que inflacionam o componente de fluxo de aprovação sem necessariamente refletir uma melhora real na aprendizagem dos alunos.

Conforme aponta [Silva 2025], o ensino médio da rede estadual de Goiás teve, entre 2005 e 2021, um acréscimo de 15,5% na taxa de aprovação, atingindo 98,4% — um valor que beira a aprovação total. Tais aspectos impõem desafios adicionais à análise comparativa entre estados, especialmente quando se busca correlacionar indicadores de desempenho como o IDEB com dados autorrelatados e comportamentais da PeNSE.

2.2. Aprendizado de máquina

Diante da complexidade dos dados e do grande volume de informações disponíveis, recorre-se ao Aprendizado de Máquina para explorar padrões ocultos e identificar relações não triviais entre as variáveis. O Aprendizado de Máquina é uma subárea da Inteligência Artificial que propõe o desenvolvimento de sistemas capazes de aprender padrões a partir de dados. Essa aprendizagem pode ser classificada em duas categorias principais [Naeem et al. 2023]:

- (i) **Aprendizado Supervisionado:** Utiliza conjuntos de dados rotulados, onde cada amostra de entrada possui um rótulo de saída conhecido para inferir uma função de mapeamento. Após o treinamento, o modelo pode generalizar e prever rótulos para novos dados não rotulados. Dentre os principais algoritmos, destacam-se a Regressão Linear e *Random Forest* [Naeem et al. 2023].
- (ii) **Aprendizado Não Supervisionado:** Opera sobre dados não rotulados, identificando automaticamente padrões, agrupamentos e estruturas intrínsecas. Os resultados requerem interpretação posterior por especialistas. Neste trabalho, utilizamos o algoritmo K-means [Hernández-Leal et al. 2021]. Dentre os algoritmos de Aprendizado de Máquina não supervisionado, o K-Means é uma técnica amplamente reconhecida. Ele é utilizado sem hipóteses prévias, os dados semelhantes são agrupados automaticamente, sem a necessidade de rótulos prévios, o que o torna ideal para conjuntos de dados heterogêneos, como os da PeNSE. Seu funcionamento pode ser descrito pelos seguintes passos [Corcovia e Alves 2019]:
 - 1. Particiona N pontos de dados em K grupos (clusters);
 - 2. Atribui cada ponto ao cluster cujo centróide está mais próximo (distância euclidiana);
 - 3. Atualiza iterativamente os centróides (média dos pontos no cluster);
 - 4. Repete até convergência (minimizando a soma das distâncias quadradas intra-cluster).

Vale ressaltar que, assim como qualquer algoritmo, o K-Means possui limitações. Para determinar a quantidade ideal de agrupamentos, recorre-se à técnicas como o Método

do Cotovelo, o coeficiente de Silhouette e os índices de Davies-Bouldin e Calinski-Harabasz. Dentre esses métodos, o principal, e mais utilizado para essa finalidade, é o Método do Cotovelo [Lenz et al. 2020]. Esse método utiliza métricas para identificar o valor ideal de k , a partir do ponto em que o aumento do número de clusters deixa de gerar ganhos significativos [Cui et al. 2020]. Já o coeficiente de Silhouette avalia a coesão e a separação dos clusters, enquanto os índices de Davies-Bouldin e Calinski-Harabasz medem, respectivamente, a compacidade/separação e a dispersão intra e intercluster [Furlanetto et al. 2022]. A combinação desses métodos permite identificar, de maneira mais robusta, a quantidade de clusters que melhor representa a estrutura dos dados.

Para fins de visualização e interpretação da separação e da coesão dos grupos gerados, utilizam-se algumas técnicas, dentre elas o *t-Distributed Stochastic Neighbor Embedding* (t-SNE), uma técnica de redução de dimensionalidade voltada à visualização de dados em duas ou três dimensões, preservando estruturas locais e globais dos dados [Maaten e Hinton 2008]. Já o *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP) também realiza redução de dimensionalidade com base na preservação das vizinhanças locais, destacando-se pela escalabilidade e pela capacidade de gerar separações mais nítidas entre clusters [Diaz-Papkovich et al. 2021]. Essas técnicas de visualização, somadas às métricas de validação anteriormente descritas, contribuem para garantir a qualidade e a confiabilidade dos grupos gerados.

3. Metodologia

Foram utilizados os microdados das PeNSEs (2009, 2015 e 2019) e os indicadores do IDEB dos mesmos anos, estabelecendo a associação a partir das unidades da federação (UF) e da esfera administrativa. A metodologia foi estruturada nas etapas ilustradas no fluxograma da Figura 1, que pode ser resumida da seguinte forma: seleção de dados, análise exploratória, pré-processamento, clusterização e análise dos resultados.

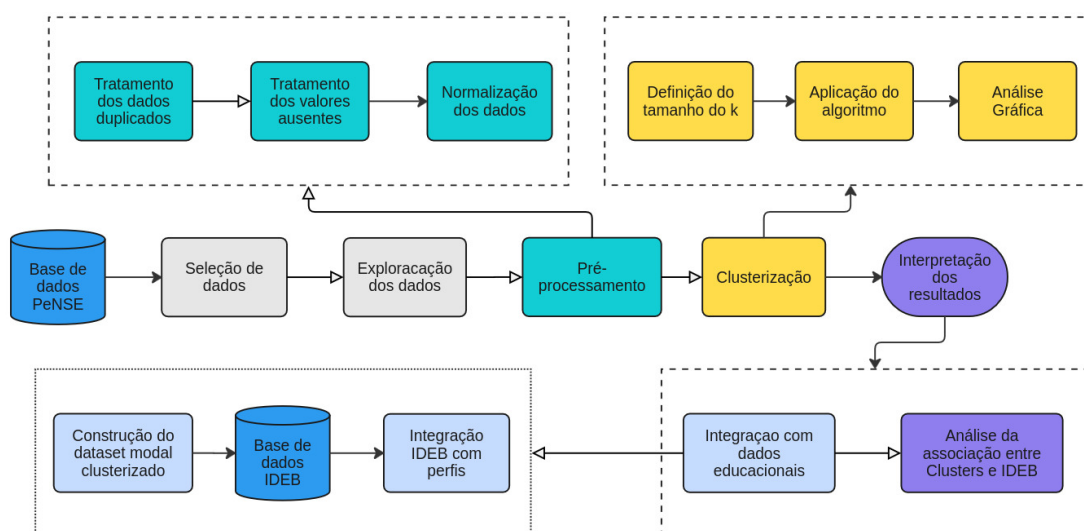


Figura 1. Fluxograma da metodologia adotada no processo de análise de associação dos dados educacionais

3.1. Seleção e preparo dos dados

Os dados utilizados neste estudo foram obtidos a partir de duas fontes públicas oficiais: PeNSE e IDEB. A escolha dos anos analisados deu-se pela convergência de informações entre as duas bases de dados. A seleção de dados das bases da PeNSE ocorreu da seguinte maneira: foram selecionados apenas os estudantes do 9º ano do Ensino Fundamental e do 3º ano do Ensino Médio de todo o país — com exceção do ano de 2009, que não contempla informações sobre o Ensino Médio. Essa escolha foi feita para alinhar-se aos resultados do IDEB, que considera os mesmos anos de escolaridade. Todas as variáveis disponíveis dos estudantes foram consideradas, exceto aquelas identificadas como irrelevantes para o processo de clusterização, como variáveis sintéticas e estatísticas desenvolvidas pela PeNSE, além de variáveis relacionadas à estrutura escolar, que não fazem parte do escopo desta pesquisa.

Já a seleção dos dados do IDEB foi realizada apenas após a etapa de clusterização, uma vez que sua seleção depende da esfera (pública ou privada) e da unidade federativa correspondente a cada cluster. A partir dessas informações, foram selecionados os dados do IDEB com base na esfera, no ano, na unidade federativa (UF) e na média dos resultados dos anos finais do Ensino Fundamental e do Ensino Médio. Ressalta-se que, para o ano de 2009, considerou-se exclusivamente o 9º ano do Ensino Fundamental, a fim de manter a compatibilidade metodológica com a PeNSE 2009, que, em sua primeira edição, não contemplava dados referentes ao Ensino Médio.

3.2. Exploração dos dados

Foram utilizados métodos de estatística descritiva, como média, mediana, moda, valores mínimos e máximos, quartis, percentis, variância e desvio padrão. Além disso, os dados foram visualizados por meio de histogramas, e não foram encontrados *outliers*. Essas técnicas foram importantes para o reconhecimento dos dados em questão, além de evidenciar a necessidade de métodos mais avançados para a identificação de padrões ocultos.

3.3. Pré-processamento

O pré-processamento dos dados foi conduzido em quatro etapas sequenciais:

Tratamento de duplicatas: Identificação e remoção de registros duplicados no conjunto de dados, garantindo a exclusividade dos registros.

Tratamento de valores ausentes: Verificação da presença de valores ausentes em todas as variáveis. Optou-se por não excluir variáveis com dados faltantes, preservando informações potencialmente relevantes. Os valores ausentes foram tratados por imputação: moda para variáveis categóricas numéricas e média para variáveis numéricas contínuas, devido ao baixo percentual de ausência na maioria das variáveis. O preenchimento foi realizado com o algoritmo *SimpleImputer*, da biblioteca *scikit-learn*².

Normalização de dados: A normalização foi realizada por meio do algoritmo *Normalizer*, disponível na biblioteca *scikit-learn*, nas variáveis que continham escalas de valores (tempo).

²Documentação do *Scikit learn*: <https://scikit-learn.org/>

Análise de outliers: Foram identificados outliers pelo intervalo interquartil (IQR) e escore robusto de Z, porém optou-se por mantê-los, pois podem representar grupos raros e relevantes para a formação dos clusters.

3.4. Clusterização com K-Means

A clusterização com o algoritmo K-Means foi conduzida em três etapas sequenciais:

Determinação do número ótimo de clusters: Aplicou-se o K-Means aos dados para k entre 2 e 10, com inicialização *k-means++*. O número de clusters foi definido principalmente pelo método do cotovelo (*elbow method*), enquanto o coeficiente de silhueta, o índice de Calinski-Harabasz e o índice de Davies-Bouldin serviram para confirmar essa escolha. Em casos de divergência ou empate entre as métricas, adotou-se o método do cotovelo como critério final.

Aplicação do algoritmo K-Means: O K-Means foi aplicado com o valor de k definido na etapa anterior, inicialização *k-means++*, *n_init* = 10 e máximo de 300 iterações. O algoritmo foi configurado para parar quando não houvesse mudança significativa nos centróides (critério de tolerância padrão do *scikit-learn*). Para garantir reprodutibilidade, foram testadas múltiplas *seeds*, sendo a final fixada em 42. Assim, com os mesmos dados e parâmetros, o algoritmo produzirá resultados idênticos; porém, *seeds* diferentes podem gerar resultados distintos.

Análise gráfica: Para visualizar e interpretar a separação e coesão dos grupos, aplicou-se o algoritmo t-SNE para projeção bidimensional (2D), visando representar com maior fidelidade as relações de proximidade entre indivíduos no espaço original de alta dimensionalidade. Complementarmente, utilizou-se o algoritmo UMAP para projeção tridimensional (3D), a fim de explorar possíveis estruturas não lineares e padrões de agrupamento em outra perspectiva de visualização.

3.5. Análise dos Resultados

Após a clusterização dos dados, a análise foi conduzida pelas seguintes etapas:

Construção do dataset clusterizado: Após a clusterização do dataset bruto, contendo uma coluna indicando o cluster de pertencimento de cada observação, optou-se por representar cada grupo por meio da moda de suas variáveis, em vez de utilizar médias ou proporções ponderadas. Essa escolha se justifica pelo caráter predominantemente categórico das variáveis, no qual a moda representa de forma mais adequada o valor mais frequente e característico de cada cluster, além de simplificar a análise e a interpretação dos grupos identificados.

Integração com Dados Educacionais: Com base nos grupos obtidos, considerando a unidade federativa, esfera administrativa (pública ou privada) e o ano da pesquisa, associaram-se os valores do IDEB correspondentes. Como não foi possível identificar se os estudantes de cada grupo pertenciam ao Ensino Fundamental ou ao Ensino Médio (pois na metodologia deste estudo foram clusterizados juntos), utilizou-se a média das notas do IDEB das duas etapas de ensino para cada grupo. Importante ressaltar que essa associação foi feita apenas em nível agregado por UF, esfera administrativa e ano, caracterizando-se como ecológica, sem pareamento individual e portanto não permitindo inferência causal.

Criação do dataset de variáveis divergentes: Aplicou-se um filtro para manter apenas as variáveis discriminantes entre os clusters, com o objetivo de identificar divergências e comparar as diferenças entre os grupos.

Geração do arquivo final: O arquivo final foi elaborado em formato XLSX, contendo três abas. A primeira apresenta o dataset descrito no item 1, acrescido dos dados do IDEB mencionados no item 2. A segunda contém o dataset com as variáveis divergentes, conforme descrito no item anterior. Por fim, a terceira aba apresenta o dicionário com os valores correspondentes às respostas de cada variável, a fim de auxiliar na interpretação dos resultados.

Método de interpretação: Para a geração dos insights apresentados nos resultados, a interpretação foi realizada por meio da comparação entre o cluster associado ao pior IDEB e o cluster associado ao melhor IDEB, desconsiderando-se os clusters com valores intermediários do índice neste estudo. Vale ressaltar que, para o ano de 2009, houve empate no IDEB dos agrupamentos estaduais de GO e MA; nesse caso, o desempate foi feito com base no IDEB total do estado, o que resultou na escolha de MA como o cluster com o pior IDEB.

4. Resultados e Discussão

Este capítulo apresenta os principais resultados obtidos a partir da metodologia descrita no capítulo anterior, incluindo as etapas de análise exploratória, pré-processamento, definição da quantidade de grupos, clusterização e identificação de padrões nos períodos analisados. Após a seleção dos dados, foram obtidos 63.411 registros em 2009, 100.547 em 2015 e 48.404 em 2019. No pré-processamento, identificaram-se 377 registros duplicados em 2009, 19 em 2015 e nenhum em 2019. Apenas 2015 apresentou *missing values* (1.736 casos), tratados por imputação da média nas variáveis numéricas contínuas.

Assim, a definição do número de grupos resultou nos valores apresentados na Tabela 1, e a aplicação do K-Means foi realizada com base nessas métricas. Em 2009 e 2015, as métricas indicaram 2 ou 3 clusters, sendo adotado $k = 3$ com base no método do cotovelo e na convergência dos indicadores. Em 2019, observou-se melhor desempenho para 4 ou 5 clusters, sendo escolhido $k = 4$ por apresentar melhor equilíbrio entre as métricas e maior consistência com o método do cotovelo.

Tabela 1. Métricas de Avaliação para Definição do Número de Grupos do K-Means

Métrica	2009	2015	2019
Cotovelo	3	3	4
Coeficiente de Silhueta	2	2	4
Calinski Harabasz	2	2	5
Índice de Davies-Bouldin	2	2	4

Após o cálculo das métricas de validação, foram geradas projeções bidimensionais e tridimensionais por meio dos algoritmos t-SNE e UMAP, visando representar a estrutura dos dados e complementar a avaliação da coesão e separação entre grupos (Tabela 1).

Em 2009 (Figura 2), observaram-se três agrupamentos bem definidos e com baixa sobreposição. Em 2015 (Figura 3), manteve-se a formação de três clusters, porém com

maior dispersão intra-grupo. Já em 2019 (Figura 4), a presença de quatro clusters indicou novas segmentações e maior distanciamento entre alguns grupos, sugerindo mudanças nos padrões comportamentais e educacionais ao longo do período analisado.

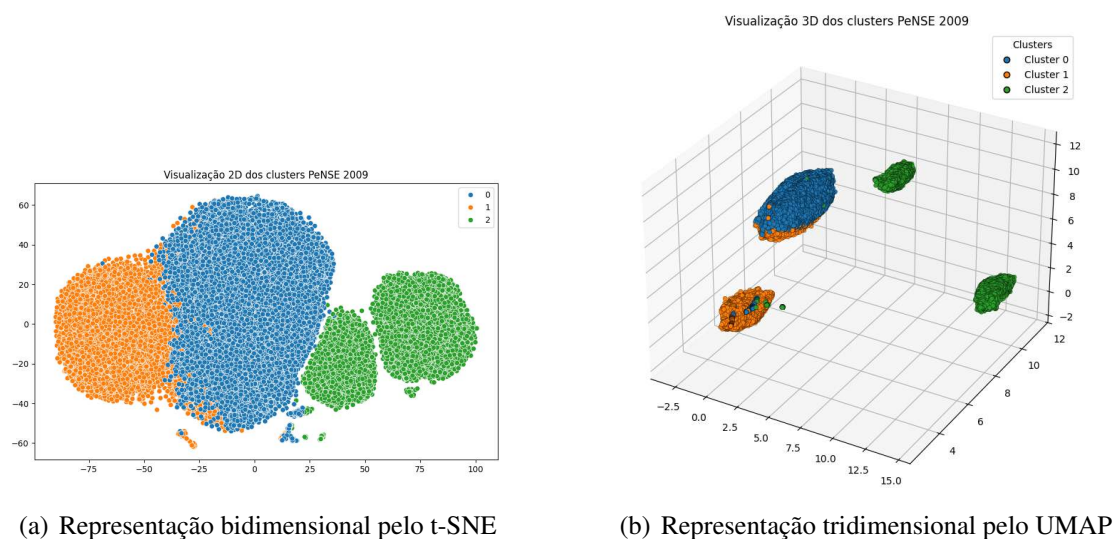


Figura 2. Clusterizações obtidas para 2009.

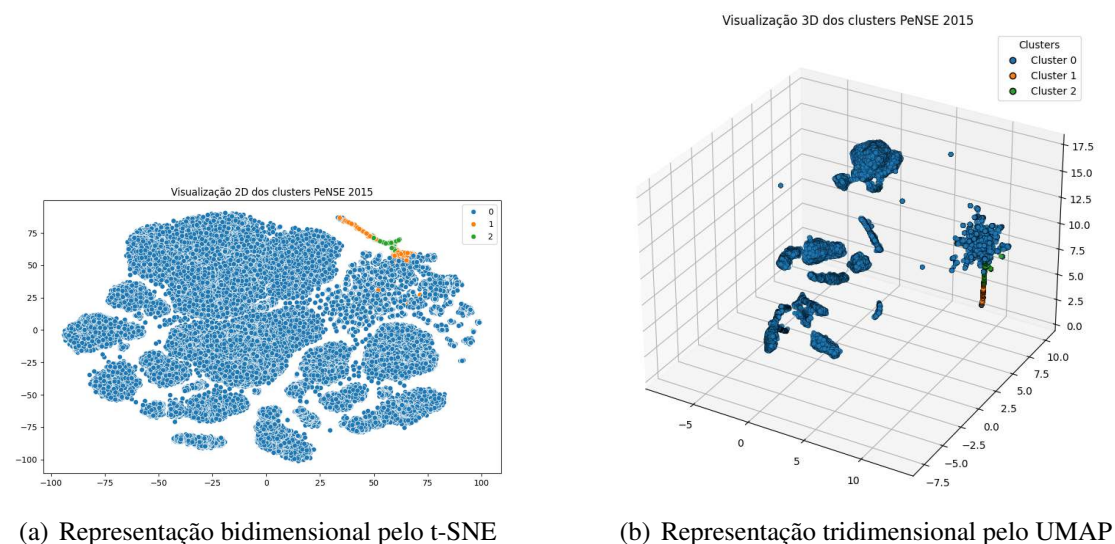


Figura 3. Clusterizações obtidas para 2015.

4.1. Análise por período

A análise dos ciclos avaliados (2009, 2015 e 2019) revelou os principais padrões marcantes, organizados nos seguintes eixos: Fatores Sociais e Econômicos (i), Supervisão Parental (ii), Atividade Física (iii), Saúde Emocional/Autoimagem (ix), Drogas/Álcool/Tabagismo (x) e Alimentação (xi). A seguir, descrevem-se as tendências observadas, complementadas pelo gráfico da Figura 5, que compara os estados com melhores e piores desempenhos no IDEB, utilizando valores binários (-1 e +1) para destacar os contrastes.

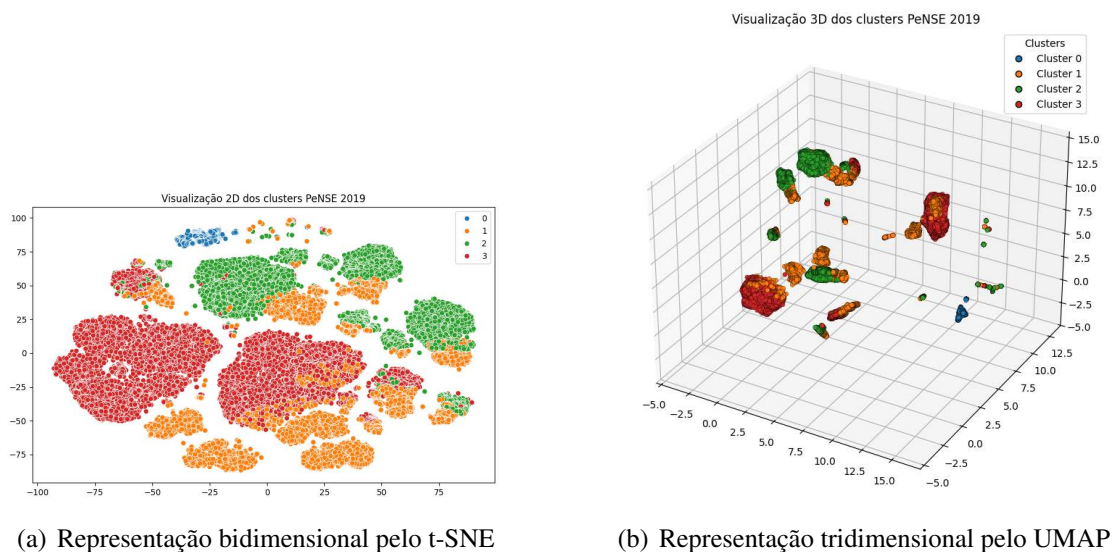


Figura 4. Clusterizações obtidas para 2019.

- (i) **Fatores Sociais e Econômicos:** Os grupos de estudantes de estados com melhores resultados no IDEB (MG) apresentaram melhores condições socioeconômicas do que os de estados com pior desempenho (MA). Portanto, grupos com fatores sociais e econômicos mais favoráveis se associaram consistentemente ao grupo com melhor desempenho no IDEB. Ressalta-se, no entanto, que esse padrão foi identificado apenas em 2009, o que limita a generalização da análise.
- (ii) **Supervisão Parental:** Os grupos de estudantes dos estados com melhores resultados no IDEB (MG, SP e GO) relataram maior acompanhamento parental, diferentemente dos estados com piores desempenhos (MA, PI e RJ). A maior supervisão parental se associou consistentemente ao grupo de melhor IDEB.
- (iii) **Atividade Física:** Os estudantes de estados com melhores resultados no IDEB (MG, SP e GO) relataram maior frequência de atividades físicas, enquanto os de estados com piores desempenhos (MA, PI e RJ) não apresentaram esse padrão. A prática regular de atividade física se associou consistentemente ao grupo de melhor IDEB.
- (iv) **Saúde Emocional/Autoimagem:** Indicadores de saúde emocional e autoimagem também se associaram aos resultados educacionais. Estudantes de estados com melhor desempenho no IDEB, como GO e SP, relataram maior bem-estar emocional e autoestima, respectivamente, em comparação àqueles de estados com desempenho inferior, como RJ e PI. Essas variáveis se associaram de forma consistente aos agrupamentos com melhores resultados no IDEB. No entanto, a limitação da disponibilidade dessas variáveis — saúde emocional apenas em 2019 e autoimagem apenas em 2015 — reduz a robustez e a generalização dos achados.
- (v) **Drogas/Álcool/Tabagismo:** O consumo precoce de substâncias como álcool, tabaco e outras drogas mostrou-se associado aos estados com pior desempenho no IDEB (como PI e RJ), enquanto a iniciação mais tardia esteve presente nos estados com melhores resultados (como SP e GO). Especificamente, estudantes de estados com IDEB mais elevado relataram contato com álcool e drogas em idades mais avançadas, além de iniciação mais tardia no tabagismo — variável

disponível apenas no ciclo de 2019. Esses comportamentos se associaram de forma consistente aos agrupamentos com melhor desempenho educacional.

- (vi) **Alimentação:** Estudantes de estados com melhor desempenho no IDEB (SP e GO) relataram maior consumo de frutas e/ou legumes, enquanto os de estados com piores resultados (PI e RJ) apresentaram menor adesão a esse padrão. A alimentação mais saudável se associou consistentemente ao grupo de melhor IDEB.

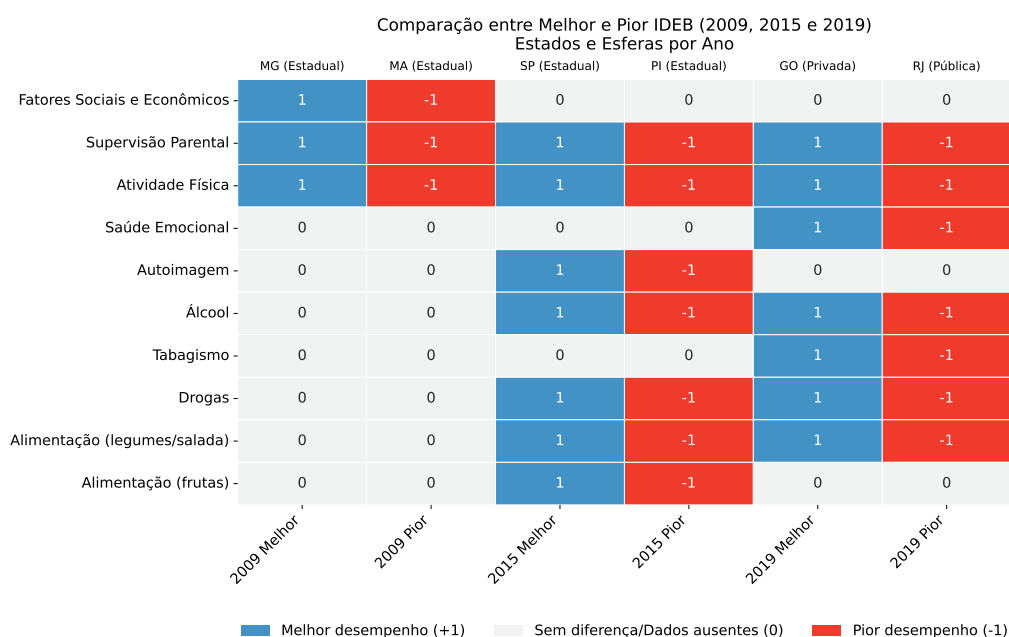


Figura 5. Comparação entre Melhor e Pior IDEB por Estados e Esferas

4.2. Padrões recorrentes entre os levantamentos

Conforme demonstrado na seção anterior, diversas variáveis apresentaram recorrência entre as edições analisadas da PeNSE — apesar das limitações impostas pela heterogeneidade dos questionários —, evidenciando padrões consistentes entre os levantamentos transversais repetidos. Para a construção da generalização dos comportamentos observados, foram selecionados os fatores que se repetiram em pelo menos duas edições da pesquisa. Os resultados indicam que estudantes com alimentação inadequada, baixa supervisão parental, prática insuficiente de atividade física e exposição precoce ao álcool e às drogas estiveram consistentemente associados a estados com menores desempenhos no IDEB. Em contrapartida, estudantes que não apresentavam essas características tenderam a estar associados a estados com melhores resultados no índice.

5. Considerações Finais

Este estudo investigou quais padrões de fatores comportamentais e de estilo de vida, medidos pela PeNSE, podem ser identificados entre os estados brasileiros e como se relacionam ao desempenho no IDEB. Para isso, aplicou-se o algoritmo K-means aos dados de 2009, 2015 e 2019, considerando estudantes da rede pública e privada em

cada estado. Os resultados indicaram 3 agrupamentos em 2009 e 2015, e 4 em 2019. Observou-se que perfis associados à baixa supervisão parental, alimentação inadequada, sedentarismo e exposição precoce ao álcool e drogas estiveram relacionados aos piores desempenhos no IDEB, enquanto estados com melhores resultados apresentaram maior supervisão e menor exposição a comportamentos de risco.

A abordagem adotada apresenta limitações associadas ao caráter ecológico da análise, ao uso de dados autorrelatados da PeNSE e à ausência de pareamento individual com o IDEB. Diferenças entre os ciclos da pesquisa e o uso de médias agregadas podem introduzir distorções. Além disso, possíveis vieses amostrais, multicolinearidade e a sensibilidade do K-Means à escala e à inicialização podem influenciar os agrupamentos obtidos.

Apesar das limitações inerentes à abordagem ecológica e exploratória adotada, o estudo atingiu seus objetivos ao identificar perfis distintos de estudantes associados ao desempenho no IDEB, considerando diferenças regionais e entre as redes pública e privada. Os resultados evidenciaram padrões recorrentes entre fatores comportamentais, estilo de vida e desempenho educacional, destacando que perfis caracterizados por baixa supervisão parental, alimentação inadequada, sedentarismo e exposição precoce ao álcool e drogas estiveram associados aos menores indicadores educacionais. Em contrapartida, estados com melhores desempenhos apresentaram maior supervisão familiar, hábitos alimentares mais adequados e menor exposição a comportamentos de risco.

Os achados reforçam o potencial da integração entre dados da PeNSE, indicadores do IDEB e técnicas de *Machine Learning*, oferecendo uma abordagem exploratória replicável para investigar fatores associados ao desempenho escolar em âmbito nacional. Além da contribuição científica, o estudo fornece subsídios para políticas públicas voltadas à educação e saúde escolar, auxiliando na priorização de ações preventivas, no fortalecimento do apoio familiar e comunitário e na focalização territorial de recursos e estratégias educacionais. Os resultados também apresentam relevância prática ao possibilitar a identificação de contextos com maior presença de perfis de risco, contribuindo para intervenções educacionais e de saúde. Entre as possíveis aplicações, destacam-se a integração entre políticas educacionais, apoio familiar e saúde escolar, bem como a utilização dos perfis identificados como instrumento de focalização territorial para auxiliar na distribuição de recursos e estratégias por UF e esfera administrativa.

Como perspectivas futuras, recomendam-se a utilização de dados mais granulares e individualizados, respeitando a LGPD, a aplicação de outros algoritmos de clusterização e modelos preditivos, bem como a validação dos agrupamentos com apoio de especialistas da área educacional. Também se destaca o potencial de ampliar a análise para níveis municipais e escolares, permitindo investigações mais precisas sobre desigualdades regionais e fatores associados ao desempenho educacional.

Conclui-se que foi possível identificar padrões regionais e recorrências entre os diferentes levantamentos transversais da PeNSE, relacionando comportamentos estudantis ao desempenho educacional, apesar das limitações quanto à causalidade e à generalização. O estudo contribui para a compreensão desses padrões em dados nacionais, fornece evidências úteis para gestores e abre caminhos para futuras investigações com novas abordagens analíticas e diferentes níveis de detalhamento dos dados.

6. Uso de Inteligência Artificial

Ferramentas de Inteligência Artificial Generativa foram utilizadas como apoio à revisão gramatical e ao aprimoramento da redação deste trabalho. Os autores revisaram integralmente o conteúdo e assumem total responsabilidade pelas informações apresentadas.

7. Disponibilidade dos Artefatos

Com o objetivo de favorecer a reprodutibilidade da pesquisa, todos os artefatos utilizados neste estudo estão disponíveis no repositório público: <https://github.com/guilhermecamboim/pense-clustering-analysis>. O repositório inclui o código-fonte, o dicionário de variáveis harmonizadas, os parâmetros utilizados na execução dos algoritmos, as versões das bibliotecas empregadas, os critérios de exclusão adotados, os resultados das análises de sensibilidade e as instruções para reprodução dos experimentos. Os microdados da PeNSE e os indicadores do IDEB utilizados são públicos e podem ser obtidos nos portais oficiais do IBGE e do INEP.

8. Agradecimentos

Os autores agradecem o apoio financeiro do Instituto Nacional de Ciência e Tecnologia em Inteligência Artificial na Educação Desplugada (INCT IA.Edu), financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), processo nº 408483/2024-5 e pela Fundação de Amparo à Pesquisa do Estado de Alagoas (FAPEAL), processo nº 60030.0000001481/2025.

Referências

- Alves Filho, E. M. A. (2024). A importância da alimentação e atividade física na educação infantil. *Revista Ibero-Americana de Humanidades, Ciências e Educação*, 10(4):1527–1528.
- Corcovia, L. O. e Alves, R. S. (2019). Aprendizagem de máquina e mineração de dados: avaliação de métodos de aprendizagem. *Revista Interface Tecnológica*, 16(1):90–101.
- Costa, E., Baker, R. S., Amorim, L., Magalhães, J., e Marinho, T. (2012). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.
- Cui, M. et al. (2020). Introduction to the k-means clustering algorithm based on the elbow method. *Accounting, Auditing and Finance*, 1(1):5–8.
- da Silva, G. C., da Cunha, R. S. B. P., Rodrigues, R. L., e Maciel, A. M. A. (2023). Utilização da mineração de dados educacionais na educação infantil. *Revista Novas Tecnologias na Educação*, 21(1):251–260.
- de Araújo, J. M., de Almeida, F. M., Gonzaga Martin, D., Marques Ferreira, M. A., e Rodrigues de Faria, E. (2021). Fatores escolares como determinantes do desempenho dos alunos da educação básica. *Linhas Críticas*, 27.
- Diaz-Papkovich, A., Anderson-Trocmé, L., e Gravel, S. (2021). A review of umap in population genetics. *Journal of Human Genetics*, 66(1):85–91.

- Furlanetto, G. C., de Carvalho, V. O., Baldassin, A., e Manacero, A. (2022). Algoritmos de agrupamento aplicados à detecção de fraudes. In *Escola Regional de Alto Desempenho de São Paulo (ERAD-SP)*, pages 29–32. SBC.
- Gomes, M. M. (2018). Fatores que facilitam e dificultam a aprendizagem. *Revista Educação Pública, Rio de Janeiro*, 18(14):28–38.
- Hernández-Leal, E., Duque-Méndez, N. D., e Cechinel, C. (2021). Unveiling educational patterns at a regional level in colombia: data from elementary and public high school institutions. *Heliyon*, 7(9).
- IBGE (2009). Pesquisa nacional de saúde do escolar (pense). <https://www.ibge.gov.br/estatisticas/sociais/saude/9134-pesquisa-nacional-de-saude-do-escolar.html>. Acesso em: 2026-08-04.
- INEP (2025). Índice de desenvolvimento da educação básica (ideb). <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb>. Acesso em: 2026-08-04.
- Lenz, M., NEUMAN, F., Santarelli, R., e Salvador, D. (2020). Fundamentos de aprendizagem de máquina. *Porto Alegre: SAGAH*.
- Maaten, L. v. d. e Hinton, G. (2008). Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605.
- Mello, F. C. M., Silva, J. L. d., Oliveira, W. A. d., Prado, R. R. d., Malta, D. C., e Silva, M. A. I. (2017). A prática de bullying entre escolares brasileiros e fatores associados, pesquisa nacional de saúde do escolar 2015. *Ciência & Saúde Coletiva*, 22:2939–2948.
- Naeem, S., Ali, A., Anam, S., e Ahmed, M. M. (2023). An unsupervised machine learning algorithms: Comprehensive review. *International Journal of Computing and Digital Systems*.
- OCDE (2022). Pisa 2022 results (volume i and ii) - country notes: Brazil. Acesso em 13 jul. 2024.
- Reis, A. A. C. d., Malta, D. C., e Furtado, L. A. C. (2018). Desafios para as políticas públicas voltadas à adolescência e juventude a partir da pesquisa nacional de saúde do escolar (pense). *Ciência & saúde coletiva*, 23(9):2879–2890.
- Silva, F. M. (2025). A educação estadual de goiás à beira da aprovação automática dos estudantes. *Revista Terceiro Incluído*, 15(1):e15104–e15104.
- Soares, C. A. M., Leão, O. A. d. A., Freitas, M. P., Hallal, P. C., e Wagner, M. B. (2023). Tendência temporal de atividade física em adolescentes brasileiros: análise da pesquisa nacional de saúde do escolar de 2009 a 2019. *Cadernos de Saúde Pública*, 39(10):e00063423.
- Soares, Maria Livanete, R. B. (2018). Desestrutura familiar e desinteresse escolar: uma avaliação multidimensional. *Revista Atlante: Cuadernos de Educación y Desarrollo (septiembre 2018)*. <https://www.eumed.net/rev/atlante/2018/09/desestructura-familiar.html/hdl.handle.net/20.500.11763>.

UNICEF et al. (2023). As múltiplas dimensões da pobreza na infância e na adolescência no Brasil. *Brasília, DF: Unicef.*