UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Programa de Pós-Graduação em Computação

Dissertação

ANÁLISE DE FATORES ASSOCIADOS AO DESEMPENHO EDUCACIONAL NO BRASIL: UMA ABORDAGEM POR AGRUPAMENTO COM K-MEANS NOS DADOS DA PENSE E IDEB

Guilherme de Barros Camboim

Pelotas, 2025

Guilherme de Barros Camboim

ANÁLISE DE FATORES ASSOCIADOS AO DESEMPENHO EDUCACIONAL NO BRASIL: UMA ABORDAGEM POR AGRUPAMENTO COM K-MEANS NOS DADOS DA PENSE E IDEB

Dissertação apresentada ao Programa de Pós-Graduação em Computação do Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Mestre em Ciência da Computação.

Orientador: Prof. Dr. Tiago Thompsen Primo

Pelotas, 2025

Universidade Federal de Pelotas / Sistema de Bibliotecas
Catalogação da Publicação

C176a Camboim, Guilherme de Barros

Análise de fatores associados ao desempenho educacional no Brasil [recurso eletrônico] : uma abordagem por agrupamento com K-means nos dados da PeNSE e IDEB / Guilherme de Barros Camboim ; Tiago Thompsen Primo, orientador. — Pelotas, 2025.
87 f. : il.

Dissertação (Mestrado) — Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, 2025.

1. Aprendizagem de máquina. 2. K-means. 3. Educação. 4. PeNSE. I. Primo, Tiago Thompsen, orient. II. Título.

CDD 005

Elaborada por Simone Godinho Maisonave CRB: 10/1733

Guilherme de Barros Camboim

ANÁLISE DE FATORES ASSOCIADOS AO DESEMPENHO EDUCACIONAL NO BRASIL: UMA ABORDAGEM POR AGRUPAMENTO COM K-MEANS NOS DADOS DA PENSE E IDEB

Dissertação aprovada, como requisito parcial, para obtenção do grau de Mestre em Ciência da Computação, Programa de Pós-Graduação em Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 19 de setembro de 2025

Banca Examinadora:

Prof. Dr. Tiago Thompsen Primo (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Profa. Dra. Patricia Augustin Jaques Maillard
Doutora em Computação pela Universidade Federal do Rio Grande do Sul.

Profa. Dra. Ana Marilza Pernas Fleischmann
Doutora em Computação pela Universidade Federal do Rio Grande do Sul.

Dedico este trabalho ao meu pai, Adilson.

AGRADECIMENTOS

A construção deste trabalho contou com a colaboração de diversas pessoas que contribuíram, direta ou indiretamente, para sua realização. À minha esposa, Danielle Souza, agradeço pelo incentivo, paciência e companheirismo. Ao meu orientador, Tiago Thompsen Primo, sou profundamente grato pela presença, paciência, dedicação, disposição e, sobretudo, pela orientação na indicação de caminhos e no estímulo à germinação de ideias.

À Universidade — minha segunda casa nesta jornada — e ao dedicado e solícito corpo docente, que tanto contribuíram para a minha formação, deixo registrado meu sincero agradecimento.

Por fim, a todos aqueles que a vida colocou em meu caminho e que, de alguma forma, me ajudaram a desenvolver qualidades essenciais para que eu chegasse até aqui, expresso minha mais profunda gratidão.

Gosto de ser gente porque, inacabado, sei que sou um ser condicionado mas, consciente do inacabamento, sei que posso ir mais além dele.
— PAULO FREIRE

RESUMO

CAMBOIM, Guilherme de Barros. **ANÁLISE DE FATORES ASSOCIADOS AO DESEMPENHO EDUCACIONAL NO BRASIL: UMA ABORDAGEM POR AGRUPAMENTO COM K-MEANS NOS DADOS DA PENSE E IDEB**. Orientador: Tiago Thompsen Primo. 2025. 87 f. Dissertação (Mestrado em Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2025.

Os resultados do PISA 2022 evidenciaram desafios significativos para a educação brasileira, com a maioria dos estudantes não atingindo os níveis mínimos de proficiência em matemática, leitura e ciências — situação também refletida nos indicadores do IDEB. Diversos fatores, como condições habitacionais, emocionais, sociais e a formação docente, contribuem para esse cenário. Apesar da relevância dessas avaliações, ainda há pouco uso de técnicas de clusterização aplicadas a bases nacionais como subsídio à formulação de políticas públicas. Este estudo buscou identificar padrões de comportamentos estudantis, a partir da Pesquisa Nacional de Saúde do Escolar (PeNSE), que se relacionam com diferentes trajetórias de desempenho no IDEB. Aplicou-se o algoritmo K-Means aos microdados da PeNSE de 2009, 2015 e 2019, para identificar padrões comportamentais e relacioná-los aos desempenhos educacionais. Foram considerados estudantes do 9º ano e do 3º ano, organizados por unidade federativa e esfera administrativa, com dados de inquérito amostral baseados em autorrelato. O pré-processamento incluiu limpeza de inconsistências, imputação de valores faltantes e padronização das variáveis. O número de clusters foi definido por métricas como cotovelo, silhouette, Calinski–Harabasz e Davies–Bouldin. Identificaram-se entre três e quatro clusters por edição, e os resultados mostraram que baixa supervisão parental, alimentação inadequada, sedentarismo e uso precoce de álcool e drogas estiveram associados aos piores resultados no IDEB, sobretudo nas escolas públicas. Apesar das limitações — como a heterogeneidade dos ciclos da PeNSE e a ausência de vínculo direto entre comportamento e desempenho —, o estudo revela padrões recorrentes e fornece subsídios relevantes para políticas públicas mais eficientes. Fatores como supervisão parental, prática de exercícios e alimentação adequada emergem como prioritários em futuras intervenções.

Palavras-chave: aprendizagem de máquina; k-means; educação; PeNSE.

ABSTRACT

CAMBOIM, Guilherme de Barros. **PATTERN ANALYSIS IN BRAZILIAN EDUCATIONAL PERFORMANCE: AN UNSUPERVISED LEARNING APPROACH USING DATA FROM THE PENSE SURVEY**. Advisor: Tiago Thompsen Primo. 2025. 87 f. Dissertation (Masters in Computer Science) – Technology Development Center, Federal University of Pelotas, Pelotas, 2025.

The 2022 PISA results revealed significant challenges for Brazilian education, with most students failing to reach minimum proficiency levels in mathematics, reading, and science — a situation also reflected in IDEB indicators. Various factors, such as housing conditions, emotional and social aspects, and teacher training, contribute to this scenario. Despite the relevance of these assessments, there is still limited use of clustering techniques applied to national datasets to support educational policy development. This study aimed to identify patterns of student behavior, based on data from the National School Health Survey (PeNSE), that relate to different educational performance trajectories in IDEB. The K-Means clustering algorithm was applied to PeNSE microdata from 2009, 2015, and 2019 to identify behavioral patterns and associate them with educational outcomes. Students from the 9th grade of elementary school and the 3rd year of high school were analyzed, organized by federal unit and administrative sphere, using self-reported survey data. Preprocessing included cleaning inconsistencies, imputing missing values, and standardizing variables. The number of clusters was determined using metrics such as elbow, silhouette, Calinski–Harabasz, and Davies–Bouldin. Between three and four clusters were identified per edition, and the results showed that low parental supervision, poor diet, sedentary behavior, and early exposure to alcohol and drugs were associated with lower IDEB performance, especially in public schools. Despite limitations — such as the heterogeneity of PeNSE cycles and the absence of a direct link between behavior and performance — the study reveals recurring patterns and provides valuable insights for more effective public policies. Factors such as parental supervision, physical activity, and healthy eating emerge as priorities for future interventions.

Keywords: machine learning; k-means; education; PeNSE.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACM	*Association for Computing Machinery*
ANOVA	Análise de Variância
IBGE	Instituto Brasileiro de Geografia e Estatística
IDEB	Índice de Desenvolvimento da Educação Básica
IDH	Índice de Desenvolvimento Humano
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais
IRQ	Intervalo entre Quartis
LGPD	Lei Geral de Proteção de Dados Pessoais
MEC	Ministério da Educação
OCDE	Organização para a Cooperação e Desenvolvimento Econômico
PeNSE	Pesquisa Nacional de Saúde do Escolar
PIB	Produto Interno Bruto
PLANEA	*Plan Nacional para la Evaluación de los Aprendizajes*
PISA	Programa Internacional de Avaliação de Alunos
SAEB	Sistema de Avaliação da Educação Básica
SOL	*SBC Open Lib*
t-SNE	*t-Distributed Stochastic Neighbor Embedding*
UMAP	*Uniform Manifold Approximation and Projection for Dimension Reduction*
UF	Unidade Federativa
UNICEF	Fundo das Nações Unidas para a Infância
WCSS	*Within-Cluster Sum of Squares*

SUMÁRIO

1 INTRODUÇÃO

Relatórios recentes da Organização para a Cooperação e Desenvolvimento Económico (OCDE) e do Instituto Nacional de Estudos e Pesquisas Educacionais (INEP) mostram dificuldades persistentes de proficiência e grande heterogeneidade entre unidades federativas (UFs) e redes de ensino (Souza; Chagas; Cassia et al., 2019). Evidências nacionais também indicam que comportamentos de risco, condições socioeconômicas e características do contexto escolar e familiar se relacionam às diferenças de aprendizado (Alves, 2023).

No contexto brasileiro, um relatório do Fundo das Nações Unidas para a Infância (UNICEF) de 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 (UNICEF, 2023). Essas crianças enfrentam restrições significativas relacionadas à renda, à segurança alimentar, ao acesso à informação e à educação, além da exposição precoce ao trabalho infantil, da precariedade habitacional e da ausência de infraestrutura básica, como abastecimento de água potável e saneamento adequado. O relatório também mostrou disparidades regionais marcantes: enquanto menos de 40% das crianças residentes em São Paulo e no Distrito Federal sofrem algum tipo de privação, esse percentual ultrapassa 90% em estados como Rondônia e Amapá.

Em consonância com esse diagnóstico estrutural, os resultados do Programa para Avaliação Internacional de Estudantes (PISA) 2022 revelaram que 73% dos estudantes brasileiros não atingiram o nível mínimo de proficiência em matemática, enquanto 50% apresentaram desempenho insatisfatório em leitura e 55% em ciências — todos significativamente abaixo das médias da OCDE (Ocde, 2022). Tais resultados refletem a influência de múltiplos determinantes sociais, econômicos e institucionais. Conforme argumenta Gomes (2018), condições habitacionais e sanitárias inadequadas, vulnerabilidades socioeconômicas, ambiente familiar desestruturado, fatores emocionais, bem como limitações na formação e valorização docente, constituem um ecossistema adverso à aprendizagem.

Estudos recentes corroboram essa perspectiva multifatorial. Alves Filho (2024) destaca que uma alimentação equilibrada e a prática regular de atividades físicas não apenas contribuem para a saúde geral, mas também são determinantes diretos no desenvolvimento cognitivo. Já Soares; Bernardo Junior (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. Esses achados reforçam a complexidade do problema e desafiam abordagens reducionistas que desconsideram o entrelaçamento de dimensões sociais, emocionais e comportamentais no processo de aprendizagem.

Apesar do volume de estudos sobre determinantes do desempenho escolar, faltam investigações que integrem a Pesquisa Nacional de Saúde do Escolar (PeNSE) e o Índice de Desenvolvimento da Educação Básica (IDEB) em um mesmo arcabouço analítico, com vistas a identificar perfis multivariados via algoritmos de clusterização e relacioná-los a níveis de desempenho em escala nacional.

Considerando isso, neste estudo, formulou-se a seguinte pergunta de pesquisa: Quais padrões de fatores sociais, econômicos, comportamentais e escolares podem ser identificados nos microdados da PeNSE por meio do algoritmo K-Means, e como esses padrões se relacionam aos níveis do IDEB nas unidades da federação?

Para responder a essa questão, o *pipeline* metodológico compreendeu: limpeza e imputação de dados; padronização de variáveis; definição do número de grupos a partir de métricas internas; e aplicação do algoritmo K-Means aos microdados da PeNSE — coletados em 2009, 2015 e 2019 —, correlacionando-se os agrupamentos identificados com os resultados do IDEB nos mesmos períodos. O desempenho educacional foi definido a partir dos indicadores do IDEB nas etapas avaliadas (9º ano do ensino fundamental e 3º ano do ensino médio). Os fatores analisados corresponderam aos domínios observados na PeNSE, como ambiente familiar, hábitos de saúde, comportamentos de risco e variáveis de contexto escolar, excluindo-se da análise questões relativas à estrutura escolar e focando apenas nos aspectos relacionados aos estudantes. A unidade de análise considerou a unidade federativa (UF) e a esfera administrativa (pública ou privada).

Ressalta-se que não foi realizado pareamento individual entre respondentes da PeNSE e indicadores do IDEB, tampouco inferência causal. O objetivo central foi mapear fatores associados ao desempenho educacional e analisar sua variação temporal, de modo a evidenciar padrões recorrentes que se associam positiva ou negativamente à educação brasileira ao longo do tempo, fornecendo subsídios para o aprimoramento das políticas públicas, embora se reconheça a defasagem temporal entre os diagnósticos internacionais mais recentes e a janela dos dados efetivamente utilizados.

À luz desses objetivos, a presente dissertação organiza-se da seguinte forma: o

Capítulo 1 apresentou a contextualização do problema, a justificativa e a formulação da questão de pesquisa e ainda neste capítulo, apresentam-se os objetivos do estudo; o Capítulo 2 aborda a fundamentação teórica, destacando os principais conceitos relacionados à PeNSE e ao algoritmo K-Means; o Capítulo 3 descreve detalhadamente a metodologia utilizada, explicitando cada etapa do processo analítico; o Capítulo 4 apresenta e discute os resultados obtidos; e, por fim, o Capítulo 5 reúne as conclusões do estudo, apontando suas contribuições e sugerindo direções para futuras investigações. Dessa forma, encerra-se este capítulo introdutório, estabelecendo as bases para o desenvolvimento da análise que se seguirá.

1.1 Objetivos

Este estudo propõe uma abordagem exploratória baseada em técnicas de aprendizado de máquina para identificar padrões ocultos nos dados da PeNSE. A seguir, apresentam-se o objetivo geral e os objetivos específicos que norteiam este trabalho.

1.1.1 Objetivo Geral

Identificar, categorizar e caracterizar grupos com base em características sociodemográficas, utilizando métodos de agrupamento nos microdados das PeNSEs de 2009, 2015 e 2019, e relacioná-los ao IDEB por unidade federativa e esfera administrativa.

1.1.2 Objetivos Específicos

1. Identificar perfis de estudantes com base em suas características sociodemográficas aplicando técnicas de clusterização;

2. Realizar a associação dos perfis encontrados aos resultados do IDEB, pareando ecologicamente por ano, unidade federativa e esfera administrativa;

3. Discutir os padrões encontrados e obter insights com base nas análises;

4. Disponibilizar os insights e discussões realizadas, ainda que de maneira exploratória e sem relação causal, como insumos para estudos que possam servir de subsídio para a formulação de políticas públicas.

2 FUNDAMENTAÇÃO TEÓRICA

Esta seção apresenta os fundamentos teóricos que sustentam a pesquisa. Inicialmente, definem-se os principais construtos operacionais utilizados no estudo. O desempenho educacional é avaliado por meio dos indicadores do IDEB, considerando o 9º ano do ensino fundamental e o 3º ano do ensino médio. Os fatores correspondem aos domínios observados na PeNSE, abrangendo comportamentos de risco, condições socioeconômicas e características do contexto escolar e familiar, entre outros. A clusterização é entendida como uma técnica de análise de dados que agrupa elementos semelhantes de forma automática; neste estudo, o algoritmo K-Means é utilizado para identificar agrupamentos de estudantes com perfis comportamentais e contextuais similares, permitindo analisar padrões de fatores que se associam a diferentes níveis de desempenho educacional. Posteriormente, apresentam-se os conceitos relacionados à correlação entre variáveis e articula-se a relação entre os pontos abordados, de modo a oferecer uma visão integrada da proposta do estudo. Por fim, discute-se o estado atual da arte.

2.1 Fatores que Influenciam o Desempenho Escolar

A aprendizagem é 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). Diante da diversidade de influências sobre a aprendizagem, este estudo limita-se à análise dos fatores que dizem respeito diretamente aos alunos, conforme apontado por Gomes (2018), entre os quais se destacam:

(i) **Condições Habitacionais e Sanitárias:** fatores como superlotação na residência e falta de condições adequadas de higiene podem interferir na concentração e no desempenho escolar dos alunos.

(ii) **Fatores Afetivos e Psicológicos:** esses fatores mostram que a educação não é apenas uma questão de conteúdo acadêmico, mas também envolve aspectos

emocionais e psicológicos que são fundamentais para o sucesso dos alunos, tais como relação professor-aluno, autoestima e motivação, entre outros.

(iii) **Fatores Sociais e Econômicos:** a situação financeira das famílias pode limitar o acesso a recursos essenciais, como alimentação adequada, vestuário e condições de vida, impactando diretamente o rendimento escolar.

(iv) **Ambiente Familiar:** comportamentos inadequados por parte dos responsáveis, como violência doméstica, desemprego e desestruturação familiar, podem afetar o comportamento e a motivação dos alunos, contribuindo para a evasão e a repetência escolar.

(v) **Fatores Físicos e Mentais:** dificuldades de locomoção e necessidades específicas não atendidas nas escolas podem limitar a aprendizagem, especialmente quando as instituições não estão preparadas para receber alunos com essas condições.

Esses elementos, em conjunto, criam um ambiente que pode influenciar significativamente o processo de ensino-aprendizagem. Nesse contexto, constata-se que a aprendizagem ocorre muito além dos limites da sala de aula. Considerando essa complexidade, torna-se relevante investigar como diferentes fatores — sociais, econômicos, comportamentais e escolares — se organizam em padrões específicos entre os estudantes, servindo de base para a análise dos determinantes do desempenho educacional.

2.2 Métricas Educacionais no Brasil

Nesse sentido, a Pesquisa Nacional de Saúde do Escolar (PeNSE) — um levantamento amostral que, desde 2009, coleta dados sobre saúde, comportamentos de risco, contexto social e condições socioeducacionais de crianças e adolescentes — fornece informações valiosas para a compreensão dessa variável complexa: a educação. A PeNSE possui periodicidade eventual e abrangência nacional (IBGE, 2009). Dada a riqueza e a multidimensionalidade de seu escopo, a pesquisa vem sendo amplamente utilizada em diferentes áreas, como mostra Mello; Silva; Oliveira; Prado; Malta; Silva (2017), que analisou a prática de bullying entre escolares brasileiros e fatores associados. Soares; Leão; Freitas; Hallal; Wagner (2023) sintetizou as tendências temporais da atividade física em adolescentes. Já, Reis; Malta; Furtado (2018) demonstrou os desafios enfrentados pelas políticas públicas voltadas à adolescência e juventude.

Apesar de sua relevância, é fundamental reconhecer algumas limitações metodológicas da PeNSE, como a natureza autorrelatada das informações — que pode

introduzir vieses de memória ou de desejabilidade social — e as variações amostrais significativas entre os estados, que comprometem a comparabilidade e a generalização dos resultados.

Considerando as limitações e potencialidades da PeNSE, a integração com indicadores educacionais, como o IDEB, possibilita uma análise associativa exploratória mais robusta. O Índice de Desenvolvimento da Educação Básica (IDEB), criado em 2007, é um indicador sintético que combina dois componentes fundamentais da qualidade educacional: o aprendizado escolar — medido pelas proficiências em Língua Portuguesa e Matemática no SAEB (aplicado no último ano do ensino fundamental e no ensino médio) — e o fluxo escolar, calculado a partir das taxas de aprovação obtidas no Censo Escolar. Essa métrica permite o monitoramento integrado da eficácia dos sistemas de ensino, articulando os resultados de aprendizagem com a progressão adequada dos estudantes (INEP, 2025). O indicador apresenta resultados das redes pública e privada nos níveis nacional, estadual, municipal e por escola.

Vale ressaltar que, embora amplamente utilizado em políticas públicas e estudos educacionais, esse indicador da educação brasileira possui limitações que devem ser consideradas. Conforme demonstrado por Alves; Soares (2013), 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. Neste trabalho, optou-se por estabelecer uma relação entre a PeNSE e o IDEB a partir da unidade federativa, uma vez que a PeNSE não informa o nome da escola ou do município analisado, disponibilizando apenas a região e a unidade federativa como referência para essa vinculação.

2.3 Aprendizagem de Máquina

Diante da sua complexidade e do volume de informações disponíveis, recorreu-se ao Aprendizado de Máquina para explorar padrões ocultos e 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; Ali; Anam; Ahmed, 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; Ali; Anam; Ahmed, 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, sendo outras alternativas comuns o DBSCAN, entre outros (Hernández-Leal; Duque-Méndez; Cechinel, 2021).

2.3.1 Algoritmos de Clusterização

Segundo Jain; Dubes (1988), a clusterização, ou análise de agrupamentos, consiste em uma técnica de aprendizado não supervisionado cujo objetivo é identificar estruturas e padrões ocultos em conjuntos de dados, organizando-os em grupos (*clusters*) de acordo com o grau de similaridade entre as observações. Trata-se de um método particularmente adequado para revelar relações latentes e segmentações naturais nos dados.

Na literatura, diversos algoritmos de clusterização têm sido amplamente empregados. Entre os mais tradicionais, destacam-se os métodos particionais, como o *K-Means* — adotado neste trabalho —, que divide o conjunto em k grupos a partir da minimização da distância entre os pontos e seus centróides (Macqueen, 1967; Lloyd, 1982). O *K-Medoids*, por sua vez, constitui uma variação mais robusta, pois utiliza elementos reais do conjunto como representantes de cada cluster, reduzindo a sensibilidade a outliers (Kaufman; Rousseeuw, 1990).

Além dos métodos particionais, destacam-se os algoritmos hierárquicos, que constroem uma estrutura de agrupamento em forma de dendrograma e permitem a análise em diferentes níveis de granularidade (Johnson, 1967). Outra categoria relevante são os métodos baseados em densidade, como o *HDBSCAN*, ou probabilísticos, como o *GMM*, capazes de identificar agrupamentos de acordo com a densidade ou distribuição dos pontos, lidando inclusive com clusters de formas arbitrárias e com a presença de ruído (Ester; Kriegel; Sander; Xu, 1996).

Conforme Jain (2010), esses algoritmos têm sido amplamente aplicados em áreas diversas, como educação, bioinformática, mineração de textos, marketing e ciências sociais, devido à sua capacidade de organizar grandes volumes de dados em estruturas compreensíveis e úteis para análises subsequentes. Dessa forma, configuram-se como ferramentas adequadas para o presente trabalho.

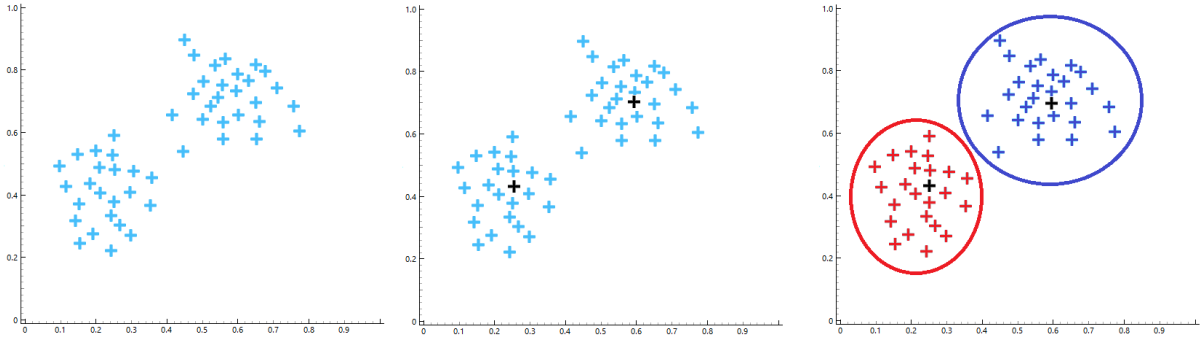

Figura 1 – Representação gráfica da sequência de funcionamento do K-Means.
Fonte: Elaborado pelo autor

2.3.2 Algoritmo K-Means

Dentre os algoritmos de Aprendizado de Máquina não supervisionado, o K-Means é uma técnica amplamente reconhecida. Ele é utilizado sem hipóteses prévias, o que o torna ideal para conjuntos de dados heterogêneos, como os da PeNSE. Seu funcionamento é ilustrado na Figura 1 e pode ser descrito da seguinte forma (Witten; Frank; Hall; Pal; Data, 2016):

1. Particiona N pontos de dados em K grupos (clusters);
2. Atribui cada ponto ao cluster cujo centroide está mais próximo (distância euclidiana);
3. Atualiza iterativamente os centroides (média dos pontos no cluster);
4. Repete até convergência (minimizando a soma das distâncias quadradas intra-cluster).

Dessa forma, dados semelhantes são agrupados automaticamente, sem a necessidade de rótulos prévios. Vale ressaltar que, assim como qualquer algoritmo, o K-Means apresenta limitações, como a necessidade de definição prévia do número de clusters e a menor eficiência na identificação de grupos não esféricos. Por essa razão, recorrem-se a métodos para determinar a quantidade ideal de agrupamentos, sendo o principal — e mais utilizado para essa finalidade — o Método do Cotovelo (Lenz; Neuman; Santarelli; Salvador, 2020).

Primeiro, o Método do Cotovelo avalia o melhor valor de k utilizando três indicadores, entre eles o WCSS, que mede o quão próximos os pontos de cada grupo estão do seu centróide. O ponto ideal é aquele em que aumentar o número de grupos deixa de proporcionar uma melhora significativa nessa proximidade (Cui et al., 2020). Em seguida, o coeficiente de Silhouette mede a coesão e a separação dos clusters, sendo que valores próximos a 1 indicam uma boa definição dos agrupamentos (Furlanetto;

Carvalho; Baldassin; Manacero, 2022). Por outro lado, o índice Davies-Bouldin avalia a compacidade dos clusters e a separação entre eles, de modo que valores menores indicam melhor desempenho (Furlanetto; Carvalho; Baldassin; Manacero, 2022). Ademais, o índice Calinski-Harabasz compara a dispersão intra-cluster com a dispersão inter-cluster, em que valores mais altos sugerem melhor qualidade dos clusters (Furlanetto; Carvalho; Baldassin; Manacero, 2022). Por fim, a combinação desses métodos permite identificar de maneira robusta a quantidade de clusters que melhor representa a estrutura dos dados.

No contexto educacional, o uso do K-Means tem ganhado destaque em várias áreas, como *learning analytics*. Por exemplo, Torcate; Barbosa; Oliveira rodrigues (2020a) utilizaram o K-Means para analisar dificuldades de aprendizagem na educação básica, identificando as principais dificuldades dos grupos de alunos em matemática.

Cientes das limitações, aplicações e potencialidades do algoritmo K-Means, sua escolha neste estudo justifica-se tanto pela aplicabilidade na segmentação de dados educacionais quanto pela necessidade de capturar a heterogeneidade dos perfis estudantis e suas interações com fatores de saúde, comportamento e contexto socioeconômico. Embora o K-Means tenha sido originalmente desenvolvido para variáveis numéricas contínuas e apresente restrições na análise de dados mistos, sua adoção é sustentada pelo caráter exploratório deste trabalho e por ser um método amplamente consolidado e bem compreendido na literatura, o que favorece a interpretação e a reprodutibilidade dos resultados. Reconhece-se, contudo, que a presença de variáveis categóricas pode reduzir a precisão dos agrupamentos, razão pela qual os resultados devem ser interpretados com cautela. Ainda assim, a utilização do K-Means nesta etapa busca fornecer uma base comparativa robusta para análises posteriores, que poderão empregar algoritmos mais específicos, como o K-Prototypes, K-Medoids ou DBSCAN, mediante ajustes apropriados aos dados.

2.3.3 Técnicas de visualização

O *t-Distributed Stochastic Neighbor Embedding* (t-SNE) é uma técnica recente desenvolvida para a visualização de dados em alta dimensão. Seu objetivo é atribuir a cada ponto dos dados uma posição em um mapa bidimensional ou tridimensional. Em essência, o t-SNE é uma variação do *Stochastic Neighbor Embedding* (SNE) que busca criar uma projeção capaz de preservar a estrutura significativa dos dados originais. Essa técnica destaca-se por gerar representações que revelam padrões em múltiplas escalas, capturando tanto a estrutura local quanto a global — como a presença de agrupamentos em diferentes níveis. (Maaten; Hinton, 2008)

O *Uniform Manifold Approximation and Projection for Dimension Reduction* (UMAP), por sua vez, visa encontrar uma representação de baixa dimensionalidade

dos dados a partir da preservação das vizinhanças locais. O método identifica os vizinhos mais próximos de cada ponto e os representa por meio de um grafo ponderado. Ao priorizar a preservação da topologia das vizinhanças em vez das distâncias absolutas, o UMAP tende a agrupar indivíduos mais relacionados. Entre suas vantagens estão a maior escalabilidade em relação ao t-SNE, a capacidade de preservar tanto estruturas locais quanto globais e a frequência com que produz separações de clusters mais nítidas. No entanto, o método pode ser sensível à escolha de hiperparâmetros, exigindo ajustes cuidadosos. (Diaz-papkovich; Anderson-trocmé; Gravel, 2021)

2.4 Análise estatística correlacional

De posse dos perfis obtidos a partir da clusterização, a análise estatística correlacional — técnica utilizada para avaliar a relação entre duas ou mais variáveis, determinando se elas se movem juntas e qual a intensidade dessa relação (embora correlação não implique causalidade (Sampaio; Bosco, 2015)) — pode auxiliar na interpretação dos perfis encontrados. Dentre as técnicas para avaliação da correlação entre duas variáveis quantitativas, a mais difundida é o coeficiente de correlação de Pearson, que pressupõe a normalidade dos dados. Por outro lado, coeficientes como Spearman e Kendall não fazem essa pressuposição. Por essa razão, é essencial verificar a normalidade dos dados antes da escolha do método de correlação (Miot, 2018).

2.4.1 Normalidade dos dados

A normalidade dos dados refere-se à condição em que a distribuição amostral segue o modelo da distribuição normal ou gaussiana. Essa distribuição caracteriza-se por ser simétrica, em formato de sino, com média, mediana e moda coincidentes, além de curtose próxima ao valor teoricamente esperado. A verificação da normalidade é fundamental, pois a validade de diversos procedimentos estatísticos depende dessa suposição. A avaliação pode ser realizada por meio de testes formais, como Shapiro-Wilk, Kolmogorov-Smirnov e Lilliefors, ou ainda pela análise dos coeficientes de assimetria e curtose, comparando os valores observados com aqueles esperados em uma distribuição normal (Hatem; Zeidan; Goossens; Moreira, 2022). Cada abordagem apresenta vantagens e limitações; neste estudo, optou-se pela utilização dos coeficientes de Assimetria (i) e Curtose (ii) devido ao tamanho das amostras, o que os torna mais adequados para esse cenário.

(i) Assimetria (Skewness): Conforme Santos; Ferreira (2003), a assimetria corresponde a uma medida da forma da distribuição dos dados, indicando o grau e a direção de afastamento em relação à simetria. Distribuições simétricas apresentam valores igualmente distribuídos em torno da média, enquanto distribuições assimétricas possuem caudas mais longas em um dos lados. Valores positivos

de assimetria indicam maior alongamento da cauda direita, enquanto valores negativos indicam maior alongamento da cauda esquerda. Em termos práticos, ao traçar uma curva semelhante à distribuição normal a partir da média, a medida de skewness expressa o quanto os dados se afastam dessa distribuição.

(ii) Curtose (Kurtosis): Conforme Santos; Ferreira (2003), a curtose expressa o grau de concentração dos dados em torno da média e nas caudas da distribuição. Distribuições com curtose elevada (leptocúrticas) apresentam caudas mais pesadas e maior concentração central em comparação à distribuição normal. Em contraste, distribuições com baixa curtose (platicúrticas) apresentam caudas mais leves e formato mais achatado. A distribuição normal, considerada mesocúrtica, é tomada como referência, com curtose teórica igual a três. Em termos práticos, quanto maior a curtose, maior a presença de outliers ou a tendência da distribuição em produzi-los.

2.4.2 Métodos de correlação

A partir da definição da normalidade dos dados, aplicam-se os métodos de correlação adequados à natureza dos dados. Primeiramente, o coeficiente de correlação de Pearson é uma medida paramétrica que avalia a força e a direção da relação linear entre duas variáveis quantitativas contínuas. Seus valores variam entre -1 e $+1$, em que $+1$ indica correlação linear perfeita e 0 ausência de relação. O uso de Pearson é recomendado quando as variáveis apresentam distribuição normal ou próxima do normal, sendo calculado a partir da covariância, normalizada pelo produto dos desvios-padrão (Shaqiri; Iljazi; Kamberi; Ramani-halili, 2023).

Em seguida, o coeficiente de correlação de Spearman constitui uma medida não paramétrica que avalia a associação entre duas variáveis com base na ordenação (*ranks*) dos valores, verificando o quanto a relação pode ser descrita por uma função monotônica. Também variando entre -1 e $+1$, o Spearman é indicado quando os dados não seguem distribuição normal ou quando se suspeita de relações monotônicas não lineares. Por ser o análogo não paramétrico de Pearson, seus valores tendem a ser próximos aos deste (Shaqiri; Iljazi; Kamberi; Ramani-halili, 2023).

Por outro lado, o coeficiente de Kendall (Tau de Kendall) é igualmente uma medida não paramétrica que mensura a associação entre duas variáveis por meio da proporção de pares de observações concordantes e discordantes. Ele pode ser interpretado como a quantidade mínima de trocas necessárias para transformar uma ordenação na outra. Seu uso é recomendado para dados ordinais ou quando se busca uma medida mais robusta a outliers e pequenas variações na ordem. Entretanto, em comparação com Pearson e Spearman, Kendall tende a apresentar valores mais conservadores para o mesmo conjunto de dados (Shaqiri; Iljazi; Kamberi; Ramani-halili, 2023).

Vale destacar que, conforme Cohen (2013), correlações em torno de 10% são consideradas de magnitude pequena, em torno de 30% de magnitude média e a partir de 50% de magnitude elevada. Nesse sentido, valores a partir de aproximadamente 25% podem ser interpretados como correlações de efeito médio, alinhando-se à literatura estatística.

2.5 Articulação entre blocos

A articulação entre os três blocos — PeNSE, IDEB e K-means — sustenta a proposta central deste estudo: explorar, por meio de técnicas de agrupamento, a relação entre fatores contextuais e comportamentais de crianças e adolescentes brasileiros e o desempenho educacional em diferentes estados do país. Ao agrupar estados com perfis similares de estudantes com base nos dados da PeNSE e correlacionar esses grupos com variações no IDEB ao longo do tempo, o estudo visa compreender como condições de saúde, comportamento e contexto social podem se associar a trajetórias educacionais diferenciadas. Essa estratégia metodológica, embora limitada por questões de agregação e pela ausência de vínculo individual entre as bases (o que impede estabelecer uma relação de causalidade), permite lançar luz sobre padrões sistêmicos e recorrentes, oferecendo insights relevantes para o desenho de políticas públicas mais sensíveis à diversidade do cenário educacional brasileiro.

2.6 Estado da Arte

O estudo sobre o estado da arte foi conduzido por meio de uma revisão sistemática, realizada a partir de buscas com palavras-chave combinadas a operadores booleanos (AND/OR) nas seguintes bases de dados: IEEE Xplore, Association for Computing Machinery (ACM), ScienceDirect e SBC Open Lib (SOL). A Tabela 1 detalha os termos de busca específicos utilizados em cada base de dados. Esta estratégia de busca resultou na identificação de 208 trabalhos relevantes.

Dentre os trabalhos encontrados, o trabalho do Torcate; Barbosa; Oliveira rodrigues (2020b) é um relato de experiência focado na Educação Básica, que utiliza *Learning Analytics* e Mineração de Dados Educacionais em conjunto com a estratégia de aprendizagem não supervisionada K-Means. O objetivo principal é identificar as dificuldades de aprendizagem em oito conteúdos de matemática do 6º ano. A coleta de dados foi realizada através de jogos digitais, e a análise, utilizando o K-Means, agrupou os alunos por desempenho, evidenciando os conteúdos mais problemáticos e fornecendo feedback concreto para intervenções pedagógicas.

Assim como Maia; Andrade; Fernandes (2021) que implementou o algoritmo de aprendizado não supervisionado K-means nos microdados do Enem 2018 para agru-

par candidatos ao ensino superior com base em suas notas de proficiência nas cinco áreas. O K-means dividiu os candidatos em dois clusters, sendo o Cluster 1 classificado como o de menor desempenho. A análise posterior revelou que o Cluster 1 possuía percentuais mais elevados de candidatos que satisfaziam os critérios para adesão total às cotas (oriundos de escola pública, baixa renda e autodeclarados pretos, pardos ou indígenas), demonstrando a relação entre notas e desigualdade socioeconômica. (Maia; Andrade; Fernandes, 2021)

Além desses, o trabalho mais alinhado com os objetivos desta pesquisa foi o artigo intitulado "*Academic Achievement in Mathematics of Higher-Middle Education Students in Veracruz: An Approach Based on Computational Intelligence*"(Céspedes-gonzález; Escobar; Jiménez; Molero-castillo, 2023). Este estudo analisou o desempenho em matemática de estudantes do ensino médio e superior utilizando dados da avaliação PLANEA - um exame padronizado mexicano que avalia o desempenho acadêmico nos níveis básico e médio superior - para identificar padrões de similaridade entre estudantes da região de Veracruz, correlacionando essas características com os resultados obtidos nas provas. Com isso, foram obtidos 5 clusters de escolas, revelando que condições socioeconômicas (como marginalização e turno) estão fortemente associadas a resultados acadêmicos, com desigualdades marcantes entre regiões.

No entanto, cabe destacar que o referido artigo limitou-se a analisar apenas 12 variáveis da base PLANEA, não contemplando aspectos comportamentais ou psicossociais dos estudantes, diferindo assim do trabalho proposto.

Tabela 1 – Resultados das Estratégias de Busca nas Bases de Dados

Input	Base	Resutados
("machine learning"OR "unsupervised learning"OR "clustering"OR "K-Means") AND ("education"OR "educational data mining") AND ("school performance"OR "student achievement") AND ("data analysis")	IEEE Xplore	*19*
("machine learning"OR "unsupervised learning"OR "clustering"OR "K-Means") AND ("education"OR "educational data mining") AND ("school performance"OR "student achievement") AND ("data analysis")	ACM	*176*
("machine learning"AND "education") AND ("student achievement"OR "school performance") AND ("unsupervised learning"OR "clustering"OR "K-Means") AND ("socioeconomic factors"OR "educational data analysis")	ScienceDirect	*13*
("machine learning"AND "education") AND ("student achievement"OR "school performance") AND ("unsupervised learning"OR "clustering"OR "K-Means") AND ("socioeconomic factors"OR "educational data analysis")	SOL	*0*

Fonte: Elaborada pelo autor.

3 METODOLOGIA DA ABORDAGEM PROPOSTA

Este estudo caracteriza-se como quantitativo e exploratório-descritivo. Foram utilizados os microdados da PeNSE (2009, 2015 e 2019) e os indicadores do IDEB (2015 e 2019), tendo como unidade de análise as unidades da federação (UF). A metodologia foi estruturada nas etapas ilustradas no fluxograma da Figura 2, com o código-fonte disponível no Apêndice D, e pode ser resumida da seguinte forma: preparo dos dados $\rightarrow$ definição do número de clusters (k) $\rightarrow$ aplicação do algoritmo K-Means $\rightarrow$ avaliação gráfica $\rightarrow$ caracterização dos perfis de cada cluster $\rightarrow$ associação aos resultados do IDEB $\rightarrow$ análises subsequentes. Ressalta-se que todos os processos foram aplicados, inicialmente, apenas à base da PeNSE, sendo o IDEB incorporado posteriormente.

As análises foram conduzidas em linguagem Python, no ambiente Google Colab (Python 3, Google Compute Engine backend), configurado com 12,7 GB de memória RAM e 107,7 GB de armazenamento em disco. Importa destacar que, embora os dados utilizados sejam públicos e anonimizados, foram observadas integralmente as diretrizes da Lei Geral de Proteção de Dados (LGPD). Ressalta-se que não houve qualquer acesso a informações pessoais ou sensíveis, uma vez que se trata de dados secundários disponibilizados por instituições oficiais, utilizados exclusivamente para fins de pesquisa acadêmica.

3.1 Dados e Fontes

Os dados utilizados neste estudo foram obtidos a partir de duas fontes públicas oficiais, descritas no Apêndice E. A primeira delas foi o IBGE, de onde se extraíram os microdados das edições da PeNSE de 2009, 2015 e 2019. A edição de 2009 estava disponível em formato TXT, enquanto as edições posteriores foram disponibilizadas em formatos CSV e XLS. Cabe destacar que a edição de 2015 foi dividida em duas amostras, ambas acessíveis no site do IBGE; neste estudo, optou-se pela utilização da amostra 1, por conter uma maior quantidade de informações. No total, os dados apresentam 187 variáveis em 2009, 293 em 2015 e 306 em 2019, abrangendo informações sobre aspectos comportamentais, socioeconômicos, saúde física e mental,

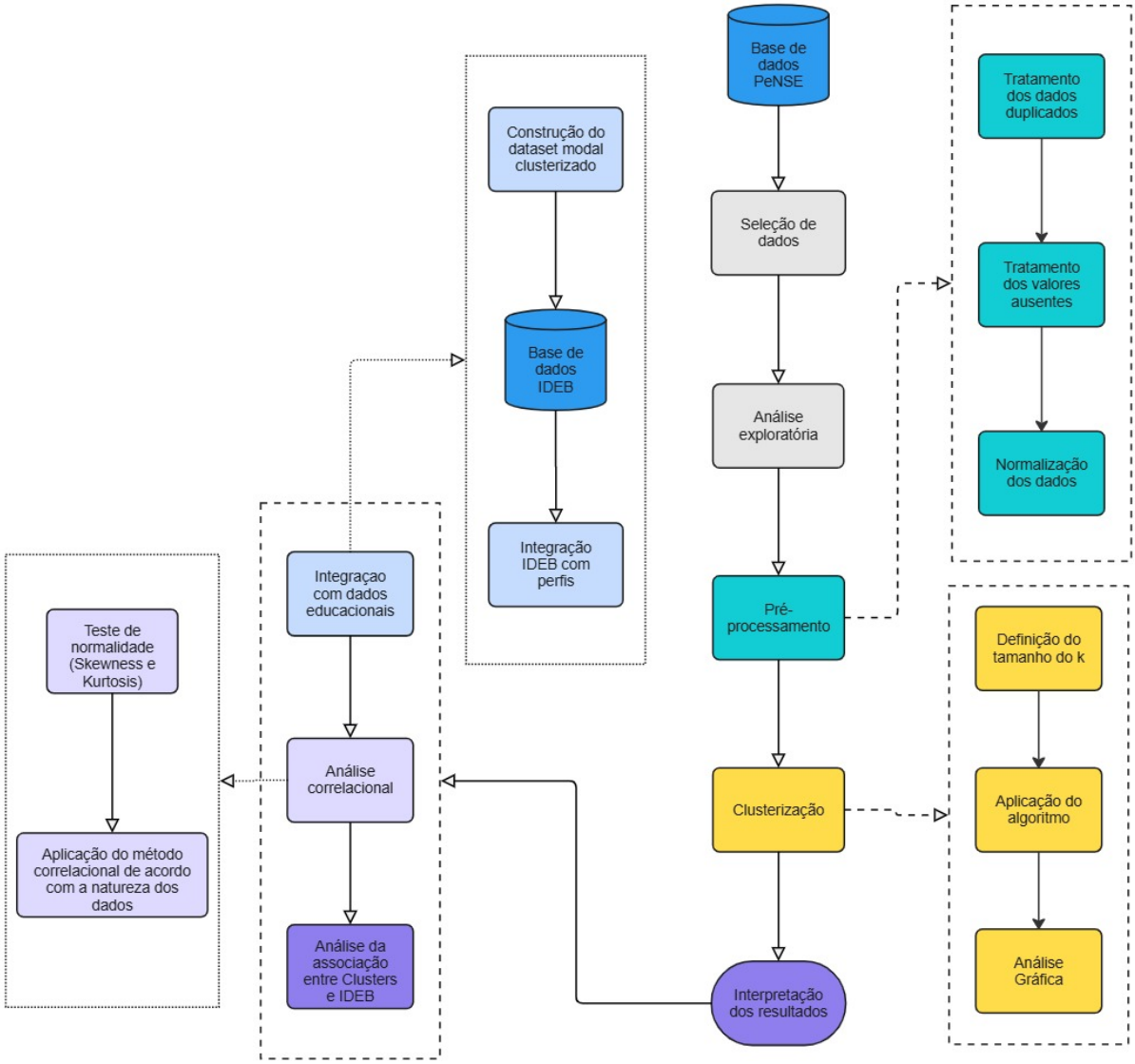

Figura 2 – Fluxograma da metodologia adotada
Fonte: Elaborado pelo autor

ambiente familiar, entre outros. As descrições detalhadas das variáveis podem ser consultadas nos Apêndices A, B e C.

A segunda fonte utilizada foi o INEP/MEC, responsável pela disponibilização dos dados do IDEB referentes aos anos de 2009, 2015 e 2019, de modo a garantir a convergência temporal com as edições da PeNSE. Esses dados foram selecionados por meio de consulta visual no Power BI disponibilizado pela instituição.

3.2 Seleção de dados

O tamanho amostral inicial da PeNSE foi de 63.411 registros em 2009, 102.072 em 2015 e 165.838 em 2019. Consideraram-se apenas os estudantes do 9º ano do Ensino Fundamental e do 3º ano do Ensino Médio de todo o país — com exceção de 2009, que não contempla informações sobre o Ensino Médio. Essa escolha deve-se à convergência de informações com o IDEB, que, conforme explicado na Seção 2.2, abrange os resultados da prova do SAEB, aplicada somente no último ano do ensino fundamental e do ensino médio. Foram excluídas as variáveis de identificação, as de desenho amostral e aquelas julgadas prejudiciais ao processo de clusterização, cuja lista encontra-se apresentada a seguir para (i) 2009, (ii) 2015 e (iii) 2019. Após a aplicação desses critérios, não houve perdas em 2009, enquanto os percentuais de exclusão corresponderam a 1,49% em 2015 e 70,81% em 2019. A seleção final resultou em 104 variáveis para 2009, 164 para 2015 e 188 para 2019. Vale ressaltar que, embora a PeNSE disponibilize pesos amostrais para ajustar a representatividade populacional, estes não foram utilizados nesta análise, pois o foco do estudo é identificar padrões internos entre variáveis e realizar a clusterização dos participantes. Dessa forma, as comparações permanecem válidas dentro da amostra, mas a generalização direta dos resultados para a população nacional é limitada.

(i) 2009: B00P01, B01P03, B01P04, B01P05, CAPITAL, ID, QUESTIONARIO, COD_MUNICIPIO, ESTALOCA, B01P03M, TURMA, TURMAS, MATRIC, FREQ, PESQ, PESQ_CRIT, FREQ_CRIT, PESO_ESCOLA, N_TURMAS, PROBSELTURMA, PESO_TURMA, PART, N_PART, SEXO, PESO_AJU_FREQ, PESO_AJU_SEXO, DEP2, GR_IDADE, INST_MAE, FRUTA, PARENTAL, COME_RESP, RESP_SABE, RESP_FUMO, FALTA_AULA, AGRESSAO, COMEU_FEIJAO, COMEU_FRUTA, COMEU_GULOS, BEBEU_REFRI, TV, PC, EXPCIG, FUMAREG, EXPALC, BEBEREG, EMBRIAGUEZ, EXPDRO, EXPSEX, AULASEF, BULLYING, PRESERV, SEGTRAJ, SEGESC, SENTINSEG, CINTO, DIRIGIU, MOTALCOOL, ESCOVA, DORDENTE, IMAGEM, PRODPESO, FEIJAO, BATATA, SALG, HAMB, LEGU, SALADA, LEGUC, BISCSALG, BISCDOCE, GULO, FRUFRE, LEITE, REFRI, ESCORATI, Q2B01P01, Q2B01P02, Q2B01P01I, Q2B01P02I, DEF_ALT, DEF_PES e EXC_PES.

(ii) 2015: Todas iniciadas com "VE", ANOPESQ, ESTRATOGEOREG, ESTRATO_EXP, PESO,

ALUNO, ESCOLA, TURMA, PAIS, MUNICIPIO_CAP, REGEOGR, TIPO_MUNIC, VB01004, VB0100, VB0102 e VB01021.

(iii) 2019: Todas iniciadas com "E", B00004, B01005, B01003, B01004, REGIAO, MUNICIPIO_CAP, TIPO_MUNIC, SITUACAO, DEP_ADMIN, ESCOLA, TURMA, ALUNO, ANO_TURMA, ESTRATO, IND_EXPANSAO, PESO_ALUNO_FREQ, PESO_INICIAL, POSEST, TOTAIS_POSEST e B01021A.

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 estado e na média dos resultados dos anos finais do Ensino Fundamental e do Ensino Médio. Ressalta-se novamente 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.

3.3 Análise exploratória

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

3.4 Pré-processamento

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

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

2. **Análise de valores ausentes**: Verificação sistemática da presença de valores ausentes em todas as variáveis. Optou-se por não estabelecer limites de exclusão com base na proporção de dados faltantes, a fim de preservar o máximo de informação disponível e evitar o descarte de indicadores potencialmente relevantes. Dessa forma, todos os valores ausentes foram tratados por meio de imputação: para variáveis categóricas numéricas, utilizou-se a moda da coluna, preservando a categoria mais frequente; para variáveis numéricas contínuas, empregou-se a média, garantindo a manutenção da tendência central da distribuição. A escolha desses métodos se justifica pelo baixo percentual de valores

ausentes na maioria das variáveis. O preenchimento foi realizado com o algoritmo *SimpleImputer*, da biblioteca *scikit-learn*.

3. **Normalização de dados**: A normalização foi realizada por meio do algoritmo *Normalizer*, disponível na biblioteca *scikit-learn*[1]. As variáveis normalizadas estão listadas a seguir, de acordo com o período analisado:

   (i) **2009**: `TEMPATIV`;

   (ii) **2015**: `TEMPODESLOC`, `TEMPOEDFIS`, `TEMPOEXTRA`, `TEMPOTOTAL`, `TEMPOEST`;

   (iii) **2019**: `TEMPODESLOC`, `TEMPOEDFIS`, `TEMPOEXTRA`, `TEMPOTOTAL`.

4. **Análise de outliers**: Durante a etapa de verificação de valores extremos em variáveis numéricas contínuas, descritas no item anterior, foram identificados outliers com base no critério do intervalo interquartil (IQR) e no escore robusto de Z. No entanto, optou-se por não removê-los, pois, no contexto desta pesquisa, esses valores podem representar grupos raros, mas relevantes, de estudantes. A exclusão desses casos poderia levar à perda de informações importantes para a formação dos clusters, já que padrões atípicos refletem parte da diversidade comportamental e socioeconômica da população analisada. Dessa forma, os outliers foram mantidos na base, preservando a heterogeneidade dos dados e permitindo que eventuais agrupamentos minoritários fossem detectados pelo algoritmo.

3.5 Clusterização com K-Means

A clusterização com o algoritmo K-Means, aplicada a todas as variáveis de cada período — exceto aquelas previamente removidas, conforme descrito na Seção 3.2 — foi conduzida em duas etapas sequenciais:

1. **Determinação do número ótimo de clusters**: Inicialmente, aplicou-se o algoritmo K-Means aos dados para um intervalo de 2 a 10 clusters, utilizando a inicialização *k-means++* para melhorar a convergência. O valor de k foi escolhido primariamente pelo método do cotovelo (*elbow method*). As demais métricas — coeficiente de silhueta, índice de Calinski-Harabasz e índice de Davies-Bouldin — foram utilizadas para ratificar a escolha e fornecer suporte adicional à definição do número de clusters. Em casos de divergência ou empate entre as métricas, o método do cotovelo foi adotado como critério final de desempate na escolha de k.

[1] Documentação *Normalizer*: `https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.Normalizer.html`

2. **Aplicação do algoritmo K-Means**: O K-Means foi então aplicado utilizando o valor de k definido na etapa anterior, com inicialização *k-means++*, n_init = 10, e limite máximo de 300 iterações. O algoritmo foi configurado para interromper a execução quando não houvesse mudança significativa nos centroides (critério de parada baseado em tolerância padrão do *scikit-learn*). Para assegurar a reprodutibilidade dos resultados obtidos a partir do algoritmo em análise, foram testadas múltiplas *seeds*, sendo a final fixada em 42. Assim, quando aplicados os mesmos dados e parâmetros, o algoritmo produzirá resultados idênticos; contudo, a utilização de *seeds* distintas poderá conduzir a resultados diferentes.

3. **Análise gráfica**: Para fins de visualização e interpretação da separação e da coesão dos grupos, aplicou-se o algoritmo t-SNE para projeção bidimensional (2D), com o objetivo de representar de forma mais fiel as relações de proximidade entre os indivíduos no espaço original de alta dimensionalidade. Em complemento, utilizou-se o algoritmo UMAP para projeção tridimensional (3D), com o intuito de explorar possíveis estruturas não lineares e padrões de agrupamento sob uma perspectiva adicional de visualização.

3.6 Análise dos Resultados

Após a clusterização dos dados, a análise foi conduzida em seis etapas:

1. **Construção do dataset clusterizado**: A partir do dataset bruto, para caracterizar os clusters, optou-se por utilizar a moda de cada variável dentro de cada grupo, em vez de calcular médias ou proporções ponderadas. Essa escolha se justifica pelo caráter predominantemente categórico das variáveis, onde a moda representa de forma mais direta o valor mais frequente e informativo em cada cluster. Além disso, a simplicidade dessa abordagem facilita a interpretação e a comunicação dos resultados.

2. **Integração com Dados Educacionais**: Com base nos grupos obtidos, considerando a unidade federativa, a esfera do agrupamento (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, utilizou-se a média das notas do IDEB das duas etapas para cada grupo. Ressalta-se que, em 2009, não havia dados do Ensino Médio na pesquisa, de modo que foram utilizados apenas os valores do IDEB do Ensino Fundamental para esse ano. Importante ressaltar que essa associação foi feita apenas em nível agregado por UF, esfera e ano, caracterizando-se como ecológica, sem pareamento individual, e portanto não permitindo inferência causal.

3. **Análise correlacional**: Conhecidas a UF e a esfera de cada agrupamento, inferiram-se, a partir do dataset bruto (sem a redução de cada variável ao seu valor de moda) após a clusterização, os valores do IDEB para cada grupo. Em seguida, testou-se a normalidade de cada coluna por meio das medidas de Skewness e Kurtosis. De acordo com a distribuição de cada variável (normal ou não normal), aplicaram-se os métodos de correlação: Pearson para dados normais, e Spearman ou Kendall para dados não normais. Em todas as análises, cada coluna foi relacionada à coluna do IDEB previamente inferida, sendo reportadas apenas as correlações iguais ou superiores a 25%.

4. **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.

5. **Geração do arquivo final**: O arquivo final foi elaborado em formato XLSX, contendo seis abas. A primeira apresenta o dataset descrito no item 1, acrescido dos dados do IDEB mencionados no item 2. A segunda aba contém o dataset com variáveis divergentes, conforme descrito no item anterior. As demais abas reúnem os resultados das correlações de Pearson, Spearman e Kendall e, por fim, o dicionário com o valor de cada resposta de cada variável (Apêndice E), a fim de auxiliar na interpretação final.

6. **Método de interpretação**: Para a geração dos insights apresentados nos resultados, a interpretação foi feita por meio da comparação dos comportamentos do cluster com o pior IDEB e do cluster com o melhor IDEB, desconsiderando-se na análise os clusters com IDEBs intermediários. 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

Os resultados apresentados a seguir referem-se à análise dos microdados da PeNSE (2009, 2015 e 2019), vinculados aos indicadores correspondentes do IDEB. O recorte foi definido a partir das unidades da federação e da esfera administrativa, de modo a possibilitar comparações entre diferentes períodos e contextos educacionais. Ressalta-se que a vinculação entre as bases possui caráter ecológico, sem pareamento individual de estudantes, ou seja, os dados foram analisados em nível agregado e não de forma individualizada. Essa abordagem foi adotada para possibilitar a compreensão da associação entre as variáveis dos estudantes e o desempenho observado no indicador do IDEB.

Com base neste recorte, este capítulo apresenta os principais achados obtidos a partir da metodologia descrita no capítulo anterior. Os resultados estão organizados em diferentes partes, iniciando com os procedimentos de análise exploratória e de pré-processamento, que incluíram a quantificação dos registros, a identificação de valores duplicados e a verificação de *missing values*. Em seguida, são apresentados os resultados da etapa de clusterização, que envolveu a definição do número ideal de grupos e a visualização dos clusters formados em cada período analisado. A análise correlacional aparece na sequência, com a investigação de possíveis relações entre variáveis comportamentais e de desempenho educacional, utilizando diferentes coeficientes de correlação conforme a natureza e a distribuição dos dados.

Na última parte, são discutidos os padrões comportamentais identificados em cada agrupamento, tanto de forma isolada para cada período quanto de maneira longitudinal. Ressalta-se, novamente, que as descrições das variáveis estão disponíveis no Apêndice E, o que pode ser útil para uma melhor compreensão dos achados.

4.1 Análise exploratória e Pré-processamento

Após a seleção dos dados, avaliou-se a quantidade de registros em cada edição, resultando em 63.411 registros em 2009, 100.547 em 2015 e 48.404 em 2019. Na etapa de pré-processamento, foram identificados 377 registros duplicados em 2009,

19 em 2015 e nenhum em 2019. A análise de *missing values* indicou a ocorrência apenas na edição de 2015, totalizando 1.736 casos. Esses valores estavam restritos às variáveis numéricas contínuas, que foram posteriormente normalizadas e estão descritas na Seção 3.4. O tratamento consistiu na imputação pela média das variáveis correspondentes.

4.2 Clusterização

A aplicação do algoritmo K-Means a todas as variáveis de cada período — exceto aquelas descritas na Seção 3.2 —, bem como a definição da quantidade de clusters, conforme apresentado na Seção 3.5, resultou nos valores apresentados na Tabela 2. Essa tabela apresenta, para cada valor de k testado, os resultados das métricas internas de avaliação — *silhouette médio*, Calinski-Harabasz, Davies-Bouldin e o método do cotovelo. Para os anos de 2009 e 2015, observou-se uma tendência para 2 ou 3 clusters, com a maioria das métricas indicando melhor desempenho para k igual a 2. No entanto, o método do cotovelo, amplamente utilizado na literatura e corroborado pela consistência de três indicadores internos, sugeriu k igual a 3, razão pela qual essa configuração foi adotada como critério de desempate.

No caso de 2019, as métricas apontaram maior adequação para 4 ou 5 clusters. Considerando que k igual a 4 apresentou melhor equilíbrio entre os valores de *silhouette*, Calinski-Harabasz e Davies-Bouldin, além da indicação clara do método do cotovelo, optou-se por essa solução. Dessa forma, o critério de seleção combinou a análise comparativa das métricas internas com o método do cotovelo, priorizando a configuração em que houve maior convergência entre os indicadores.

Tabela 2 – Métricas de Avaliação para Definição do Número de Grupos do K-Means
Fonte: Elaborado pelo autor

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

4.2.1 Analise gráfica

Após o cálculo das métricas de validação dos agrupamentos, foram geradas visualizações em projeções bidimensionais e tridimensionais, utilizando-se, respectivamente, os algoritmos t-SNE e UMAP. Essas técnicas de redução de dimensionalidade foram aplicadas com o objetivo de representar, em espaços de menor dimensão, a estrutura dos dados de alta dimensionalidade, preservando, tanto quanto possível, as relações de proximidade entre as observações. Essa abordagem possibilitou uma ava-

liação visual complementar da coesão e separação entre grupos, contribuindo para a validação qualitativa dos resultados obtidos pelas métricas quantitativas apresentadas na Tabela 2.

No ano de 2009, como ilustrado na Figura 3, observa-se a formação de três agrupamentos visualmente distintos, com fronteiras bem delimitadas e baixa sobreposição, indicando boa separabilidade entre os perfis identificados. Em 2015, conforme apresentado na Figura 4, manteve-se a configuração de três clusters; entretanto, nota-se maior dispersão intra-grupo, o que sugere aumento da variabilidade interna e possível diversificação dos perfis estudantis. Por fim, em 2019, como mostra a Figura 5, a configuração com quatro clusters evidencia novas segmentações nos padrões de agrupamento, com maior distância relativa entre determinados grupos. Essa reorganização espacial pode refletir mudanças estruturais nos comportamentos e contextos dos estudantes ao longo do período analisado, indicando transformações nas dinâmicas subjacentes aos fatores educacionais e comportamentais considerados.

4.3 Análise correlacional

Estabeleceu-se uma linha de corte para a análise, considerando apenas correlações iguais ou superiores a 25%. Nesse contexto, os períodos de 2009 e 2015 não apresentaram correlações que atendessem a esse critério, enquanto apenas o período de 2019 exibiu correlações estatisticamente significativas, cujos resultados são apresentados a seguir. Os maiores coeficientes de correlação de Pearson, calculados para as variáveis com distribuição normal, estão ilustrados na Figura 6, acompanhados de suas respectivas interpretações na Tabela 3.

Já entre as variáveis com distribuição não normal, os maiores valores de correlação de Spearman estão apresentados na Figura 7, com suas respectivas interpretações dispostas na Tabela 4. Os resultados da correlação de Kendall são mostrados na Figura 8, e suas interpretações encontram-se na Tabela 5.

Em suma, a correlação de Pearson identificou associação entre maior consumo de álcool e menores notas no IDEB, enquanto Spearman e Kendall encontraram associação entre maior consumo de cigarro e menores notas no IDEB. Esses dois últimos coeficientes também indicaram relação entre exposição sexual e piores notas no IDEB. Essa última associação pode ser explicada pelo fato de que os grupos foram formados a partir de estudantes do ensino fundamental e do ensino médio, sendo que o ensino fundamental apresenta, em média, maiores notas no IDEB e, estatisticamente, esse perfil de estudante tende a não estar sexualmente ativo. Conforme ilustra (Felisbinomendes; Araújo; Oliveira; Vasconcelos; Vieira; Malta, 2021), que analisou a mesma base de dados (PeNSE 2019), a idade média da primeira relação sexual no Brasil é de 16 anos para homens e 18 para mulheres, coincidindo com o período do ensino

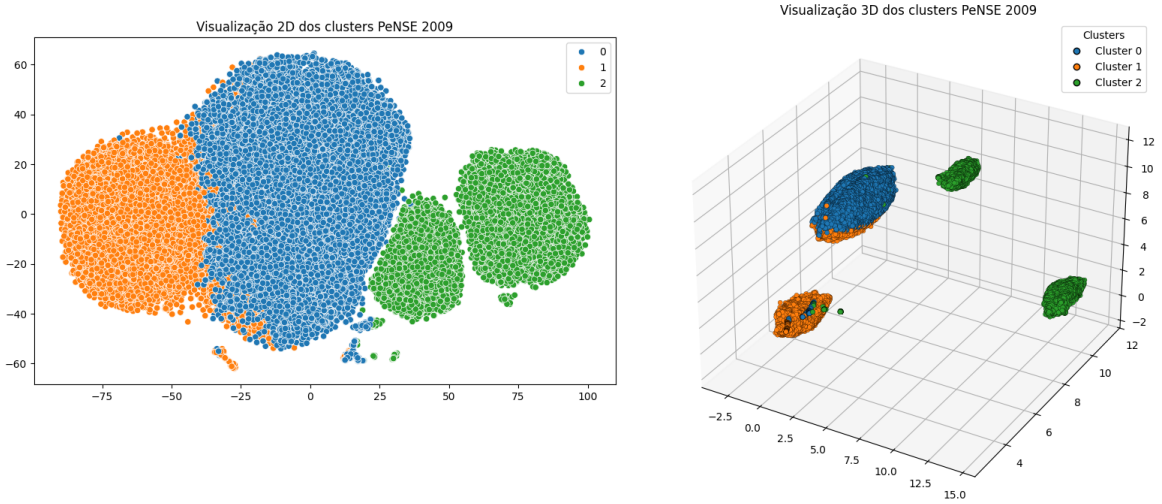

Figura 3 – Clusterização obtida — PeNSE 2009: representação bidimensional pelo t-SNE e tridimensional pelo UMAP.
Fonte: Elaborado pelo autor

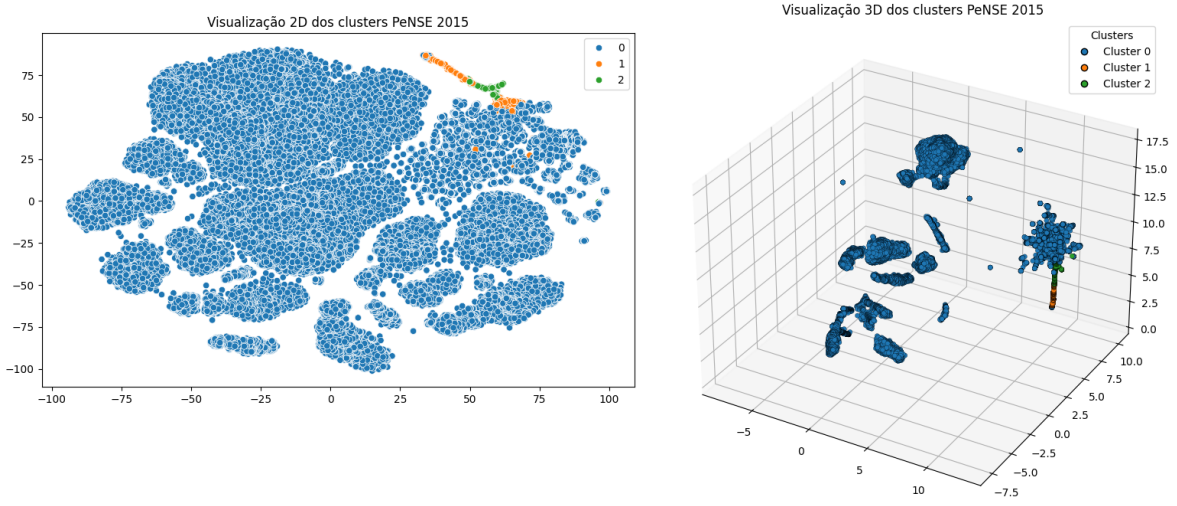

Figura 4 – Clusterização obtida — PeNSE 2015: representação bidimensional pelo t-SNE e tridimensional pelo UMAP.
Fonte: Elaborado pelo autor

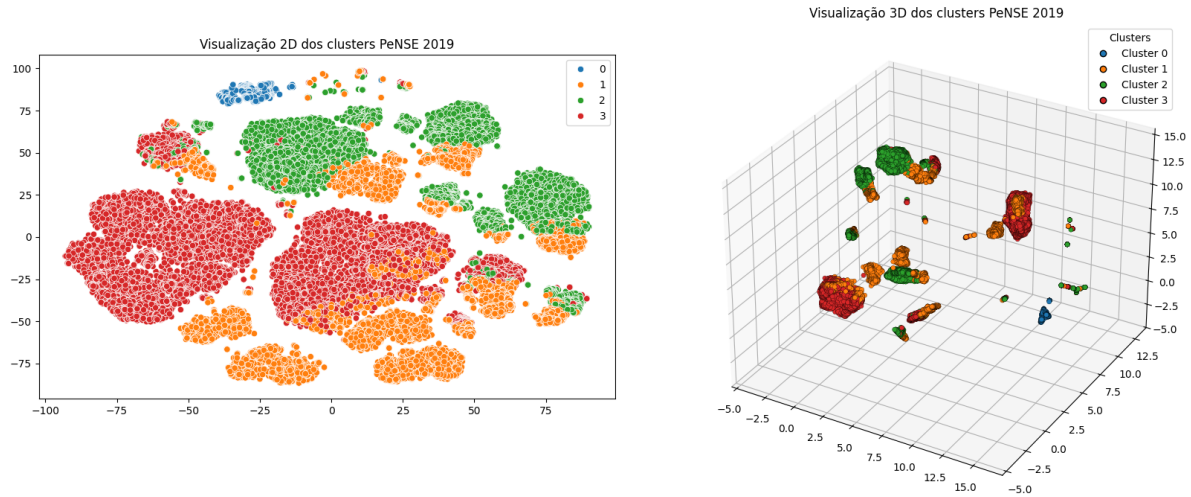

Figura 5 – Clusterização obtida — PeNSE 2019: representação bidimensional pelo t-SNE e tridimensional pelo UMAP.
Fonte: Elaborado pelo autor

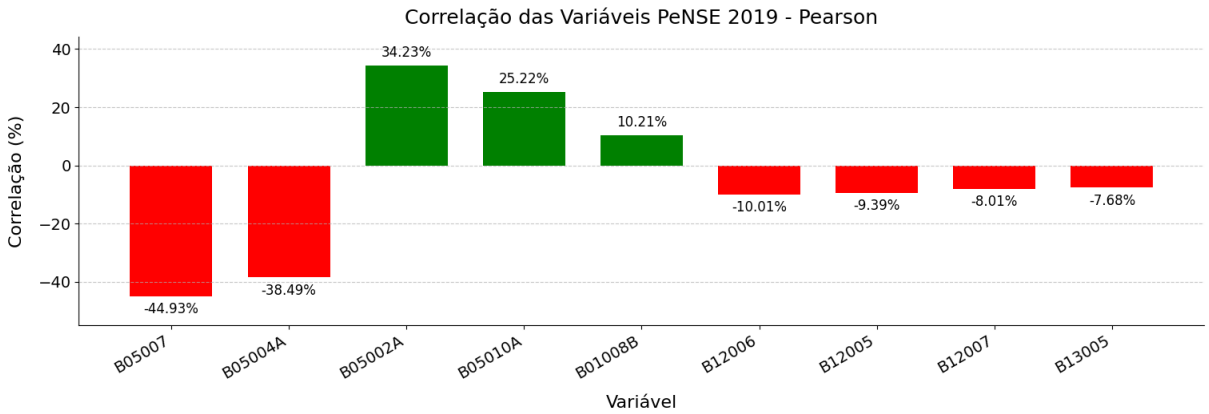

Figura 6 – Análise correlacional Pearson
Fonte: Elaborado pelo autor

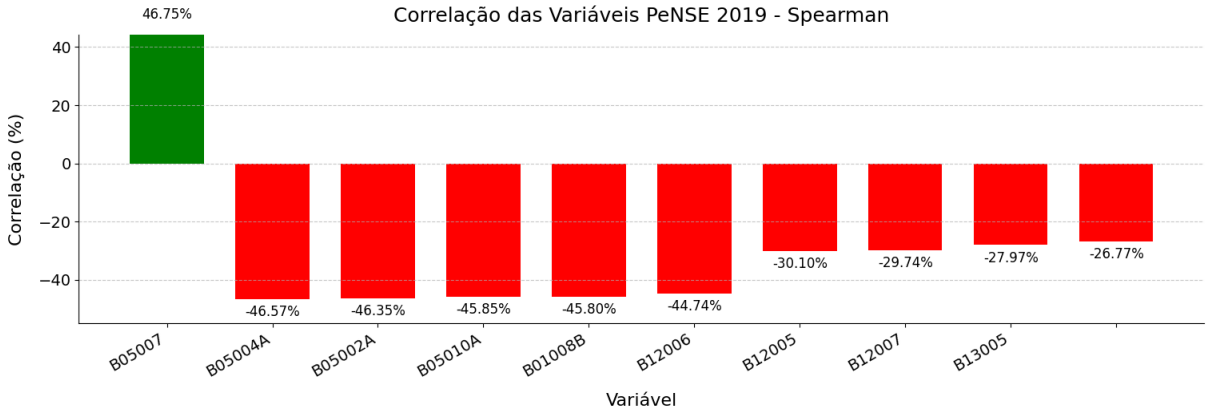

Figura 7 – Análise correlacional Spearman
Fonte: Elaborado pelo autor

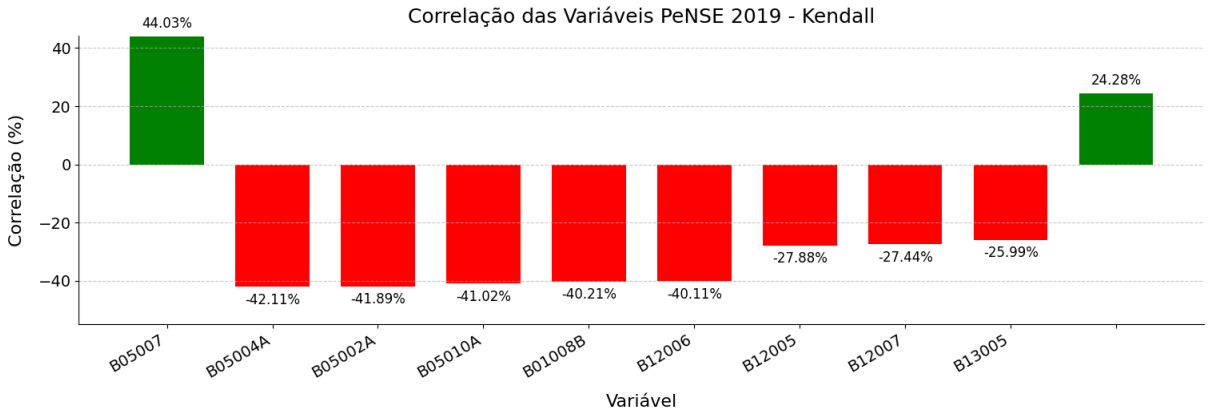

Figura 8 – Análise correlacional Kendall
Fonte: Elaborado pelo autor

médio, que possui IDEB menor.

4.4 Padrões encontrados

Após obter os clusters das PeNSEs de 2009, 2015 e 2019 e seus respectivos índices do IDEB, foi possível classificar os grupos em categorias de IDEB alto e IDEB baixo e realizar a análise por extremos dos indicadores. A pesquisa evidenciou as divergências entre esses dois grupos em cada edição da PeNSE (2009, 2015 e 2019), conforme ilustrado na Figura 9. Além disso, por meio de uma análise longitudinal, investigaram-se tanto as relações consistentes ao longo dos três ciclos avaliados quanto as particularidades emergentes em cada período, sempre à luz da literatura sobre os determinantes educacionais. Os principais resultados obtidos são apresentados nos tópicos a seguir:

4.4.1 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), Su-

pervisão 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 9, que compara os estados com melhores e piores desempenhos no IDEB, utilizando valores binários (-1 e +1) para destacar os contrastes.

(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.

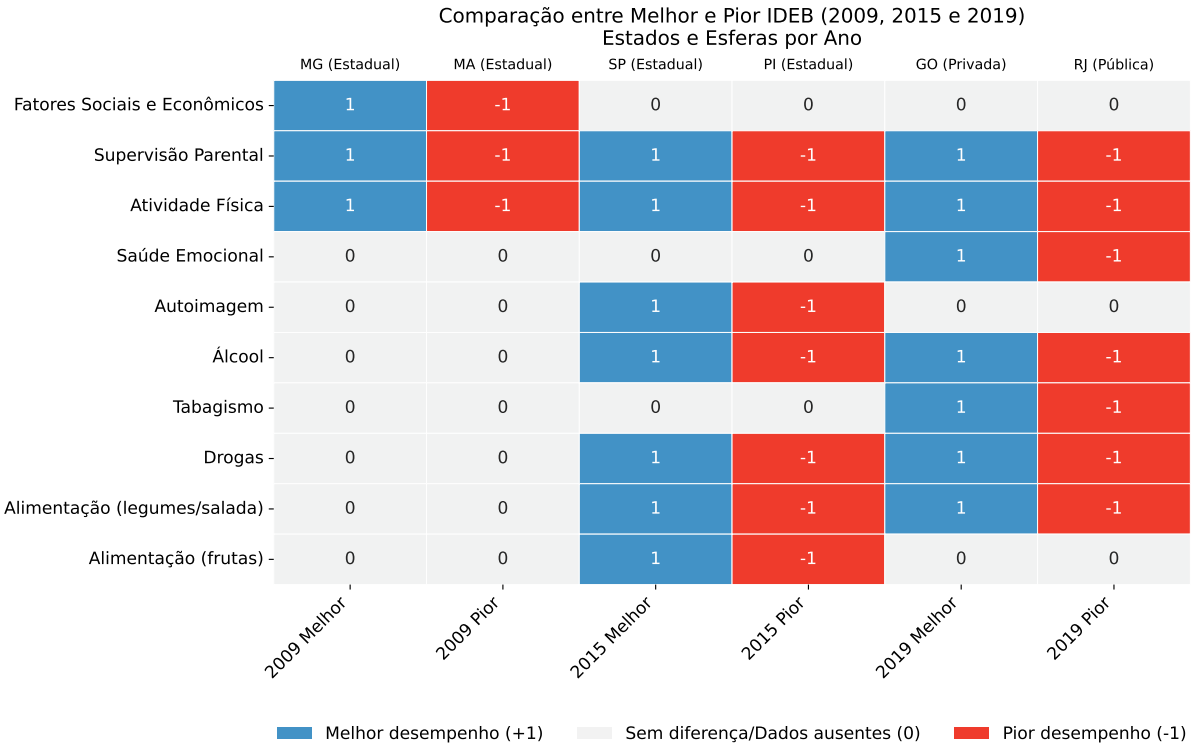

Figura 9 – Comparação entre Melhor e Pior IDEB (2009, 2015 e 2019). Estados e Esferas por Ano.
Fonte: Elaborado pelo autor

4.4.2 Padrões longitudinais

Conforme demonstrado na seção anterior, diversas variáveis apresentaram recorrência ao longo dos ciclos analisados — apesar das limitações impostas pela heterogeneidade dos questionários da PeNSE —, evidenciando padrões longitudinais relevantes. Para a construção da generalização dos comportamentos compilados dos ciclos, conforme ilustrado na Figura 10, foram selecionados os fatores que se repetiram em pelo menos dois períodos da pesquisa. Os resultados indicam que, ao longo do tempo analisado, grupos de estudantes com alimentação inadequada, baixa supervisão parental, prática insuficiente de atividade física e exposição precoce ao álcool e às drogas se associaram consistentemente a estados com desempenhos inferiores no IDEB. Por outro lado, os estudantes que não apresentavam essas características tenderam a estar em estados com melhores resultados no índice.

Tabela 3 – Correlações entre variáveis e IDEB - Pearson
Fonte: Elaborado pelo autor

Correlação	Variável	Interpretação
-0,449	B05007 - Na sua vida, quantas vezes você bebeu tanto que ficou realmente bêbado(a)?	Maior frequência de episódios de embriaguez está associada a menores desempenhos educacionais.
-0,385	B05004A - Nos últimos 30 dias, em quantos dias você tomou pelo menos um copo ou uma dose de bebida alcoólica?	O aumento no número de dias com consumo de álcool no último mês correlaciona-se negativamente com o IDEB.
0,342	B05002A - Alguma vez na vida você tomou um copo ou uma dose de bebida alcoólica?	Estudantes que nunca consumiram álcool apresentam, em média, melhores resultados no IDEB.
0,252	B05010A - Nos últimos 30 dias, algum dos seus amigos bebeu alguma bebida alcoólica na sua presença?	A não exposição ao consumo de álcool por parte de amigos no ambiente social recente está relacionada a melhores desempenhos educacionais.

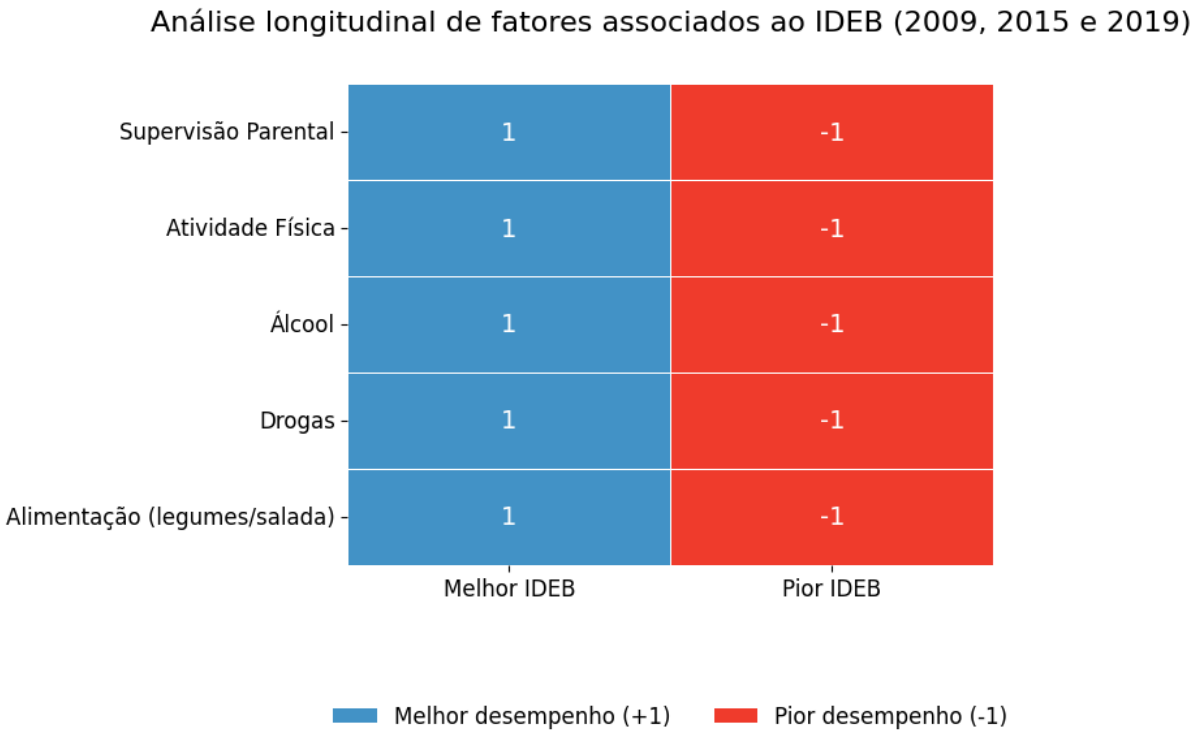

Figura 10 – Resultados longitudinais da análise comparativa entre os clusters das PeNSEs e seus respectivos índices do IDEB (2009, 2015, 2019)
Fonte: Elaborado pelo autor

Tabela 4 – Correlações entre variáveis e IDEB - Spearman
Fonte: Elaborado pelo autor

Correlação	**Variável**	**Interpretação**
0,468	B08001 - Você já teve relação sexual alguma vez?	Estudantes que não iniciaram a vida sexual apresentaram, em média, melhores resultados no IDEB.
-0,466	B08011A - Usou camisinha na PRIMEIRA RELAÇÃO SEXUAL?	A ausência do uso de preservativo na primeira relação sexual está negativamente associada ao desempenho educacional.
-0,464	B08006A - Usou camisinha na ÚLTIMA RELAÇÃO SEXUAL?	O não uso de preservativo na relação sexual mais recente correlaciona-se com menores níveis de IDEB.
-0,459	B08007 - Usou outro método contraceptivo (não camisinha) na ÚLTIMA relação?	A não utilização de métodos contraceptivos alternativos à camisinha na última relação sexual está relacionada a pior desempenho educacional.
-0,458	B08002 - Idade na primeira relação sexual	Início mais precoce da vida sexual está associado a menores desempenhos no IDEB.
-0,447	B08015 - Já usou pílula do dia seguinte?	A utilização da contracepção de emergência apresenta correlação negativa com o desempenho educacional.
-0,301	B08014 - Como conseguiu a camisinha na última vez?	A obtenção do preservativo através de meios alternativos está associado a menores desempenhos no IDEB.
-0,297	B08013A - Alguma vez engravidou?	Estudantes que não vivenciaram gravidez apresentam maior desempenho no IDEB.
-0,280	B08016 - Como conseguiu a pílula do dia seguinte na última vez?	A obtenção da pílula através de meios alternativos está associado a menores desempenhos no IDEB.
-0,268	B04002A - Idade quando fumou cigarro pela primeira vez	Início mais precoce do tabagismo apresenta relação negativa com os resultados no IDEB.

Tabela 5 – Correlações entre variáveis e IDEB - Kendall
Fonte: Elaborado pelo autor

Correlação	Variável	Interpretação
0,440	B08001 - Você já teve relação sexual (transou) alguma vez?	Estudantes que não iniciaram a vida sexual apresentam, em média, melhores desempenhos educacionais (IDEB).
-0,421	B08011A - Você ou seu(sua) parceiro(a) usou camisinha (preservativo) na primeira relação sexual?	A ausência de uso de preservativo na primeira relação sexual está associada a menores níveis de desempenho educacional.
-0,419	B08006A - Na última vez que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou camisinha (preservativo)?	O não uso de preservativo na última relação sexual mostra associação negativa com o desempenho educacional.
-0,410	B08007 - Na última vez que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou algum outro método para evitar a gravidez que não seja camisinha?	A ausência de uso de métodos contraceptivos alternativos na última relação sexual correlaciona-se com menores índices de IDEB.
-0,402	B08015 - Alguma vez na vida, você ou sua parceira já usou pílula do dia seguinte (contracepção de emergência)?	O não uso de contracepção de emergência após relações sexuais apresenta correlação negativa com o desempenho educacional.
-0,401	B08002 - Que idade você tinha quando teve relação sexual (transou) pela primeira vez?	Início mais precoce da vida sexual tende a estar relacionado a menores níveis de desempenho no IDEB.
-0,279	B08013A - Alguma vez na vida você engravidou, mesmo que a gravidez não tenha chegado ao fim?	Estudantes que não vivenciaram gravidez apresentam maior desempenho educacional.
-0,274	B08014 - Nesta última vez que você teve relação sexual (transou), como você conseguiu a camisinha (preservativo)?	A obtenção do preservativo por intermédio de alguém fora do círculo social está associada a menores desempenhos no IDEB.
-0,260	B08016 - Na última vez que você ou sua parceira usou pílula do dia seguinte (contracepção de emergência), como conseguiu?	A obtenção da pípula por intermédio de alguém fora do círculo social está associada a menores desempenhos no IDEB.
0,243	B04001 - Alguma vez na vida, você já fumou cigarro, mesmo uma ou duas tragadas?	Início mais precoce do tabagismo apresenta relação negativa com os resultados no IDEB.

5 CONSIDERAÇÕES FINAIS

Este estudo partiu da seguinte questão orientadora: quais padrões de fatores comportamentais e de estilo de vida, medidos pela PeNSE, podem ser identificados entre os estados brasileiros e como esses padrões se relacionam ao desempenho no IDEB? Para respondê-la, aplicou-se o algoritmo de clusterização K-means aos dados da PeNSE referentes aos anos de 2009, 2015 e 2019, considerando separadamente os grupos de estudantes por esfera administrativa (pública ou privada) em cada estado. Os agrupamentos obtidos foram então comparados aos respectivos desempenhos estaduais no IDEB e analisados quanto à correlação estatística entre as variáveis e o índice educacional.

Os resultados indicaram três agrupamentos em 2009, três em 2015 e quatro em 2019. Apesar da heterogeneidade entre os períodos analisados e do fato de que, em 2009, poucos padrões foram observados, identificou-se que, longitudinalmente, perfis caracterizados por baixa supervisão parental, alimentação inadequada, sedentarismo e exposição precoce ao álcool e a drogas se associaram aos estados com piores resultados no IDEB. Por outro lado, estados com melhores desempenhos apresentaram perfis marcados por maior supervisão, hábitos alimentares mais adequados e menor exposição a comportamentos de risco.

Esses achados mostraram-se mais consistentes nos dados de 2019, quando as associações alcançaram significância estatística após a correção para múltiplos testes, confirmando que os padrões identificados estão associados às trajetórias de desempenho educacional, conforme evidenciado pelos dados analisados.

A abordagem adotada apresenta limitações relevantes que precisam ser reconhecidas. Em primeiro lugar, trata-se de um estudo de natureza associativa ecológica, no qual a vinculação entre a PeNSE e o IDEB foi estabelecida em nível agregado. Essa opção metodológica implica o risco de falácia ecológica, já que os dados da PeNSE são autorrelatados pelos estudantes, enquanto o IDEB reflete médias populacionais por unidade da federação e esfera administrativa, sem pareamento individual. Além disso, a ausência de controle de variáveis, decorrente da heterogeneidade entre os ciclos da PeNSE e das diferenças nos questionários ao longo do tempo, limita aná-

lises mais precisas e generalizáveis. O uso de médias que combinam resultados do último ano do ensino fundamental e do ensino médio para representar o desempenho educacional também pode introduzir distorções na interpretação dos achados. Já, as limitações do ponto de vista estatístico, há ainda possíveis vieses decorrentes do não uso ou uso parcial dos pesos amostrais da PeNSE, bem como da sensibilidade do algoritmo K-Means às escalas das variáveis e ao processo de inicialização, fatores que podem influenciar a estabilidade dos agrupamentos formados. Soma-se a isso o potencial efeito da multicolinearidade entre as variáveis comportamentais consideradas, o que pode reduzir a clareza na interpretação dos padrões identificados.

Apesar das limitações inerentes à abordagem exploratória adotada, o estudo cumpre satisfatoriamente seus objetivos ao identificar perfis distintos de estudantes com base em características sociodemográficas e ao associar esses perfis aos resultados do IDEB, considerando as variações por ano, unidade federativa e esfera administrativa. As análises realizadas permitiram discutir padrões recorrentes de associação — ainda que sem pretensão de inferência causal — entre comportamentos estudantis e desempenho educacional, gerando insights relevantes sobre as dinâmicas que permeiam o contexto escolar brasileiro.

Do ponto de vista prático, os resultados obtidos configuram-se como um ponto de partida para a formulação de hipóteses mais específicas e a condução de estudos futuros capazes de aprofundar as relações aqui observadas. As evidências levantadas podem ser traduzidas em orientações estratégicas para a gestão pública, contribuindo para o desenho e o direcionamento de políticas educacionais e de saúde. Entre as possíveis aplicações, destacam-se:

(i) a priorização de estratégias preventivas em contextos nos quais perfis de risco se mostram mais frequentes;

(ii) a integração de políticas de saúde escolar com mecanismos de apoio familiar e comunitário;

(iii) e a utilização dos perfis identificados como instrumento de focalização territorial, auxiliando a alocação de recursos por unidade da federação e esfera administrativa.

Adicionalmente, fatores como supervisão parental, prática de exercícios físicos e alimentação adequada, embora já amplamente reconhecidos pela literatura, emergem neste estudo como elementos empiricamente promissores para orientar tais diretrizes — desde que confirmados por investigações subsequentes com outros desenhos metodológicos. Assim, o trabalho contribui ao oferecer uma base empírica inicial e um arcabouço analítico exploratório que podem servir de subsídio para a formulação de políticas públicas mais focalizadas e baseadas em evidências.

Diante das limitações identificadas e das potenciais aplicações práticas deste estudo, recomenda-se, para pesquisas futuras:

(i) a obtenção de dados mais granulares, com identificação individual dos estudantes da PeNSE — respeitando a LGPD — e vinculação direta ao desempenho no IDEB, de modo a possibilitar inferências causais e análises mais precisas em recortes geográficos menores, como municípios, distritos e escolas.

(ii) a exploração de informações relativas às escolas dos estudantes, disponíveis na base PeNSE mas não utilizadas neste trabalho, permitindo a geração de clusters que incorporem características institucionais.

(iii) a aplicação de outros algoritmos de clusterização particionais, como o K-Modes, ou ainda algoritmos baseados em densidade, como o HDBSCAN, ou modelos probabilísticos, como o GMM. Recomenda-se, adicionalmente, comparar os resultados obtidos por esses algoritmos com os apresentados neste trabalho, discutindo a convergência qualitativa entre diferentes soluções de agrupamento.

(iv) a incorporação de variáveis contextuais dos estados — como IDH, PIB, Censo Escolar, investimentos em educação e indicadores de desigualdade — de forma a enriquecer a análise atualmente centrada apenas na vivência dos alunos.

(v) a validação dos agrupamentos com apoio de especialistas da área educacional, bem como o desenvolvimento de modelos preditivos capazes de associar as notas do IDEB às diferentes regiões e prever o desempenho de novos perfis com base nos comportamentos previamente identificados.

(vi) o uso de procedimentos de remoção de variáveis de baixa variância e o tratamento da colinearidade (por exemplo, por meio do VIF ou da análise de correlações elevadas), favorecendo maior robustez e interpretabilidade dos modelos.

(vii) a avaliação da estabilidade dos agrupamentos por meio de repetições com diferentes valores de *seeds* e/ou estratégias de bootstrap, estimando a consistência dos resultados e relatando a concordância média entre rótulos (por exemplo, com o índice de Jaccard).

(viii) a realização de análises de correlações parciais, controlando por covariáveis como idade, sexo e etapa escolar, a fim de isolar efeitos específicos dos comportamentos estudantis sobre o desempenho educacional.

(ix) a incorporação de pesos amostrais, ausentes nesta análise devido ao caráter exploratório do estudo, mas relevantes para ampliar a representatividade e a validade inferencial dos resultados em pesquisas futuras.

(x) a condução de análises de clusterização segmentadas por nível de ensino — distinguindo os anos finais do ensino fundamental e o ensino médio — possibilitando resultados mais específicos e comparações úteis para subsidiar políticas públicas direcionadas a cada etapa.

(xi) a investigação dos perfis associados a valores intermediários do IDEB, evitando foco exclusivo nos extremos e permitindo captar nuances do desempenho educacional que podem revelar padrões menos evidentes, mas igualmente relevantes.

Em síntese, este trabalho oferece uma contribuição exploratória e metodológica ao associar dados da PeNSE ao desempenho educacional medido pelo IDEB, articulando aportes em três planos complementares. No plano científico, destaca-se a aplicação de técnicas de clusterização a bases nacionais educacionais, a integração das informações da PeNSE com os indicadores do IDEB e o desenvolvimento de um desenho replicável em estudos futuros. No plano metodológico, evidencia-se um pipeline transparente que abrange desde o pré-processamento dos dados, a escolha do número de clusters por métricas internas, a caracterização detalhada dos perfis até o vínculo ecológico com os resultados do IDEB, com o código integral disponibilizado no Apêndice D. No plano aplicado, os resultados fornecem insumos relevantes para redes de atenção e gestores, ao explicitar perfis de maior risco ou proteção que podem orientar ações intersetoriais, como prevenção ao uso de substâncias, educação em saúde e apoio familiar e escolar.

Ao retomar os objetivos propostos, conclui-se que foi possível identificar padrões longitudinais e regionais de associação entre comportamentos dos escolares e resultados educacionais, ainda que com limitações quanto à causalidade e à generalização dos achados. O fortalecimento desta linha de pesquisa, por meio da incorporação de métodos mais robustos e de dados mais detalhados, possui potencial para gerar contribuições substantivas ao entendimento e à melhoria das políticas públicas voltadas à educação no Brasil. Este estudo avança a compreensão de padrões combinados em dados nacionais, oferece evidências que podem ser úteis para gestores e abre trilhas metodológicas seguras para pesquisas futuras, mantendo-se fiel ao que foi efetivamente investigado.

REFERÊNCIAS

ALVES, A. S. A INFLUÊNCIA DO MEIO SOCIOECONÓMICO NA APRENDIZAGEM DA CRIANÇA. , [S.l.], 2023.

ALVES Filho, E. M. A. A IMPORTÂNCIA DA ALIMENTAÇÃO E ATIVIDADE FÍSICA NA EDUCAÇÃO INFANTIL. **Revista Ibero-Americana de Humanidades, Ciências e Educação**, [S.l.], v.10, n.4, p.1527–1528, 2024.

ALVES, M. T. G.; SOARES, J. F. Contexto escolar e indicadores educacionais: condições desiguais para a efetivação de uma política de avaliação educacional. **Educação e pesquisa**, [S.l.], v.39, n.01, p.177–194, 2013.

CÉSPEDES-GONZÁLEZ, Y.; ESCOBAR, A. D. O.; JIMÉNEZ, J. D. R.; MOLERO-CASTILLO, G. Academic Achievement in Mathematics of Higher-Middle Education Students in Veracruz: An Approach Based on Computational Intelligence. In: INTERNATIONAL CONFERENCE IN SOFTWARE ENGINEERING RESEARCH AND INNOVATION (CONISOFT), 2023., 2023. **Anais...** [S.l.: s.n.], 2023. p.177–185.

COHEN, J. **Statistical power analysis for the behavioral sciences**. [S.l.]: routledge, 2013.

CUI, M. et al. Introduction to the k-means clustering algorithm based on the elbow method. **Accounting, Auditing and Finance**, [S.l.], v.1, n.1, p.5–8, 2020.

de Araújo, J. M. et al. Fatores escolares como determinantes do desempenho dos alunos da educação básica. **Linhas Críticas**, [S.l.], v.27, 2021.

DIAZ-PAPKOVICH, A.; ANDERSON-TROCMÉ, L.; GRAVEL, S. A review of UMAP in population genetics. **Journal of Human Genetics**, [S.l.], v.66, n.1, p.85–91, 2021.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: SECOND INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 1996. **Proceedings...** [S.l.: s.n.], 1996. p.226–231.

FELISBINO-MENDES, M. S.; ARAÚJO, F. G.; OLIVEIRA, L. V. A.; VASCONCELOS, N. M. d.; VIEIRA, M. L. F. P.; MALTA, D. C. Comportamento sexual e uso de preservativos na população brasileira: análise da Pesquisa Nacional de Saúde, 2019. **Revista Brasileira de Epidemiologia**, [S.l.], v.24, p.e210018, 2021.

FURLANETTO, G. C.; CARVALHO, V. O. de; BALDASSIN, A.; MANACERO, A. Algoritmos de agrupamento aplicados à detecção de fraudes. In: ESCOLA REGIONAL DE ALTO DESEMPENHO DE SÃO PAULO (ERAD-SP), 2022. **Anais...** [S.l.: s.n.], 2022. p.29–32.

GOMES, M. M. Fatores que facilitam e dificultam a aprendizagem. **Revista Educação Pública, Rio de Janeiro**, [S.l.], v.18, n.14, p.28–38, 2018.

HATEM, G.; ZEIDAN, J.; GOOSSENS, M.; MOREIRA, C. Normality testing methods and the importance of skewness and kurtosis in statistical analysis. **BAU Journal-Science and Technology**, [S.l.], v.3, n.2, p.7, 2022.

Hernández-Leal, E.; Duque-Méndez, N. D.; Cechinel, C. Unveiling educational patterns at a regional level in Colombia: data from elementary and public high school institutions. **Heliyon**, [S.l.], v.7, n.9, 2021.

IBGE. **Pesquisa Nacional de Saúde do Escolar (PeNSE)**. 2009. Acesso em: October 13, 2025, `https://www.ibge.gov.br/estatisticas/sociais/saude/9134-pesquisa-nacional-de-saude-do-escolar.html`. Disponível em: <https://www.ibge.gov.br/estatisticas/sociais/saude/9134-pesquisa-nacional-de-saude-do-escolar.html>.

INEP. **Índice de Desenvolvimento da Educação Básica (IDEB)**. 2025. Acesso em: October 13, 2025, `https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb`. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb>.

JAIN, A. K. Data clustering: 50 years beyond K-means. **Pattern Recognition Letters**, [S.l.], v.31, n.8, p.651–666, 2010.

JAIN, A. K.; DUBES, R. C. **Algorithms for Clustering Data**. [S.l.]: Prentice Hall, 1988.

JOHNSON, S. C. Hierarchical clustering schemes. **Psychometrika**, [S.l.], v.32, n.3, p.241–254, 1967.

KAUFMAN, L.; ROUSSEEUW, P. J. **Finding Groups in Data**: An Introduction to Cluster Analysis. [S.l.]: Wiley, 1990.

LENZ, M.; NEUMAN, F.; SANTARELLI, R.; SALVADOR, D. Fundamentos de aprendizagem de máquina. **Porto Alegre: SAGAH**, [S.l.], 2020.

LLOYD, S. Least squares quantization in PCM. **IEEE transactions on information theory**, [S.l.], v.28, n.2, p.129–137, 1982.

MAATEN, L. v. d.; HINTON, G. Visualizing data using t-SNE. **Journal of machine learning research**, [S.l.], v.9, n.Nov, p.2579–2605, 2008.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations**. [S.l.]: University of California Press, 1967.

MAIA, M. M.; ANDRADE, L. H. F. de; FERNANDES, S. K-means na análise de características socioeconômicas de candidatos ao ensino superior. **Anais do Encontro de Computação do Oeste Potiguar ECOP/UFERSA (ISSN 2526-7574)**, [S.l.], v.2, n.5, 2021.

MELLO, F. C. M.; SILVA, J. L. d.; OLIVEIRA, W. A. d.; PRADO, R. R. d.; MALTA, D. C.; SILVA, M. A. I. A prática de bullying entre escolares brasileiros e fatores associados, Pesquisa Nacional de Saúde do Escolar 2015. **Ciência & Saúde Coletiva**, [S.l.], v.22, p.2939–2948, 2017.

MIOT, H. A. **Análise de correlação em estudos clínicos e experimentais**. 2018. [S.l.]: SciELO Brasil, 2018. 275–279p. v.17, n.4.

NAEEM, S.; ALI, A.; ANAM, S.; AHMED, M. M. An Unsupervised Machine Learning Algorithms: Comprehensive Review. **International Journal of Computing and Digital Systems**, [S.l.], v.13, n.1, p.911–921, 2023.

OCDE. **PISA 2022 Results (Volume I and II) - Country Notes**: Brazil. 2022. Acesso em 13 jul. 2024. Disponível em: <https://www.oecd.org/en/publications/pisa-2022-results-volume-i-and-ii-country-notes$_e d6 f b c c 5 - e n / b r a z i l _6 1 6 9 0 6 4 8 - e n . h t m l >$.

REIS, A. A. C. d.; MALTA, D. C.; FURTADO, L. A. C. Desafios para as políticas públicas voltadas à adolescência e juventude a partir da Pesquisa Nacional de Saúde do Escolar (PeNSE). **Ciência & saúde coletiva**, [S.l.], v.23, n.9, p.2879–2890, 2018.

SAMPAIO, N.; BOSCO, E. Aplicações da correlação e regressão linear. **Associacao Educacional Dom Boco**, [S.l.], 2015.

SANTOS, A. C. d.; FERREIRA, D. F. Definição do tamanho amostral usando simulação Monte Carlo para o teste de normalidade baseado em assimetria e curtose: I. Abordagem univariada. **Ciência e Agrotecnologia**, [S.l.], v.27, p.432–437, 2003.

SHAQIRI, M.; ILJAZI, T.; KAMBERI, L.; RAMANI-HALILI, R. Differences Between The Correlation Coefficients Pearson, Kendall And Spearman. **Journal of Natural Sciences and Mathematics of UT**, [S.l.], v.8, n.15-16, p.392–397, 2023.

SILVA, F. M. A Educação Estadual de Goiás à Beira da Aprovação Automática dos Estudantes. **Revista Terceiro Incluído**, [S.l.], v.15, n.1, p.e15104–e15104, 2025.

SOARES, C. A. M.; LEÃO, O. A. d. A.; FREITAS, M. P.; HALLAL, P. C.; WAGNER, M. B. Tendência temporal de atividade física em adolescentes brasileiros: análise da Pesquisa Nacional de Saúde do Escolar de 2009 a 2019. **Cadernos de Saúde Pública**, [S.l.], v.39, n.10, p.e00063423, 2023.

SOARES, M. L.; Bernardo Junior, R. 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**, [S.l.], v.11763, 2018.

SOUZA, T. M.; CHAGAS, A. M.; CASSIA, A. Rita de et al. O Índice de Desenvolvimento da Educação Básica (Ideb): uma década de monitoramento da qualidade da educação. **Revista Com Censo: Estudos Educacionais do Distrito Federal**, [S.l.], v.6, n.2, p.57–62, 2019.

TORCATE, A. S.; BARBOSA, J. C. F.; OLIVEIRA RODRIGUES, C. M. de. Utilizando o learning analytics com o k-means para análise de dificuldades de aprendizagem na educação básica. In: WORKSHOP DE INFORMÁTICA NA ESCOLA (WIE), 2020. **Anais...** [S.l.: s.n.], 2020. p.31–40.

TORCATE, A. S.; BARBOSA, J. C. F.; OLIVEIRA RODRIGUES, C. M. de. Utilizando o learning analytics com o k-means para análise de dificuldades de aprendizagem na educação básica. In: WORKSHOP DE INFORMÁTICA NA ESCOLA (WIE), 2020. **Anais...** [S.l.: s.n.], 2020. p.31–40.

UNICEF. As múltiplas dimensões da pobreza na infância e na adolescência no Brasil. **Brasília, DF: Unicef**, [S.l.], 2023.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J.; DATA, M. Practical machine learning tools and techniques. In: DATA MINING, 2016. **Anais...** [S.l.: s.n.], 2016. v.2, n.4, p.403–413.

Apêndices

APÊNDICE A – Descrição das Variáveis utilizadas PeNSE 2009

Ressalta-se que as fontes contendo o dicionário das respostas das variáveis encontram-se no Apêndice E.

Variável	**Descrição**
ID	Identificação da escola
QUESTIONARIO	identificação do tipo de questionario
B00P01	Prezado(a) estudante, você concorda em participar dessa pesquisa?
B01P01	Qual é o seu sexo?
B01P02	Qual a sua cor ou raça?
B01P03	Qual a sua idade?
B01P04	Qual o mês do seu aniversário?
B01P05	Em que ano você nasceu?
B01P06	Você mora com sua mãe?
B01P07	Você mora com seu pai?
B01P08	Até que nível de ensino(grau) sua mãe estuda ou estudou?
B01P12	Na sua casa tem televisão?
B01P13	Na sua casa tem geladeira?
B01P14	Na sua casa tem fogão?
B01P15	Na sua casa tem forno de microondas?
B01P16	Na sua casa tem máquina de lavar roupa? (Não considere o tanquinho)
B01P17	Na sua casa tem telefone fixo (convencional)?
B01P18	Você tem celular?
B01P19	Na sua casa tem aparelho de DVD?
B01P20	Na sua casa tem computador?
B01P21	Na sua casa tem algum computador ligado à Internet?
B01P22	Alguém que mora na sua casa tem carro?
B01P23	Alguém que mora na sua casa tem moto?
B01P24	Dentro da sua casa tem banheiro?
B01P25	Quantos banheiros com chuveiro tem dentro da sua casa?

Variável	Descrição
B01P26	Tem empregado(a) doméstico(a) recebendo dinheiro para fazer o trabalho em sua casa, cinco ou mais dias por semana?
B02P01	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu feijão?
B02P02	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu batata frita? (Incluir a batata de pacote)
B02P03	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu salgados fritos? Exemplo: coxinha de galinha, quibe frito, pastel frito, acarajé, etc.
B02P04	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu hambúrguer, salsicha, mortadela, salame, presunto, nuggets ou linguiça?
B02P05	NOS ULTIMOS 7 DIAS, em quantos dias você comeu pelo menos um tipo de legume ou verdura, excluindo batata e aipim (mandioca)? Exemplo: couve, abóbora, chuchu, brócolis, espinafre, etc.
B02P06	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu salada crua? Exemplo: alface ou tomate ou cenoura ou pepino ou cebola etc.
B02P07	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu legumes ou verduras cozidos na comida ou sopa, excluindo batata e mandioca? Exemplo: couve, abóbora, chuchu, brócolis, espinafre, etc.
B02P08	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu biscoitos salgados ou bolachas salgadas?
B02P09	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu biscoitos doces ou bolachas doces?
B02P10	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu guloseimas (doces, balas, chocolates, chicletes, bombons ou pirulitos)?
B02P11	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu frutas frescas ou salada de frutas?
B02P12	NOS ÚLTIMOS 7 DIAS, em quantos dias você tomou leite? (Excluir leite de soja)
B02P13	NOS ÚLTIMOS 7 DIAS, em quantos dias você tomou refrigerante?
B02P14	NOS ÚLTIMOS 7 DIAS, na maioria das vezes em que você tomou refrigerante, ele foi de que tipo?

Variável	Descrição
B02P15	Ontem, em quais refeições você comeu salada crua? Exemplo: alface ou tomate ou cenoura ou pepino ou cebola etc.
B02P16	Ontem, em quais refeições você comeu legumes ou verduras cozidos, sem contar batata e aipim (mandioca/macaxeira)?
B02P17	Ontem, quantas vezes você comeu frutas frescas?
B02P18	Você costuma fazer alguma dessas refeições - almoço ou jantar - com sua mãe ou responsável?
B02P19	Você costuma comer quando está assistindo à TV ou estudando?
B03P01	NOS ÚLTIMOS 7 DIAS, em quantos dias você foi a pé ou de bicicleta para a escola?
B03P02	NOS ÚLTIMOS 7 DIAS, em quantos dias você voltou a pé ou de bicicleta da escola?
B03P03	Quando você vai a pé ou de bicicleta para a escola, quanto tempo você gasta? (CONTAR APENAS O TEMPO GASTO NA IDA OU NA VOLTA. NÃO SOMAR IDA E VOLTA)
B03P04	NOS ÚLTIMOS 7 DIAS, quantas vezes você teve aulas de educação física na escola?
B03P05	NOS ÚLTIMOS 7 DIAS, quanto tempo por dia você fez atividade física ou esporte durante as aulas de Educação Física na escola?
B03P06	NOS ÚLTIMOS 7 DIAS, sem contar as aulas de educação física da escola, em quantos dias você praticou alguma atividade física, como esportes, dança, ginástica, musculação, lutas ou outra atividade com a orientação de professor ou instrutor?
B03P07	Normalmente, quanto tempo por dia duram essas atividades que você faz com professor ou instrutor? (Não incluir as aulas de educação física)
B03P08	NOS ÚLTIMOS 7 DIAS, no seu tempo livre, em quantos dias você praticou atividade física ou esporte sem professor ou instrutor?
B03P09	Normalmente, quanto tempo por dia duram essas atividades que você faz sem professor ou instrutor?
B03P10	Se você tivesse oportunidade de fazer atividade física na maioria dos dias da semana, qual seria a sua atitude?
B03P11	Num dia de semana comum, quantas horas por dia você assiste a TV?

Variável	Descrição
B03P12	Num dia de semana comum, quantas horas por dia você joga videogame?
B03P13	Num dia de semana comum, quantas horas por dia você fica no computador?
B04P01	Alguma vez na vida, você já fumou cigarro, mesmo uma ou duas tragadas?
B04P02	Que idade você tinha quando experimentou fumar cigarro pela primeira vez?
B04P03	NOS ÚLTIMOS 30 DIAS, em quantos dias você fumou cigarros?
B04P04	NOS ÚLTIMOS 12 MESES, você tentou parar de fumar?
B04P05	NOS ULTIMOS 7 DIAS, em quantos dias outras pessoas fumaram na sua casa?
B04P06	Qual de seus pais ou responsáveis fuma?
B04P07	Se quiser, você consegue comprar cigarro na escola?
B04P08	Se você fumasse cigarros, qual seria a reação de sua família se ela ficasse sabendo?
B05P01	Alguma vez na vida, você já experimentou bebida alcoólica?
B05P02	Que idade você tinha quando experimentou bebida alcoólica pela primeira vez?
B05P03	NOS ÚLTIMOS 30 DIAS, em quantos dias você tomou pelo menos um copo ou uma dose de bebida alcoólica?
B05P04	Nos últimos 30 dias, nos dias em que você tomou alguma bebida alcoólica, quantos copos ou doses você tomou por dia?
B05P05	NOS ÚLTIMOS 30 DIAS, na maioria das vezes, como você conseguiu a bebida que tomou?
B05P06	Na sua vida, quantas vezes você bebeu tanto que ficou realmente bêbado(a)?
B05P07	Se você chegasse em casa bêbado(a), qual seria a reação de sua família se ela ficasse sabendo?
B05P08	Na sua vida, quantas vezes você teve problemas com sua família ou amigos, perdeu aulas, se machucou ou brigou porque tinha bebido?
B05P09	Alguma vez na vida, você já usou alguma droga, tais como: maconha, cocaína, crack, cola, loló, lança perfume, ecstasy etc?
B05P10	Nos últimos 30 dias, quantas vezes você usou drogas tais como maconha, cocaína, crack, cola, loló, lança perfume, ecstasy etc?

Variável	Descrição
B05P11	Que idade você tinha quando usou droga tais como maconha, cocaína, crack, cola, loló, lança perfume, ecstasy ou outra pela primeira vez?
B06P01	NOS ÚLTIMOS 30 DIAS, em quantos dias você faltou às aulas sem permissão dos seus pais ou responsáveis?
B06P02	NOS ÚLTIMOS 30 DIAS, com que frequência seus pais ou responsáveis sabiam realmente o que você estava fazendo em seu tempo livre?
B06P03	NOS ÚLTIMOS 30 DIAS, com que frequência os colegas de sua escola trataram você bem e/ou foram prestativos com você?
B06P04	NOS ÚLTIMOS 30 DIAS, com que frequência algum dos seus colegas de sua escola te esculacharam, zoaram, mangaram, intimidaram ou caçoaram tanto que você ficou magoado / incomodado / aborrecido / ofendido / humilhado?
B07P01	Você já teve relação sexual (transou) alguma vez?
B07P02	Que idade você tinha quando teve relação sexual (transou) pela primeira vez?
B07P03	Na sua vida, você já teve relação sexual (transou) com quantas pessoas?
B07P04	NOS ÚLTIMOS 12 MESES, você teve relações sexuais(transou)?
B07P05	Na última vez que você teve relação sexual(transou), você ou seu(sua) parceiro(a) usou algum método para evitar a gravidez?
B07P06	Na última vez que você teve relação sexual(transou), você ou seu(sua) parceiro(a) usou camisinha(preservativo)?
B07P07	Na escola, você já recebeu orientação sobre prevenção de gravidez?
B07P08	Na escola, você já recebeu orientação sobre Aids ou outras doenças sexualmente transmissiveis (DSTs)?
B07P09	Na escola, você já recebeu orientação sobre como conseguir camisinha(preservativo) gratuitamente?
B08P01	NOS ÚLTIMOS 30 DIAS, em quantos dias você deixou de ir à escola porque não se sentia seguro no caminho de casa para a escola ou da escola para casa?
B08P02	NOS ÚLTIMOS 30 DIAS, em quantos dias você não foi à escola porque não se sentia seguro na escola?

Variável	Descrição
B08P03	NOS ÚLTIMOS 30 DIAS, quantas vezes você foi agredido fisicamente por um adulto da sua família?
B08P04	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguém foi fisicamente agredido?
B08P05	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou arma de fogo como revólver ou espingarda?
B08P06	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou alguma outra arma como faca, canivete, peixeira, pedra, pedaço de pau ou garrafa?
B08P07	NOS ÚLTIMOS 30 DIAS, quantas vezes você usou o cinto de segurança quando estava em um carro ou outro veículo motorizado dirigido por outra pessoa (excluindo ônibus)?
B08P08	NOS ÚLTIMOS 30 DIAS, quantas vezes você usou um capacete ao andar de motocicleta?
B08P09	NOS ÚLTIMOS 30 DIAS, quantas vezes você dirigiu um veículo motorizado de transporte (carro, motocicleta, voadeira, barco) ?
B08P10	NOS ÚLTIMOS 30 DIAS, quantas vezes você andou em carro ou outro veículo motorizado dirigido por alguém que tinha consumido alguma bebida alcoólica?
B09P01	Normalmente, quantas vezes por dia você escova os dentes?
B09P02	NOS ÚLTIMOS SEIS MESES, você teve dor de dente (excluir dor de dente causada por uso de aparelho)?
B10P01	Quanto ao seu corpo, você se considera:
B10P02	O que você está fazendo em relação a seu peso?
B10P03	NOS ÚLTIMOS 30 DIAS, você vomitou ou tomou laxantes para perder peso ou evitar ganhar peso?
B10P04	NOS ÚLTIMOS 30 DIAS, você tomou algum remédio, fórmula ou outro produto para perder ou manter seu peso sem acompanhamento médico?
B11P01	O que você achou deste questionário?
Q2B01P01	Qual é o Peso do Aluno ?
Q2B01P02	Qual é a Altura do Aluno ?
COD_UF	CÓDIGO DA UF DA ESCOLA
COD_MUNICIPIO	CÓDIGO DO NUNICÍPIO DA ESCOLA
DEPEND_ADM	DEPENDÊNCIA ADMINISTRATIVA
ESTALOCA	Estrato de alocação do estudante na amostra

Variável	Descrição
B01P03M	Marca de imputação de idade
TURMA	Identificação da turma selecionada na escola
TURMAS	Número de turmas de 9º ano da escola
MATRIC	Número de alunos matriculados na turma
FREQ	Número de alunos que frequentavam regularmente as aulas na turma
PESQ	Número de alunos presentes na turma no dia da pesquisa
CAPITAL	Código de identificação da capital (mesmo código da UF)
PESQ_CRIT	Número corrigido de alunos presentes na turma no dia da pesquisa
FREQ_CRIT	Número corrigido de alunos que frequentavam regularmente as aulas na turma
PESO_ESCOLA	Peso amostral da escola (a ser usado na tabulação das variáveis de ambiente)
N_TURMAS	Número de turmas de 9º ano da escola selecionadas para a amostra
PROBSELTURMA	Probabilidade de seleção das turmas em uma dada escola
PESO_TURMA	Fator de expansão da turma
PART	Número de alunos presentes da turma, que efetivamente participaram da pesquisa
N_PART	Número de alunos presentes da turma, que se recusaram a participar da pesquisa
SEXO	Número de alunos que participaram da pesquisa e informaram seu sexo
PESO_AJU_FREQ	Fator de expansão da turma ajustado pela frequência escolar
PESO_AJU_SEXO	Fator de expansão da turma ajustado pela frequência escolar
DEP2	Dependência administrativa da escola
GR_IDADE	Grupo de idade
INST_MAE	Instrução da mãe
FRUTA	Comeu frutas frescas ou saladas de frutas
PARENTAL	Reside com os pais
COME_RESP	Refeições com os responsáveis
RESP_SABE	Responsáveis sabe o que fazia
RESP_FUMO	Responsáveis fumam
FALTA_AULA	Falta às aulas sem permissão dos pais
AGRESSAO	Agressão por adulto da família
COMEU_FEIJAO	Comeu feijão na semana

Variável	Descrição
COMEU_FRUTA	Comeu frutas na semana
COMEU_GULOS	Comeu guloseimas na semana
BEBEU_REFRI	Bebeu refrigerante na semana
TV	Horas assistindo televisão
PC	Horas no computador
EXPCIG	Experimentou cigarro
FUMAREG	Fumou nos últimos 30 dias
EXPALC	Experimentou bebida alcoólica
BEBEREG	Bebeu nos últimos 30 dias
EMBRIAGUEZ	Ficou bêbado alguma vez na vida
EXPDRO	Experimentou drogas como maconha, cocaína, crack etc
EXPSEX	Iniciação sexual
AULASEF	Aulas de educação física
BULLYING	Sofreu bulling na escola
PRESERV	Usou preservativo na última relação sexual
SEGTRAJ	Deixou de ir a escola por insegurança no trajeto
SEGESC	Deixou de ir a escola por insegurança na escola
SENTINSEG	Deixou de ir a escola por insegurança no trajeto ou na escola
CINTO	Uso do cinto de segurança
DIRIGIU	Dirigiu veículo motorizado com menos de 18 anos
MOTALCOOL	Andou em veículo motorizado dirigido por motorista alcoolizado
ESCOVA	Escova os dentes com frequência
DORDENTE	Teve dor de dente
IMAGEM	Obesidade
PRODPESO	Vomitou ou tomou laxantes para perder peso
FEIJAO	Comeu feijão quantos dias na semana
BATATA	Comeu batata frita quantos dias na semana
SALG	Comeu salgados fritos quantos dias na semana
HAMB	Comeu hambúrguer quantos dias na semana
LEGU	Comeu legume ou verdura quantos dias na semana
SALADA	Comeu salada crua quantos dias na semana
LEGUC	Comeu legume ou verdura cozidos ou sopa quantos dias na semana
BISCSALG	Comeu biscoito salgados quantos dias na semana
BISCDOCE	Comeu biscoito doces quantos dias na semana
GULO	Comeu guloseimas quantos dias na semana

Variável	**Descrição**
FRUFRE	Comeu frutas frescas ou salada de frutas quantos dias na semana
LEITE	Bebeu leite quantos dias na semana
REFRI	Bebeu refrigerante quantos dias na semana
TEMPATIV	Tempo semanal de atividade física em minutos
ESCORATI	Classificação do aluno segundo seu tempo semanal de atividade física
Q2B01P01I	Massa corporal imputada
Q2B01P02I	Altura imputada
DEF_ALT	Deficit de altura
DEF_PES	Deficit de peso
EXC_PES	Excesso de peso

APÊNDICE B – Descrição das Variáveis utilizadas PeNSE 2015

Ressalta-se que as fontes contendo o dicionário das respostas das variáveis encontram-se no Apêndice E.

Variável	**Descrição**
ANOPESQ	Ano em que a pesquisa foi realizada
PAIS	País
REGEOGR	Região geográfica
UFCENSO	Unidade da Federação
MUNICIPIO_CAP	Município
TIPO_MUNIC	Indicadora de município da capital
VB00004	Prezado(a) estudante, você concorda em participar dessa pesquisa?
VB01001	Qual é o seu sexo?
VB01002	Qual é a sua cor ou raça?
VB01003	Qual é a sua idade?
VB01004	Qual é o mês do seu aniversário?
VB01005	Em que ano você nasceu?
VB01021	Em que ano/série você está?
VB01022	Em que turno você estuda?
VB01023	Você estuda em regime integral (tem atividades escolares por 7 horas ou mais horas diárias, durante todo o período escolar)?
VB01024	Você estuda em regime de internato (a escola possui alojamento onde os alunos permanecem e dormem diariamente, durante todo o período escolar)?
VB01025	Qual o grau de escolaridade mais elevado que você pretendo concluir?
VB01026	Quando terminar o ciclo/curso que você está frequentando atualmente, você pretende?
VB01006	Você mora com sua mãe?
VB01007	Você mora com seu pai?

Variável	Descrição
VB01010A	Contando com você, quantas pessoas moram na sua casa ou apartamento?
VB01013	Na sua casa tem telefone fixo (convencional)?
VB01014	Você tem celular?
VB01015A	Na sua casa tem computador (de mesa, ou netbook, laptop, etc)?
VB01016	Você tem acesso à internet em sua casa?
VB01017	Alguém que mora na sua casa tem carro?
VB01018	Alguém que mora na sua casa tem moto?
VB01019	Quantos banheiros com chuveiro têm dentro da sua casa?
VB01020A	Tem empregado(a) doméstico(a) recebendo dinheiro para fazer o trabalho em sua casa, três ou mais dias por semana?
VB01008A	Qual nível de ensino (grau) sua mãe estudou ou estuda?
VB01011	Você tem algum trabalho, emprego ou negócio atualmente?
VB01012	Você recebe dinheiro por este trabalho, emprego ou negócio?
VB02019A	Você costuma tomar o café da manhã?
VB02017A	Você costuma almoçar ou jantar com sua mãe, pai ou responsável?
VB02018A	Você costuma comer quando está assistindo à TV ou estudando?
VB02021	Sua escola oferece comida (merenda escolar/almoço) aos alunos da sua turma? (Não considerar lanches/comida comprados na cantina)
VB02020A	Você costuma comer a comida (merenda/almoço) oferecida pela escola? (Não considerar lanches/comida comprados na cantina).
VB02001	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu feijão?
VB02002	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu salgados fritos? Exemplo: batata frita (sem contar a batata de pacote) ou salgados fritos como coxinha de galinha, quibe frito, pastel frito, acarajé etc.
VB02004A	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu pelo menos um tipo de legume ou verdura? Exemplos: alface, abóbora, brócolis, cebola, cenoura, chuchu, couve, espinafre, pepino, tomate etc. Não inclua batata e aipim (mandioca/macaxeira).
VB02010	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu guloseimas (doces, balas, chocolates, chicletes, bombons ou pirulitos)?

Variável	Descrição
VB02011	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu frutas frescas ou salada de frutas?
VB02013	NOS ÚLTIMOS 7 DIAS, em quantos dias você tomou refrigerante?
VB02022	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu alimentos industrializados/ultraprocessados salgados, como hambúrguer, presunto, mortadela, salame, linguiça, salsicha, macarrão instantâneo, salgadinho de pacote, biscoitos salgados?
VB02023	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu em restaurantes fast food, tais como lanchonetes, barracas de cachorro quentes, pizzaria etc?
VB02024	NOS ÚLTIMOS 30 DIAS, com que frequência você ficou com fome por não ter comida suficiente em sua casa?
VB02025	NOS ÚLTIMOS 30 DIAS, quantas vezes por dia você normalmente comeu frutas frescas ou salada de frutas?
VB02026	NOS ÚLTIMOS 30 DIAS, quantas vezes por dia você normalmente comeu legumes ou verduras, tais como alface, abóbora, brócolis, cebola, cenoura, chuchu, couve, espinafre, pepino, tomate etc? Não inclua batata e aipim (mandioca/macaxeira)
VB02027	NOS ÚLTIMOS 30 DIAS, quantas vezes por dia você tomou refrigerante?
VB03001A1	NOS ÚLTIMOS 7 DIAS, em quantos dias você FOI a pé ou de bicicleta para a escola?
VB03002A1	Quando você VAI para a escola a pé ou de bicicleta, quanto tempo você gasta?
VB03001A2	NOS ÚLTIMOS 7 DIAS, em quantos dias você VOLTOU a pé ou de bicicleta da escola?
VB03002A2	Quando você VOLTA da escola a pé ou de bicicleta, quanto tempo você gasta?
VB03003A	NOS ÚLTIMOS 7 DIAS, quantos dias você teve aulas de educação física na escola?
VB03005A	NOS ÚLTIMOS 7 DIAS, quanto tempo por dia você fez atividade física ou esporte durante as aulas de educação física na escola?
VB03006A	NOS ÚLTIMOS 7 DIAS, sem contar as aulas de educação física da escola, em quantos dias você praticou alguma atividade física, como esportes, dança, ginástica, musculação, lutas ou outra atividade?

Variável	Descrição
VB03007	NORMALMENTE, quanto tempo por dia duram essas atividades (como esportes, dança, ginástica, musculação, lutas ou outra atividade) que você faz? (Sem contar as aulas de educação física)
VB03011A	NOS ÚLTIMOS 7 DIAS, em quantos dias você fez atividade física por pelo menos 60 minutos (1 hora) por dia? (Some todo o tempo que você gastou em qualquer tipo de atividade física EM CADA DIA)
TEMPODESLOC	Refere-se ao tempo médio diário acumulado pelo escolar, com o deslocamento da casa para escola e da escola para casa feito a pé ou de bicicleta, nos últimos sete dias anteriores à pesquisa. Em minutos.
TEMPOEDFIS	Refere-se ao tempo médio acumulado, nos últimos sete dias anteriores à pesquisa, que o escolar fez atividade física ou esporte durante as aulas de educação física na escola. Em minutos.
TEMPOEXTRA	Refere-se ao tempo médio diário acumulado pelo escolar com a prática de alguma atividade física extraescolar como esportes, dança, ginástica, musculação, lutas ou outra atividade, nos últimos sete dias anteriores à data da pesquisa. Em minutos
TEMPOTOTAL	A atividade física acumulada foi estimada calculando o produto entre o número de dias e o tempo médio que os escolares gastam em atividades físicas, nos sete dias anteriores à pesquisa, considerando os seguintes domínios: ir e voltar da escola, aulas de educação física e outras atividades extraescolares. Em minutos.
TEMPOEST	A atividade física globalmente estimada refere-se ao número de dias que os escolares declararam fazer, pelo menos, uma hora por dia de atividade física, nos sete dias anteriores à pesquisa. Em minutos.
VB03008	Se você tivesse oportunidade de fazer atividade física na maioria dos dias da semana, qual seria a sua atitude?
VB03009A	Em um dia de semana comum, quantas horas por dia você assiste a TV? (não contar sábado, domingo e feriado)
VB03010A	Em um dia de semana comum, quanto tempo você fica sentado(a), assistindo televisão, usando computador, jogando videogame, conversando com amigos(as) ou fazendo outras atividades sentado(a)? (não contar sábado, domingo, feriados e o tempo sentado na escola)

Variável	Descrição
VB04001	Alguma vez na vida, você já fumou cigarro, mesmo uma ou duas tragadas?
VB04002	Que idade você tinha quando experimentou fumar cigarro pela primeira vez?
VB04003	NOS ÚLTIMOS 30 DIAS, em quantos dias você fumou cigarros?
VB04009	NOS ÚLTIMOS 30 DIAS, em geral, como você conseguiu seus próprios cigarros?
VB04010	NOS ÚLTIMOS 30 DIAS, alguém se recusou a lhe vender cigarros por causa de sua idade?
VB04008A	NOS ÚLTIMOS 30 DIAS, em quantos dias você usou outros produtos de tabaco: cigarros de palha ou enrolados a mão, charuto, cachimbo, cigarrilha, cigarro indiano ou bali, narguilé, rapé, fumo de mascar etc? (não incluir cigarro comum)
VB04011	Qual outro produto do tabaco você usou com mais frequência NOS ÚLTIMOS 30 DIAS?
VB04005	NOS ÚLTIMOS 7 DIAS, em quantos dias pessoas fumaram na sua presença?
VB04006A	Algum de seus pais ou responsáveis fuma?
VB05002	Alguma vez na vida você tomou uma dose de bebida alcoólica? (Uma dose equivale a uma lata de cerveja ou uma taça de vinho ou uma dose de cachaça ou uísque etc)
VB05003	Que idade você tinha quando tomou a primeira dose de bebida alcoólica? (Uma dose equivale a uma lata de cerveja ou uma taça de vinho ou uma dose de cachaça ou uísque etc)
VB05004	NOS ÚLTIMOS 30 DIAS, em quantos dias você tomou pelo menos um copo ou uma dose de bebida alcoólica? (Uma dose equivale a uma lata de cerveja ou uma taça de vinho ou uma dose de cachaça ou uísque etc)
VB05005	NOS ÚLTIMOS 30 DIAS, nos dias em que você tomou alguma bebida alcoólica, quantos copos ou doses você tomou por dia?
VB05006A	NOS ÚLTIMOS 30 DIAS, na maioria das vezes, como você conseguiu a bebida que tomou?
VB05007	Na sua vida, quantas vezes você bebeu tanto que ficou realmente bêbado(a)?
VB05009	Na sua vida, quantas vezes você teve problemas com sua família ou amigos, perdeu aulas ou brigou por que tinha bebido?
VB05010	Quantos amigos seus consomem bebida alcoólica?

Variável	Descrição
VB06001	Alguma vez na vida, você já usou alguma droga-como: maconha, cocaína, crack, loló, lança-perfume, ecstasy, oxy, etc?
VB06002	Que idade você tinha quando usou droga como maconha, cocaína, crack, cola, loló, lança-perfume, ecstasy, oxy ou outra, pela primeira vez?
VB06003A	NOS ÚLTIMOS 30 DIAS, quantos dias você usou droga como maconha, cocaína, crack, cola, loló, lança-perfume, ecstasy, oxy, etc?
VB06004A	NOS ÚLTIMOS 30 DIAS, quantos dias você usou maconha?
VB06005A	NOS ÚLTIMOS 30 DIAS, quantos dias você usou crack?
VB06006	Quantos amigos seus usam drogas?
VB07001	NOS ÚLTIMOS 30 DIAS, em quantos dias você faltou às aulas ou à escola sem permissão dos seus pais ou responsáveis?
VB07002	NOS ÚLTIMOS 30 DIAS, com que frequência seus pais ou responsáveis sabiam realmente o que você estava fazendo em seu tempo livre?
VB07003	NOS ÚLTIMOS 30 DIAS, com que frequência seus pais ou responsáveis verificaram se os seus deveres de casa (lição de casa) foram feitos?
VB07004	NOS ÚLTIMOS 30 DIAS, com que frequência seus pais ou responsáveis entenderam seus problemas e preocupações?
VB07005	NOS ÚLTIMOS 30 DIAS, com que frequência seus pais ou responsáveis mexeram em suas coisas sem a sua concordância?
VB07006	NOS ÚLTIMOS 30 DIAS, com que frequência os colegas de sua escola trataram você bem e/ou foram prestativos contigo?
VB07007	NOS ÚLTIMOS 30 DIAS, com que frequência algum dos seus colegas de escola te esculacharam, zoaram, mangaram, intimidaram ou caçoaram tanto que você ficou magoado, incomodado, aborrecido, ofendido ou humilhado?
VB07008	NOS ÚLTIMOS 30 DIAS, qual o motivo/causa de seus colegas terem te esculachado, zombado, zoado, caçoado, mangado, intimidado ou humilhado?
VB07009	NOS ÚLTIMOS 30 DIAS, você esculachou, zombou, mangou, intimidou ou caçoou algum de seus colegas da escola tanto que ele ficou magoado, aborrecido, ofendido ou humilhado?
VB07010	Você já sofreu bullying?

Variável	Descrição
VB12001	NOS ÚLTIMOS 12 MESES com que frequência tem se sentido sozinho(a)?
VB12002	NOS ÚLTIMOS 12 MESES, com que frequência você não conseguiu dormir à noite porque algo o(a) preocupava muito?
VB12003	Quantos amigos(as) próximos você tem?
VB08001	Você já teve relação sexual (transou) alguma vez?
VB08002	Que idade você tinha quando teve relação sexual (transou) pela primeira vez?
VB08011	Você usou preservativo na primeira relação sexual?
VB08003A	Na sua vida, com quantas pessoas você teve relações sexuais (transou)?
VB08005	Na última vez que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou algum método para evitar a gravidez e/ou Doenças Sexualmente Transmissíveis (DST)
VB08006	Na última vez que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou camisinha (preservativo)?
VB08007	Na última vez que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou algum outro método para evitar a gravidez (não contar camisinha)?T)
VB08012	Nesta última vez que você teve relação sexual (transou), qual outro método para evitar gravidez você ou sua parceira usou?
VB08013	Alguma vez na vida você engravidou?
VB08008	Na escola, você já recebeu orientação sobre prevenção de gravidez?
VB08009	Na escola, você já recebeu orientação sobre AIDS ou outras Doenças Sexualmente Transmissíveis (DST)?
VB08010	Na escola, você já recebeu orientação sobre como conseguir camisinha (preservativo) gratuitamente?
VB10004	NOS ÚLTIMOS 30 DIAS, com que frequência você lavou as mãos antes de comer?
VB10005	NOS ÚLTIMOS 30 DIAS, com que frequência você lavou as mãos após usar o banheiro ou o vaso sanitário?
VB10006	NOS ÚLTIMOS 30 DIAS, com que frequência você usou sabão ou sabonete quando lavou suas mãos?
VB10001A	NOS ÚLTIMOS 30 DIAS, quantas vezes por dia você usualmente escovou os dentes?

Variável	Descrição
VB10002	NOS ÚLTIMOS 6 MESES, você teve dor de dente? (excluir dor de dente causada por uso de aparelho)
VB10003	NOS ÚLTIMOS 12 MESES, quantas vezes você foi ao dentista?
VB09001	NOS ÚLTIMOS 30 DIAS, em quantos dias você deixou de ir à escola porque não se sentia seguro no caminho de casa para a escola ou da escola para casa?
VB09002	NOS ÚLTIMOS 30 DIAS, em quantos dias você não foi à escola porque não se sentia seguro na escola?
VB09006A1	NOS ÚLTIMOS 30 DIAS, com que frequência você usou cinto de segurança enquanto andava como passageiro(a) no BANCO DA FRENTE de carro/automóvel, van ou táxi?
VB09006A2	NOS ÚLTIMOS 30 DIAS, com que frequência você usou cinto de segurança enquanto andava como passageiro(a) no BANCO DE TRÁS de carro/automóvel, van ou táxi?
VB09007A	NOS ÚLTIMOS 30 DIAS, com que frequência você usou capacete ao andar de motocicleta?
VB09008	NOS ÚLTIMOS 30 DIAS, quantas vezes você dirigiu um veículo motorizado de transporte (carro, motocicleta, voadeira, barco)?
VB09009	NOS ÚLTIMOS 30 DIAS, quantas vezes você andou em carro ou outro veículo motorizado dirigido por alguém que tinha consumido alguma bebida alcoólica?
VB09003	NOS ÚLTIMOS 30 DIAS, quantas vezes você foi agredido(a) fisicamente por um adulto da sua família?
VB09004	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou arma de fogo, como revólver ou espingarda?
VB09005	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou alguma outra arma como faca, canivete, peixeira, pedra, pedaço de pau ou garrafa?
VB09010	NOS ÚLTIMOS 12 MESES, quantas vezes você foi agredido(a) fisicamente?
VB09011	NOS ÚLTIMOS 12 MESES, quantas vezes você se envolveu em briga (uma luta física)?
VB09012	NOS ÚLTIMOS 12 MESES, quantas vezes você foi seriamente ferido(a)?
VB09013A	NOS ÚLTIMOS 12 MESES, qual foi o ferimento ou a lesão mais séria que aconteceu com você?

Variável	Descrição
VB09014A	NOS ÚLTIMOS 12 MESES, qual foi a principal causa do ferimento ou da lesão mais séria que aconteceu com você?
VB09015	NOS ÚLTIMOS 12 MESES, você sofreu algum acidente de bicicleta (caiu e se machucou)?
VB09016	Alguma vez na vida você foi forçado a ter relação sexual?
VB0901701	Um(a) namorado(a)/ex-namorado(a) forçou você a ter relação sexual?
VB0901702	Um(a) amigo(a) forçou você a ter relação sexual?
VB0901703	Seu pai/mãe/padastro/madrasta forçou você a ter relação sexual?
VB0901704	Outros familiares forçaram você a ter relação sexual?
VB0901705	Um(a) desconhecido(a) forçou você a ter relação sexual?
VB0901706	Outras pessoas forçaram você a ter relação sexual?
VB13005	Como você classificaria seu estado de saúde?
VB13006	NOS ÚLTIMOS 12 MESES, quantos dias você faltou a escola por motivo(s) relacionado(s) à sua saúde?
VB13001	NOS ÚLTIMOS 12 MESES você procurou algum serviço ou profissional de saúde para atendimento relacionado à própria saúde?
VB13002A	NOS ÚLTIMOS 12 MESES, qual foi o serviço de saúde que você procurou com MAIS FREQUÊNCIA?
VB13004A	Você foi atendido NA ÚLTIMA VEZ que procurou alguma Unidade Básica de Saúde (Centro ou Posto de saúde ou Unidade de Saúde da Família/PSF), NESTES ÚLTIMOS 12 MESES?
VB13007	Qual foi o PRINCIPAL MOTIVO da sua procura na Unidade Básica de Saúde (Centro ou Posto de saúde ou Unidade de Saúde da Família/PSF) NESTA ÚLTIMA VEZ?
VB13008	Você conhece/ouviu falar sobre a campanha de vacinação contra o vírus HPV?
VB13009	Você foi vacinada contra o vírus HPV?
VB14001	NOS ÚLTIMOS 12 MESES, você teve chiado (ou piado) no peito?
VB14002	Você teve asma alguma vez na vida?
VB11006	Você considera sua imagem corporal como sendo algo:
VB11007	Como você se sente em relação ao seu corpo?
VB11001	Quanto ao seu corpo, você se considera:
VB11002	O que você está fazendo em relação a seu peso?

Variável	Descrição
VB11003	NOS ÚLTIMOS 30 DIAS, você vomitou ou tomou laxantes para perder peso ou evitar ganhar peso?
VB11004A	NOS ÚLTIMOS 30 DIAS, você tomou algum remédio, fórmula ou outro produto para perder peso, sem acompanhamento médico?
VB11005	NOS ÚLTIMOS 30 DIAS, você tomou algum remédio, fórmula ou outro produto para ganhar peso ou massa muscular sem acompanhamento médico?
VB16001A01	Você achou este questionário fácil?
VB16001A02	Você achou este questionário difícil?
VB16001A03	Você achou este questionário chato?
VB16001A04	Você achou este questionário legal?
VB16001A05	Você achou este questionário interessante?
VB16001A06	Você achou este questionário informativo?
VB16001A07	Você achou este questionário cansativo?
VB16001A08	Você achou este questionário constrangedor?
ESTRATOGEOREG	Indicador de estrato georeg
ESTRATO_EXP	Expressão do estrato
PESO	Peso do aluno de acordo com a amostra, utilizado para expansão
aluno	contador de aluno
escola	UPA (unidade primária de amostragem)
turma	USA (unidade secundária de amostragem)
V0007	Esfera Administrativa
V0008	Esfera Administrativa da escola
V0006	Situação da Escola
V0041	Tipo de Escola Privada

APÊNDICE C – Descrição das Variáveis utilizadas PeNSE 2019

Ressalta-se que as fontes contendo o dicionário das respostas das variáveis encontram-se no Apêndice E.

Variável	**Descrição**
B00004	Você concorda em participar dessa pesquisa?
B01001A	Qual é o seu sexo?
B01003	Qual é a sua idade?
B01004	Qual é o mês do seu aniversário?
B01005	Em que ano você nasceu?
B01002	Qual é a sua cor ou raça?
B01021A	Em que ano escolar você está?
B01026A1	Quando terminar o Ensino Fundamental, você pretende?
B01026A2	Quando terminar o Ensino Médio, você pretende?
B01006	Você mora com sua mãe?
B01007	Você mora com seu pai?
B01010A	CONTANDO COM VOCÊ, quantas pessoas moram na sua casa ou apartamento?
B01014	Você tem celular?
B01015B	Na sua casa tem computador ou notebook?
B01016	Você tem acesso à internet em sua casa?
B01017	Alguém que mora na sua casa tem carro?
B01018A	Alguém que mora na sua casa tem motocicleta/moto?
B01019A	Quantos banheiros completos, com vaso sanitário e chuveiro, têm dentro da sua casa?
B01020A	Tem empregado(a) doméstico(a) recebendo dinheiro para fazer o trabalho em sua casa, três ou mais dias por semana?
B01008B	Qual nível de ensino (grau) sua MÃE estudou ou estuda?
B02019A	Você costuma tomar o café da manhã?
B02017A	Você costuma almoçar ou jantar com sua mãe, pai ou responsável?

Variável	Descrição
B02018B	Nas suas refeições, com que frequência você costuma comer fazendo alguma outra coisa (assistindo à TV, mexendo no computador ou no celular)?
B02028	ONTEM, você tomou refrigerante?
B02029	ONTEM, você tomou suco de fruta em caixinha ou lata?
B02030	ONTEM, você tomou refresco em pó?
B02031	ONTEM, você tomou bebida achocolatada?
B02032	ONTEM, você tomou iogurte com sabor?
B02033	ONTEM, você comeu salgadinho de pacote (chips) ou biscoito/bolacha salgado?
B02034	ONTEM, você comeu biscoito ou bolacha doce, biscoito recheado ou bolinho de pacote?
B02035	ONTEM, você comeu chocolate, sorvete, gelatina, flan ou outra sobremesa industrializada?
B02036	ONTEM, você comeu salsicha, linguiça, mortadela ou presunto?
B02037	ONTEM, você comeu pão de forma, pão de cachorro-quente ou pão de hambúrguer?
B02038	ONTEM, você comeu margarina?
B02039	ONTEM, você comeu maionese, ketchup ou outros molhos industrializados?
B02040	ONTEM, você comeu macarrão instantâneo (miojo), sopa de pacote, lasanha congelada ou outro prato pronto comprado congelado?
B02001	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu feijão?
B02004B	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu pelo menos um tipo de legume ou verdura que não seja batata ou aipim (mandioca/macaxeira)?
B02010A	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu guloseimas doces, como balas, confeitos, chocolates, chicletes, bombons, pirulitos e outros?
B02011	NOS ÚLTIMOS 7 DIAS, em quantos dias você comeu frutas frescas ou salada de frutas?
B02013	NOS ÚLTIMOS 7 DIAS, em quantos dias você tomou refrigerante?
B02023A	NOS ÚLTIMOS 7 DIAS, em quantos deles você comeu em lanchonetes, barracas de cachorro quente, pizzaria, fast food etc?

Variável	Descrição
B02021A	Sua escola oferece comida/merenda aos alunos da sua turma? (Não considerar comida comprada na cantina)
B02020B	Você costuma comer a comida/merenda oferecida pela escola? (Não considerar comida comprada na cantina)
B02041	Você costuma comprar alimentos ou bebidas na cantina dentro da escola? (Não considerar a compra de água)
B02042	Você costuma comprar alimentos ou bebidas de vendedores de rua (camelô ou ambulante) na porta ou ao redor da escola? (Não considerar a compra de água)
B03001A1	NOS ÚLTIMOS 7 DIAS, em quantos dias você FOI a pé ou de bicicleta para a escola?
B03002A1	Quando você VAI para a escola a pé ou de bicicleta, quanto tempo você gasta?
B03001A2	NOS ÚLTIMOS 7 DIAS, em quantos dias você VOLTOU a pé ou de bicicleta da escola?
B03002A2	Quando você VOLTA da escola a pé ou de bicicleta, quanto tempo você gasta?
B03003A	NOS ÚLTIMOS 7 DIAS, quantos dias você TEVE aulas de educação física na escola?
B03005B	Quanto tempo por dia você FEZ atividade física ou praticou esporte durante as aulas de educação física na escola? Não considere o tempo gasto em atividades teóricas em sala de aula.
B03006B	NOS ÚLTIMOS 7 DIAS, sem contar as aulas de educação física da escola, em quantos dias você praticou alguma atividade física?
B03007A	Quanto tempo por dia duraram essas atividades que você fez?
B03009B	Quantas horas por dia você assiste a televisão (TV)? (NÃO contar sábado, domingo e feriado)
B03010B	Quantas horas por dia você costuma ficar sentado(a), assistindo televisão, jogando videogame, usando computador, celular, tablet ou fazendo outras atividades sentado(a)? (NÃO contar sábado, domingo, feriados ou o tempo sentado na escola)
B04001	Alguma vez na vida, você já fumou cigarro, mesmo uma ou duas tragadas?
B04002A	Que idade você tinha quando fumou cigarro pela primeira vez?
B04003	NOS ÚLTIMOS 30 DIAS, em quantos dias você fumou cigarros?

Variável	Descrição
B04009A	NOS ÚLTIMOS 30 DIAS, na maioria das vezes, como você conseguiu seus próprios cigarros?
B04010	NOS ÚLTIMOS 30 DIAS, alguém se recusou a lhe vender cigarros por causa de sua idade?
B04012	NOS ÚLTIMOS 30 DIAS, você comprou cigarro por unidade (avulso, a varejo, retalho ou cigarro solto)?
B04013	Alguma vez na vida você já experimentou narguilé (cachimbo de água)?
B04014	Alguma vez na vida você já experimentou cigarro eletrônico (e-cigarrette)?
B04015	Alguma vez na vida você já experimentou outros produtos do tabaco, SEM CONTAR narguilé e cigarro eletrônico?
B04011A01	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Não usei nenhum desses outros produtos de tabaco nos últimos 30 dias
B04011A02	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Narguilé (cachimbo de água)
B04011A03	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Cigarro eletrônico (e-cigarette)
B04011A04	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Cigarros de cravo (cigarros de Bali)
B04011A05	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Cigarros enrolados à mão (palha ou papel)
B04011A06	NOS ÚLTIMOS 30 DIAS, qual(is) desses outros produtos do tabaco você usou? - Outros
B04006B	Sua mãe, pai ou responsável fuma?
B04005A	NOS ÚLTIMOS 7 DIAS, em quantos dias pessoas fumaram em sua presença na sua casa?
B04016	NOS ÚLTIMOS 30 DIAS, algum dos seus amigos fumou na sua presença?
B05002A	Alguma vez na vida você tomou um copo ou uma dose de bebida alcoólica?
B05003A	Que idade você tinha quando tomou o primeiro copo ou dose de bebida alcoólica?
B05007	Na sua vida, quantas vezes você bebeu tanto que ficou realmente bêbado(a)?

Variável	Descrição
B05009	Na sua vida, quantas vezes você teve problemas com sua família ou amigos, perdeu aulas ou brigou por que tinha bebido?
B05004A	NOS ÚLTIMOS 30 DIAS, em quantos dias você tomou pelo menos um copo ou uma dose de bebida alcoólica?
B05005A	NOS ÚLTIMOS 30 DIAS, nos dias em que você tomou alguma bebida alcoólica, quantos copos ou doses você tomou por dia?
B05006B	NOS ÚLTIMOS 30 DIAS, na maioria das vezes, como você conseguiu a bebida que tomou?
B05011	Sua mãe, pai ou responsável bebe bebidas alcoólicas?
B05010A	NOS ÚLTIMOS 30 DIAS, algum dos seus amigos bebeu alguma bebida alcoólica na sua presença?
B06001	Alguma vez na vida, você já usou alguma droga como: maconha, cocaína, crack, cola, loló, lança-perfume, ecstasy, oxi, MD, skank e outras?
B06002A	Que idade você tinha quando usou alguma droga pela primeira vez?
B06003B	NOS ÚLTIMOS 30 DIAS, quantos dias você usou alguma droga?
B06004A	NOS ÚLTIMOS 30 DIAS, quantos dias você usou maconha?
B06005A	NOS ÚLTIMOS 30 DIAS, quantos dias você usou crack?
B06006A	NOS ÚLTIMOS 30 DIAS, algum dos seus amigos usou drogas na sua presença?
B07001	NOS ÚLTIMOS 30 DIAS, em quantos dias você faltou às aulas ou à escola sem permissão de sua mãe, pai ou responsável?
B07002	NOS ÚLTIMOS 30 DIAS, com que frequência sua mãe, pai ou responsável sabia realmente o que você estava fazendo em seu tempo livre?
B07004	NOS ÚLTIMOS 30 DIAS, com que frequência sua mãe, pai ou responsável entendeu seus problemas e preocupações?
B07006	NOS ÚLTIMOS 30 DIAS, com que frequência os colegas de sua escola trataram você bem e/ou foram prestativos com você?
B07007A	NOS ÚLTIMOS 30 DIAS, quantas vezes algum dos seus colegas de escola o esculachou, zoou, mangou, intimidou ou caçoou tanto que você ficou magoado, incomodado, aborrecido, ofendido ou humilhado?
B07008	NOS ÚLTIMOS 30 DIAS, qual o motivo/causa de seus colegas terem esculachado, zombado, zoado, caçoado, mangado, intimidado ou humilhado?

Variável	Descrição
B07011	NOS ÚLTIMOS 30 DIAS, quantas vezes algum dos seus colegas de escola se recusou a falar com você, deixou você de lado sem razão ou fez com que outros colegas deixassem de falar com você?
B07012	NOS ÚLTIMOS 30 DIAS, quantas vezes algum dos seus colegas de escola bateu (deu socos, tapas, chutes, pontapés) em você ou o machucou fisicamente de outra forma?
B07013	NOS ÚLTIMOS 30 DIAS, você se sentiu ameaçado(a), ofendido(a) ou humilhado(a) nas redes sociais ou aplicativos de celular?
B07009	NOS ÚLTIMOS 30 DIAS, você esculachou, zombou, mangou, intimidou ou caçoou algum de seus colegas da escola tanto que ele ficou magoado, aborrecido, ofendido ou humilhado?
B12003	Quantos(as) amigos(as) próximos você tem?
B12004	NOS ÚLTIMOS 30 DIAS, com que frequência você se sentiu muito preocupado com as coisas comuns do seu dia a dia como atividades da escola, competições esportivas, tarefas de casa, etc.?
B12005	NOS ÚLTIMOS 30 DIAS, com que frequência você se sentiu triste?
B12006	NOS ÚLTIMOS 30 DIAS, com que frequência você sentiu que ninguém se preocupa com você?
B12007	NOS ÚLTIMOS 30 DIAS, com que frequência você se sentiu irritado(a), nervoso(a) ou mal-humorado(a) por qualquer coisa?
B12008	NOS ÚLTIMOS 30 DIAS, com que frequência você sentiu que a vida não vale a pena ser vivida?
B08001	Você já teve relação sexual (transou) alguma vez?
B08002	Que idade você tinha quando teve relação sexual (transou) pela primeira vez?
B08011A	Você ou seu(sua) parceiro(a) usou camisinha (preservativo) NA PRIMEIRA RELAÇÃO SEXUAL?
B08006A	NA ÚLTIMA VEZ que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou camisinha (preservativo)?
B08014	Nesta última vez que você teve relação sexual (transou), como você conseguiu a camisinha (preservativo)?

Variável	Descrição
B08007	NA ÚLTIMA VEZ que você teve relação sexual (transou), você ou seu(sua) parceiro(a) usou algum outro método para evitar a gravidez que não seja camisinha (preservativo)?
B08012A	Nesta última vez que você teve relação sexual (transou), qual outro método você ou seu(sua) parceiro(a) usou para evitar gravidez?
B08015	Alguma vez na vida, você ou sua parceira já usou pílula do dia seguinte (contracepção de emergência)?
B08016	NA ÚLTIMA VEZ que você ou sua parceira usou pílula do dia seguinte (contracepção de emergência) como conseguiu?
B08013A	Alguma vez na vida você engravidou, mesmo que a gravidez não tenha chegado ao fim?
B08008A	Na escola, você já recebeu orientação sobre prevenção de gravidez?
B08009A	Na escola, você já recebeu orientação sobre prevenção de HIV/AIDS ou outras Doenças/Infecções Sexualmente Transmissíveis?
B08010A	Na escola, você já recebeu orientação sobre como conseguir camisinha (preservativo) gratuitamente?
B10004A	Com que frequência você lava as mãos antes de comer?
B10005A	Com que frequência você lava as mãos após usar o banheiro ou o vaso sanitário?
B10006A	Com que frequência você usa sabão ou sabonete quando lava suas mãos?
B10001B	Quantas vezes por dia você escova os dentes?
B10002	NOS ÚLTIMOS 6 MESES, você teve dor de dente que não tenha sido causada por uso de aparelho?
B10003	NOS ÚLTIMOS 12 MESES, quantas vezes você foi ao dentista?
B09006A1	NOS ÚLTIMOS 30 DIAS, com que frequência você usou cinto de segurança enquanto andava como passageiro(a) NO BANCO DA FRENTE de carro/automóvel, van ou táxi?
B09006A2	NOS ÚLTIMOS 30 DIAS, com que frequência você usou cinto de segurança enquanto andava como passageiro(a) NO BANCO DE TRÁS de carro/automóvel, van ou táxi?
B09007A	NOS ÚLTIMOS 30 DIAS, com que frequência você usou capacete ao andar de motocicleta/moto?

Variável	Descrição
B09008	NOS ÚLTIMOS 30 DIAS, quantas vezes você dirigiu um veículo motorizado de transporte (carro, motocicleta/moto, voadeira, barco)?
B09009	NOS ÚLTIMOS 30 DIAS, quantas vezes você andou em carro ou outro veículo motorizado dirigido por alguém que tinha consumido alguma bebida alcoólica?
B09019	NOS ÚLTIMOS 30 DIAS, quantas vezes você andou em carro ou outro veículo motorizado dirigido por alguém que usou o celular enquanto dirigia?
B09001	NOS ÚLTIMOS 30 DIAS, em quantos dias você deixou de ir à escola porque não se sentia seguro NO CAMINHO de casa para a escola ou da escola para casa?
B09002	NOS ÚLTIMOS 30 DIAS, em quantos dias você não foi à escola porque não se sentia seguro NA ESCOLA?
B09011A	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em briga com luta física?
B09004	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou arma de fogo, como revólver ou espingarda?
B09005	NOS ÚLTIMOS 30 DIAS, você esteve envolvido(a) em alguma briga em que alguma pessoa usou alguma outra arma como faca, canivete, peixeira, pedra, pedaço de pau ou garrafa?
B09012A1	NOS ÚLTIMOS 12 MESES, você sofreu algum acidente ou agressão?
B09012A2	Algum desse(s) acidente(s) ou agressão(ões) que você sofreu o(a) impediu de realizar atividades habituais (ir para a escola, trabalhar, realizar afazeres domésticos etc.)?
B09012A3	Você teve que procurar algum serviço de saúde (Pronto-socorro, emergência ou UPA, hospital, farmácia) em razão deste acidente ou agressão?
B09013B	Qual foi o ferimento ou a lesão MAIS GRAVE que você sofreu nesse acidente ou agressão?
B09014B	Qual foi a PRINCIPAL CAUSA do ferimento ou da lesão mais grave que aconteceu com você?
B09003A	NOS ÚLTIMOS 12 MESES, quantas vezes você foi agredido(a) fisicamente por sua mãe, pai ou responsável?

Variável	Descrição
B09010A1	NOS ÚLTIMOS 12 MESES, quantas vezes você foi agredido(a) fisicamente por OUTRA PESSOA que não seja sua mãe, pai ou responsável?
B09010A201	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Namorado(a), ex-namorado(a), ficante, crush
B09010A202	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Amigo(a)
B09010A203	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Outros familiares
B09010A204	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Policial
B09010A205	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Profissionais da sua escola (professor, diretor, inspetor, etc.)
B09010A206	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Desconhecido(a)
B09010A207	Quem o(a) agrediu fisicamente? (Sem contar sua mãe, pai ou responsável) - Outro
B09016A1	Alguma vez na vida alguém o(a) tocou, manipulou, beijou ou expôs partes do corpo contra a sua vontade?
B09017A101	Quem fez isso? - Namorado(a), ex-namorado(a), ficante, crush
B09017A102	Quem fez isso? - Amigo(a)
B09017A103	Quem fez isso? - Pai/mãe/padrasto/madrasta
B09017A104	Quem fez isso? - Outros familiares
B09017A105	Quem fez isso? - Desconhecido(a)
B09017A106	Quem fez isso? - Outro
B09016A2	Alguma vez na vida alguém ameaçou, intimidou ou obrigou a ter relações sexuais ou qualquer outro ato sexual contra a sua vontade?
B09017A201	Quem fez isso? - Namorado(a), ex-namorado(a), ficante, crush
B09017A202	Quem fez isso? - Amigo(a)
B09017A203	Quem fez isso? - Pai/mãe/padrasto/madrasta
B09017A204	Quem fez isso? - Outros familiares
B09017A205	Quem fez isso? - Desconhecido(a)
B09017A206	Quem fez isso? - Outro

Variável	Descrição
B09018	Que idade você tinha quando alguém ameaçou, intimidou ou obrigou a ter relações sexuais ou qualquer outro ato sexual contra a sua vontade pela primeira vez?
B13005	Como você classificaria seu estado de saúde?
B13006	NOS ÚLTIMOS 12 MESES, quantos dias você faltou a escola por motivo(s) relacionado(s) à própria saúde?
B13001	NOS ÚLTIMOS 12 MESES você procurou algum serviço ou profissional de saúde para atendimento relacionado à própria saúde?
B13002A	NOS ÚLTIMOS 12 MESES, qual foi o serviço de saúde que você procurou com MAIS FREQUÊNCIA?
B13003A	NOS ÚLTIMOS 12 MESES, quantas vezes você procurou por alguma Unidade Básica de Saúde (Centro ou Posto de saúde ou Unidade de Saúde da Família/PSF)?
B13004B	Você foi atendido NA ÚLTIMA VEZ que procurou alguma Unidade Básica de Saúde (Centro ou Posto de saúde ou Unidade de Saúde da Família/PSF)?
B13007A	Qual foi o PRINCIPAL MOTIVO da sua procura na Unidade Básica de Saúde (Centro ou Posto de saúde ou Unidade de Saúde da Família/PSF) NESTA ÚLTIMA VEZ?
B13009A	Você foi vacinado(a) contra o vírus HPV?
B13012	Por que você não foi vacinado(a) contra o vírus HPV?
B11007	Como você se sente em relação ao seu corpo?
B11001	Quanto ao seu corpo, você se considera:
B11002	O que você está fazendo em relação a seu peso?
B11003	NOS ÚLTIMOS 30 DIAS, você vomitou ou tomou laxantes para perder peso ou evitar ganhar peso?
B11004A	NOS ÚLTIMOS 30 DIAS, você tomou algum remédio, fórmula ou outro produto para perder peso, sem acompanhamento médico?
B11005	NOS ÚLTIMOS 30 DIAS, você tomou algum remédio, suplemento ou outro produto para ganhar peso ou massa muscular sem acompanhamento médico?
B16001A01	O que você achou deste questionário? - Fácil
B16001A02	O que você achou deste questionário? - Difícil
B16001A03	O que você achou deste questionário? - Chato
B16001A04	O que você achou deste questionário? - Legal
B16001A05	O que você achou deste questionário? - Interessante

Variável	**Descrição**
B16001A06	O que você achou deste questionário? - Informativo
B16001A07	O que você achou deste questionário? - Cansativo
B16001A08	O que você achou deste questionário? - Constrangedor
TEMPODESLOC	Tempo semanal em minutos de deslocamento entre casa e escola
TEMPOEDFIS	Tempo semanal em minutos de atividade física na aula de Educação Física
TEMPOEXTRA	Tempo semanal em minutos de atividade física sem ser na aula de Educação Física
TEMPOTOTAL	Tempo semanal em minutos de atividade física acumulada
UF	Qual seu estado?
ESFERA	Qual a esfera da escola?

APÊNDICE D – Código fonte

O código-fonte completo utilizado para a execução das etapas descritas nesta dissertação está disponível no repositório público do GitHub, acessível pelo link abaixo:

`https://github.com/guilhermecamboim/pense-clustering-analysis`

APÊNDICE E – Fontes e dados

Os dados completos da PeNSE, incluindo os dicionários utilizados para a interpretação dos resultados, estão disponíveis nos endereços indicados a seguir. Conforme mencionado na Seção 3.2, para o ano de 2015 foi utilizada a amostra 1, juntamente com o respectivo dicionário. Por fim, apresenta-se o link do INEP que disponibiliza o painel do IDEB no Power BI, utilizado para a obtenção dos indicadores de cada ano, organizados por estado e esfera administrativa.

PeNSE 2009: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=pense/2009/microdados

PeNSE 2015: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=pense/2015/microdados

PeNSE 2019: https://www.ibge.gov.br/estatisticas/downloads-estatisticas.html?caminho=pense/2019/microdados

IDEB: https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/ideb/resultados