Universidade Federal do Pará - UFPA
Instituto de Tecnologia - ITEC
Programa de Pós-Graduação em Engenharia Elétrica - PPGEE

Sandio Maciel dos Santos

Ciência de dados para análise da desigualdade social durante a pandemia: Uma análise com dados públicos brasileiros.

TD 08/2025

Belém - Pará - Brasil
2025

Sandio Maciel dos Santos

Ciência de dados para análise da desigualdade social durante a pandemia: Uma análise com dados públicos brasileiros.

Tese de doutorado submetida à Banca Examinadora do Programa de Pós-Graduação em Engenharia Elétrica da Universidade Federal do Pará, como requisito para a obtenção do título de Doutor em Engenharia Elétrica, na área de concentração em Computação Aplicada.

Belém - Pará - Brasil

2025

Dados Internacionais de Catalogação na Publicação (CIP) de acordo com ISBD
Sistema de Bibliotecas da Universidade Federal do Pará
Gerada automaticamente pelo módulo Ficat, mediante os dados fornecidos pelo(a) autor(a)

S237c Santos, Sandio Maciel dos.
Ciência de dados para análise da desigualdade social durante a pandemia : Uma análise com dados públicos brasileiros / Sandio Maciel dos Santos, . — 2025.
XX, 120 f. : il. color.

Orientador(a): Prof. Dr. Marcelino Silva da Silva
Tese (Doutorado) - Universidade Federal do Pará, Instituto de Tecnologia, Programa de Pós-Graduação em Engenharia Elétrica, Belém, 2025.

1. Base de Dados Aberta. 2. COVID-19. 3. Demografia Social. 4. Vulnerabilidade Social. 5. Mineração de dados. I. Título.

CDD 005

UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA

"CIÊNCIA DE DADOS PARA ANÁLISE DA DESIGUALDADE SOCIAL DURANE A PANDEMIA: UMA ANÁLISE COM DADOS PÚBLICOS BRASILEIROS"

AUTOR: **SÂNDIO MACIEL DOS SANTOS**

TESE DE DOUTORADO SUBMETIDA À BANCA EXAMINADORA APROVADA PELO COLEGIADO DO PROGRAMA DE PÓS-GRADUAÇÃO EM ENGENHARIA ELÉTRICA, SENDO JULGADA ADEQUADA PARA A OBTENÇÃO DO GRAU DE DOUTOR EM ENGENHARIA ELÉTRICA NA ÁREA DE COMPUTAÇÃO APLICADA.

APROVADA EM: 17/04/2025

BANCA EXAMINADORA:

Prof. Dr. Marcelino Silva da Silva
(Orientador – PPGEE/ITEC/UFPA)

Prof. Dr. Carlos Renato Lisboa Francês
(Avaliador Interno - PPGEE/ITEC/UFPA)

Prof.ª Dr.ª Jasmine Priscyla Leite de Araújo
(Avaliadora Interna - PPGEE/ITEC/UFPA)

Prof. Dr. Fábio Manoel França Lobato
(Avaliador Externo - UFOPA)

Prof. Dr. Renato da Silva Bandeira
(Avaliador Externo - UFOPA)

VISTO:

Prof. Dr. Diego Lisboa Cardoso

(Coordenador do PPGEE/ITEC/UFPA)

A minha Mãe
Sônia Maria Maciel dos Santos e Familiares
por sempre estarem comigo em todos os momentos.

Agradecimentos

Primeiramente, agradeço a Deus, que, durante toda a minha vida, foi um Pai de amor incalculável e proteção imensurável, concedendo-me inúmeras conquistas e graças.

Expresso também minha profunda gratidão à minha mãe, Sônia Maria, que teve um papel fundamental nesta conquista e tem sido um exemplo de determinação e dedicação em todas as áreas da minha vida, por meio de seus ensinamentos, conselhos e valores morais.

Aos professores do PPGEE e da UFOPA, agradeço pelos ensinamentos compartilhados ao longo desses dois anos de pós-graduação, em especial ao Prof. Dr. Renato Frances, ao Prof. Dr. Fábio Lobato e ao meu orientador, Prof. Dr. Marcelino Silva da Silva, por aceitarem o grande desafio de me orientar, pelo conhecimento transmitido e pela paciência demonstrada ao longo dessa jornada.

Agradeço ao CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico) e à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) por financiarem minha pesquisa por meio de bolsa de estudos.

Sou grato à Universidade Federal do Pará (UFPA) e ao Programa de Pós-Graduação em Engenharia Elétrica (PPGEE/UFPA) pelo apoio institucional, acadêmico e científico, indispensável para a realização desta pesquisa.

Por fim, agradeço novamente ao CNPq pelo suporte financeiro concedido por meio da bolsa de doutorado no âmbito da Chamada Pública nº 01/2019 — Apoio à Formação de Doutores em Áreas Estratégicas (Processo nº 142080/2020-9), e à Universidade Federal do Pará (UFPA), por meio do PPGEE, pela oportunidade acadêmica e profissional de aprendizado e por todo o apoio recebido.

"Mas, sejam fortes e não desanimem,
pois o trabalho de vocês será recompensado."
(2 Crônicas 15:7)

Resumo

O processo de análise de dados tem sido amplamente utilizado em pesquisas nas mais diversas áreas do conhecimento, incluindo estudos demográficos e investigações sobre desigualdade social durante a pandemia de COVID-19. Entre as abordagens recorrentes, destacam-se aquelas baseadas em dados públicos disponibilizados por órgãos governamentais. No Brasil, informações relevantes sobre educação, economia, saúde e aspectos sociodemográficos são acessíveis à sociedade em geral. Nesse contexto, este trabalho utiliza técnicas de ciência de dados para analisar informações sociodemográficas e educacionais brasileiras, com o objetivo de mensurar as relações de desigualdade durante a pandemia de COVID-19 e fornecer subsídios técnicos para a formulação de políticas públicas. A pesquisa foi guiada por duas hipóteses principais: (1) a desigualdade e a vulnerabilidade social desempenham um papel crucial no agravamento dos surtos epidêmicos; e (2) as classes sociais mais baixas são as mais afetadas por essas crises, em razão da falta de acesso a serviços básicos de qualidade, como saúde e educação. Com base nessas hipóteses, foram conduzidos dois estudos de caso fundamentados em dados públicos. Os resultados do primeiro estudo indicam que alunos de escolas públicas em situação de vulnerabilidade socioeconômica apresentam maior probabilidade de não realizarem o ENEM em contextos de suspensão institucional do exame, como durante estados de calamidade pública. No segundo estudo, observou-se que áreas de baixa renda e com infraestrutura urbana precária são mais suscetíveis ao aumento de casos confirmados e mortes por COVID-19. Adicionalmente, o número de moradores por residência demonstrou estar associado ao aumento da transmissão de doenças respiratórias, reforçando o papel da vulnerabilidade social como fator amplificador dos efeitos da pandemia. O primeiro estudo analisa os efeitos do fechamento das escolas na participação dos estudantes no Exame Nacional do Ensino Médio (ENEM) durante a pandemia, entre os anos de 2020 e 2022. O segundo avalia o impacto das condições de infraestrutura educacional, urbana e econômica da população brasileira por meio do Índice de Vulnerabilidade Social (IVS), buscando identificar os fatores mais associados ao aumento de casos e óbitos por COVID-19.

Palavras-chaves: Base de Dados Aberta. COVID-19. Educação. Demografia Social. Vulnerabilidade Social. Mineração de dados.

Abstract

The data analysis process has been widely used in research across various fields of knowledge, including demographic studies and investigations into social inequality during the COVID-19 pandemic. Among the common approaches, those based on public data made available by government agencies stand out. In Brazil, relevant information on education, economy, health, and sociodemographic aspects is accessible to the general public. In this context, this study employs data science techniques to analyze Brazilian sociodemographic and educational data, aiming to measure inequality relations during the COVID-19 pandemic and provide technical support for public policy formulation. The research was guided by two main hypotheses: (1) inequality and social vulnerability play a crucial role in the worsening of epidemic outbreaks; and (2) lower-income social classes are the most affected by such crises due to limited access to quality basic services, such as healthcare and education. Based on these hypotheses, two case studies were conducted using public data. The results of the first study indicate that students from public schools in situations of socioeconomic vulnerability are more likely to miss the ENEM (National High School Exam) in scenarios where the exam is institutionally suspended, such as during states of public emergency. The second study found that low-income areas with poor urban infrastructure are more susceptible to increases in confirmed COVID-19 cases and related deaths. Additionally, the number of residents per household was shown to be associated with higher transmission rates of respiratory diseases, reinforcing the role of social vulnerability as a factor that amplifies the effects of the pandemic. The first case study analyzes the effects of school closures on student participation in the ENEM during the pandemic, from 2020 to 2022. The second assesses the impact of educational, urban, and economic infrastructure conditions in the Brazilian population using the Social Vulnerability Index (SVI), seeking to identify the factors most strongly associated with increases in COVID-19 cases and deaths.

Keywords: Open data repository. COVID 19. Education. Social Demography. Social Vulnerability. Data Mining

Lista de ilustrações

Lista de tabelas

Lista de abreviaturas e siglas

AHP:	*Analytic Hierarchy Process*
ACS:	*American Community Survey*
ADH:	*Altas de Desenvolvimento Humano*
AIDS:	Síndrome da Imunodeficiência Adquirida
BBN	*Bayesian Belief Network*
COVAX:	Iniciativa de acesso global a vacinas contra a COVID-19
COVID-19:	Doença por coronavírus 2019
CGI	*Common Gateway Interface*
CRISP-DM:	*Cross Industry Standard Process for Data Mining*
DATASUS:	Departamento de Informática do Sistema Único de Saúde
DP	Desvio Padrão
EaD	Ensino a Distância
EUA:	Estados Unidos da América
ECHO	*Environmental influences on Child Health Outcomes*
ENEM:	Exame Nacional do Ensino Médio
ESPIN:	Emergência de Saúde Pública de Importância Nacional
H1N1:	Subtipo do vírus Influenza A
IBGE:	Instituto Brasileiro de Geografia e Estatística
HI:	Seguro
IgM	imunoglobulina M
IgG	Imunoglobulina G
INEP:	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IPEA:	Instituto de Pesquisa Econômica Aplicada

IVS: Índice de Vulnerabilidade Social

K10 *Kessler Psychological Distress Scale*

KDD: *Knowledge Discovery in Databases*

MCDM-Fuzzy: Tomada de Decisão Multi-Critério Difusa

MEC: Ministério da Educação

MERS: Síndrome Respiratória do Oriente Médio

NCR: Relatório de Não Conformidade

OMS: Organização Mundial da Saúde

OPAS: Organização Pan-Americana da Saúde

PDA: Plano de Dados Abertos

PIB: Produto Interno Bruto

PIPS *Performance Indicators in Primary Schools*

PNAD: Pesquisa Nacional por Amostra de Domicílios

PNAD COVID: Pesquisa Nacional por Amostra de Domicílios COVID-19

PNADC: Pesquisa Nacional por Amostra de Domicílios Contínua

PGMPY: *Probabilistic Graphical Models using Python*

PNI: Programa Nacional de Imunizações

RB: Redes Bayesianas

RR Risco relativo

SARS: Síndrome Respiratória Aguda Grave

SBI: Sociedade Brasileira de Imunologia

SIDRA: Sistema IBGE de Recuperação Automática

SIM: Sistema de Informações sobre Mortalidade

SPSS: *Statistical Package for the Social Sciences*

SRAG: Síndrome Respiratória Aguda Grave

SUS: Sistema Único de Saúde

TPC:	Tabela de Probabilidades Condicionais
TL:	Taxa de Letalidade
TM:	Taxa de Mortalidade
UF:	Unidade da Federação
UTI:	Unidade de Terapia Intensiva
MCDM:	*Multi-Criteria Decision Making*
GIS:	Sistema de Informação Geográfica
TOPSIS:	Técnicas Difusas de Preferência Sequencial por Semelhança com Soluções Ideais

Lista de símbolos

$\prod$	Letra grega (PI Maiúsculo) Produtório
$\sum$	Letra grega sigma somatório
K	Quartis
Q	Intervalos dos Quartis
SBI	Sub-Índicice de Vulnerabilidade Social
e_0	Expectativa de vida nascidos vivos
e_{65}	Expectativa de vida até 63 anos de idade
$K_{Q_{\leq 25\%}}$	Rendimento inferior ou igual a 25%
$K_{Q_{26\%-74\%}}$	Rendimento entre 26% e 74%
$K_{Q_{\geq 75\%}}$	Rendimento superiores ou iguais a 75%
ξ	Percentual de abstenção ao ENEM
$-\infty$	Valores menores
$+\infty$	Valores maiores

Sumário

1 Introdução

Neste capítulo, são apresentados a contextualização, a problemática do estudo, as hipóteses propostas na tese, os objetivos geral e específicos, as principais contribuições e a relevância do trabalho, além da organização do documento.

1.1 Contextualização

Esta tese foi desenvolvida durante o período crítico da pandemia de COVID-19, o que implicou em desafios peculiares à atividade científica. As dificuldades enfrentadas incluíram o adoecimento de pesquisadores, o acesso limitado a estruturas institucionais e, sobretudo, o atraso na liberação de bases de dados fundamentais, como o Censo Demográfico, a Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) e outras fontes oficiais. Esses entraves impactaram diretamente o cronograma e o escopo inicial da pesquisa, exigindo adaptações metodológicas e temporais.

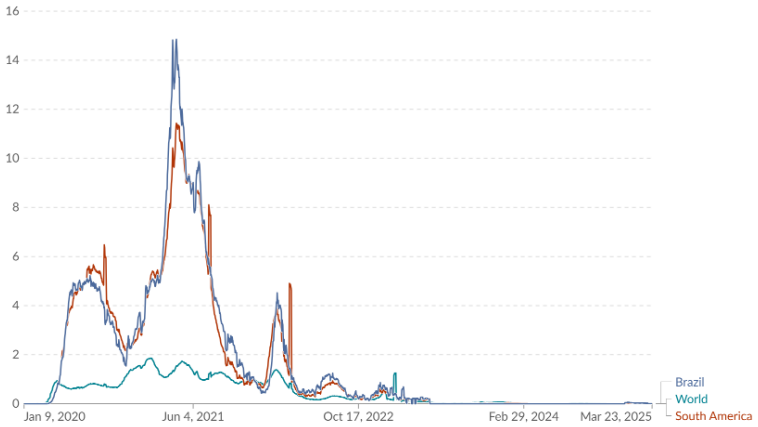
As doenças infecciosas são responsáveis por graves desequilíbrios sociais na população humana, decorrentes do surgimento de patógenos como vírus, bactérias, fungos e protozoários (SOLARTE; PEÑA; MADERA, 2006; ELLWANGER; CHIES, 2022). Esses agentes patogênicos provocam surtos epidemiológicos que, ao longo da história, resultaram em infecções massivas, mortes e distúrbios sociais generalizados (UJVARI, 2020; WALLACE, 2020). Dentre essas epidemias de grande impacto, destacam-se a tuberculose, a varíola, a febre amarela, a AIDS, a peste negra e a pandemia de influenza de 1918 (MORENS; FOLKERS; FAUCI, 2004; HAYS, 2006; UJVARI, 2020; CUNHA, 2024).

A história das epidemias é excepcionalmente rica em características, abrangendo desde a Peste Antonina até surtos de Síndrome Respiratória Aguda Grave (SARS), mais frequentes no século XX (FERNANDES, 2021). Entre esses eventos, sobressaem-se a pandemia de influenza de 1918 (DANA et al., 2020), o surto de coronavírus de 2012, conhecido como Síndrome Respiratória do Oriente Médio (MERS-CoV), e a recente pandemia causada pelo Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) (KIM et al., 2021; FOUST et al., 2020; BUENO; SOUTO; MATTA, 2021).

O SARS-CoV-2 foi inicialmente identificado em Wuhan, na China, no final de 2019 (CIOTTI et al., 2020; UDDIN et al., 2021). A Organização Mundial da Saúde (OMS) classificou a COVID-19 como pandemia de janeiro de 2020 a dezembro de 2023, período no qual foram registradas aproximadamente 657 milhões de infecções e mais de 6,68 milhões de mortes até 25 de dezembro de 2022 (World Health Organization, 2023), considerando também os casos de reinfecção (Figura 1).

Diante desse cenário, inúmeras pesquisas foram conduzidas para determinar os mecanismos de transmissão do vírus, o período de incubação, a letalidade em humanos e o impacto social da pandemia (BUENO; SOUTO; MATTA, 2021; SOTT; BENDER; BAUM, 2022). No Brasil, o primeiro caso confirmado de COVID-19 foi notificado em meados de fevereiro de 2020 (BRASIL, 2020e).

Figura 1 – Média móvel de 7 dias das novas mortes confirmadas por COVID-19 por milhão de pessoas.



Fonte: Extraído de Our World in Data, 2025.

No Brasil, a nova cepa da COVID-19 apresentou comportamentos distintos nas regiões. O Norte e o Nordeste registraram rápido colapso do sistema de saúde pública, em razão da menor capacidade hospitalar (BRASIL, 2020a; COELHO et al., 2020), enquanto o Sul e o Sudeste tiveram disseminação mais gradual (HALLAL et al., 2020). O Boletim Epidemiológico COVID-19 do Pará indicou elevadas taxas de letalidade em municípios com altos índices de vulnerabilidade social (IVS) (MATTA et al., 2021; Secretaria de Estado de Saúde Pública do Pará (SESPA), 2021).

Durante surtos epidêmicos, o processo de coleta de informações sociais incorpora novos parâmetros ou até mesmo a formulação de novas bases de dados auxiliares, capazes de representar de forma mais específica o problema em questão. Um exemplo disso é a PNAD COVID-19, que agregou, em seu escopo, informações sobre a saúde da população juntamente com dados sociodemográficos (PRAMIYANTI et al., 2020).

Dessa forma, bases de dados públicas, permanentes e auxiliares — como o Censo Demográfico, o DATASUS, a PNADC e a Pesquisa Nacional por Amostra de Domicílios COVID-19 (PNAD COVID-19), entre outras — abrangem setores sociais fundamentais para a construção de políticas públicas essenciais. Essas políticas são aplicáveis tanto à infraestrutura urbana, de forma geral, quanto às áreas da saúde e da educação, que estão diretamente relacionadas à qualidade de vida da população brasileira.

O controle e o gerenciamento das informações públicas devem ser realizados de maneira abrangente, permitindo a condução de pesquisas voltadas à identificação de

setores com alta vulnerabilidade social (BRASIL, 2018b). Essas investigações ampliam o conhecimento sobre as características sociais da população analisada, possibilitando a projeção de novos serviços ou, ainda, o fortalecimento de políticas públicas emergentes que contribuam para a redução das desigualdades sociais (BARATA, 1997; GRISOTTI, 2010).

1.2 Definição do problema e Motivação

As epidemias têm a capacidade de evidenciar, de forma mais acentuada, as desigualdades sociais entre países desenvolvidos e subdesenvolvidos, ou até mesmo dentro de um único país (MORAIS; FERES, 2024; COSTA, 2020). Essa diferença na disseminação de doenças não sazonais, como a COVID-19, está relacionada à infraestrutura tecnológica utilizada para rastrear, contabilizar e processar as informações sobre a transmissão do agente infeccioso.

Além disso, a pandemia de COVID-19 revelou o enfraquecimento da confiança pública na ciência, especialmente diante da ampla disseminação de desinformação. Muitos estudos realizados durante a pandemia apontaram de forma contundente os impactos sociais, econômicos e educacionais da crise sanitária. No entanto, grande parte dessas análises foi desconsiderada pelas esferas de tomada de decisão, o que contribuiu para o agravamento das desigualdades já existentes.

Mesmo com o fim do estado emergencial da pandemia, observa-se um arrefecimento do interesse científico em dar continuidade aos estudos sobre os seus impactos. Essa ausência de continuidade analítica pode dificultar a formulação de políticas públicas baseadas em evidências, necessárias para mitigar efeitos de médio e longo prazo, especialmente em populações historicamente vulneráveis.

Consequentemente, os estudos realizados em áreas com maior desigualdade social tendem a se tornar mais custosos, devido à escassez não apenas de infraestrutura tecnológica, mas também de infraestrutura urbana, bem como de recursos humanos e materiais adequados (CERQUEIRA et al., 2025; BARBOSA, 2020).

Essas diferenças sociais entre os países tornam-se ainda mais visíveis quando se considera a prestação de serviços básicos de saúde oferecidos à população, como consultas médicas em ambientes públicos ou privados e a distribuição de medicamentos terapêuticos essenciais para manter um nível mínimo de qualidade de vida (MORAIS; FERES, 2024; DANA et al., 2020).

Esse cenário evidencia a importância do processo de transparência e divulgação de dados em todos os países, especialmente naqueles com sistemas de informação desatualizados. A ausência de informações claras pode gerar resultados analíticos imprecisos sobre doenças epidêmicas, frequentemente resultando em decisões insustentáveis e perigosas

para as populações mais vulneráveis (BRAGA, 2023; RAAMKUMAR; TAN; WEE, 2020). A falta de dados confiáveis também pode impactar as redes laboratoriais de diagnóstico, devido à escassez de insumos básicos e à presença de informações com viés de confiabilidade, o que intensifica o processo de transmissão (BRAGA, 2023; CARDOSO et al., 2020).

A doença é um fenômeno biológico e social, fruto de debates e incertezas ao longo do tempo. Para lidar com os desafios impostos por surtos epidêmicos, como a pandemia de COVID-19 (LIMA, 2023; FIOCRUZ, 2020b), é fundamental compreender os fatores envolvidos e buscar consensos sobre as melhores abordagens. Isso envolve a análise de indicadores sociais que contribuem para a rápida disseminação da doença.

A justificativa deste trabalho destaca o uso de microdados públicos nas áreas de saúde, demografia populacional e educação. Para tanto, os dados foram tratados de maneira adequada, a fim de compreender as elevadas taxas de infecção por COVID-19 no Brasil, além de analisar a disseminação da doença no território nacional.

A motivação deste estudo é compreender os impactos da pandemia sobre a população, especialmente aquela em situação de maior vulnerabilidade, que depende de serviços públicos. O objetivo é analisar a relação entre os casos confirmados e os óbitos registrados, com o propósito de subsidiar a tomada de decisões por parte dos gestores públicos. Para essa análise, foram utilizados dados de pesquisas populacionais disponíveis em bases de dados públicas brasileiras, como o Censo Demográfico (2000, 2010 e 2022), a PNADC, o Censo Escolar e informações sobre o impacto da pandemia no Exame Nacional do Ensino Médio (ENEM).

As bases de dados mencionadas nesta seção são regulamentadas pela Política de Dados Abertos do Poder Executivo Federal, instituída pelo Decreto nº 8.777, de 11 de maio de 2016, que garante a transparência das informações analisadas.

1.3 Contribuições Técnicas, Científicas e Sociais

Esta tese, intitulada *Ciência de Dados para Análise da Desigualdade Social durante a Pandemia: Uma Análise com Dados Públicos Brasileiros*, apresenta avanços relevantes sob as perspectivas técnica, científica e social, conforme descrito a seguir.

1.3.1 Contribuições Técnicas

- Desenvolvimento de uma abordagem integrada de ciência de dados, aplicando técnicas de mineração de dados, análise estatística e modelagem probabilística para o estudo de fenômenos sociais complexos.
- Implementação de Redes Bayesianas como ferramenta de análise preditiva e inferencial

para correlacionar variáveis sociodemográficas e epidemiológicas em grandes bases públicas, ampliando o uso dessas redes em estudos de saúde pública e educação.

- Proposição de um fluxo metodológico baseado no processo CRISP-DM, adaptado para lidar com bases públicas heterogêneas e despadronizadas, como o Censo Demográfico, a PNADC, o ENEM e dados epidemiológicos da COVID-19.
- Sistematização de procedimentos de tratamento de dados abertos, incluindo estratégias de filtragem, normalização e integração de múltiplas fontes, garantindo replicabilidade e transparência metodológica.

1.3.2 Contribuições Científicas

- Ampliação do entendimento sobre a relação entre vulnerabilidades sociais, desempenho educacional e impactos da pandemia de COVID-19 no contexto brasileiro, utilizando dados empíricos de abrangência nacional.
- Produção de novos modelos de análise que combinam indicadores de infraestrutura urbana, capital humano e renda/trabalho, permitindo inferências mais refinadas sobre as desigualdades regionais no Brasil durante o período pandêmico.
- Avanço na literatura de ciência de dados aplicada às áreas de ciências sociais e saúde pública, ao demonstrar a aplicabilidade de métodos probabilísticos (Redes Bayesianas) em contextos de dados públicos heterogêneos.
- Contribuição para o debate acadêmico sobre as limitações e potencialidades do uso de dados públicos na formulação de políticas públicas baseadas em evidências.

1.3.3 Contribuições Sociais

- Geração de subsídios analíticos para gestores públicos e formuladores de políticas sociais, possibilitando o direcionamento mais eficiente de ações emergenciais em saúde e educação durante e após a pandemia.
- Identificação de padrões de desigualdade que podem orientar investimentos públicos em infraestrutura, programas educacionais e ações de mitigação de vulnerabilidades sociais.
- Estímulo à cultura de transparência e ao uso estratégico de dados públicos no Brasil, reforçando a importância do acesso aberto à informação como direito fundamental e como instrumento de redução das desigualdades sociais.

- Sensibilização acadêmica e social para a necessidade de integrar ciência de dados e ciências humanas na análise de crises sanitárias, visando à construção de sociedades mais resilientes e equitativas.

1.4 Hipóteses

- A desigualdade e a vulnerabilidade social desempenham um papel crucial no agravamento dos surtos epidêmicos. Os dados públicos sobre esses fatores podem ser utilizados como base para a formulação de políticas públicas específicas, adaptadas às necessidades de diferentes realidades sociais.
- A pandemia evidenciou que as classes sociais mais baixas são as mais afetadas por surtos epidêmicos, devido à falta de acesso a serviços básicos de qualidade, como saúde e educação. A carência desses serviços agrava as condições de vida, elevando as taxas de infecção e mortalidade.
- O uso de Redes Bayesianas pode contribuir para a detecção de padrões ocultos nas relações entre fatores sociodemográficos e desfechos da pandemia (SCUTARI; DENIS, 2018).

1.5 Objetivo Geral

O objetivo desta tese é obter, processar e analisar dados públicos provenientes de diversas fontes e temáticas — como o Atlas IVS, o BRASIL.IO, o Censo Demográfico, o DATASUS, o ENEM, a PNADC e o SIDRA — relacionados à transmissão da COVID-19 no Brasil entre 2019 e 2022. Utilizando técnicas de ciência de dados, busca-se avaliar o impacto da pandemia na qualidade de vida da população brasileira. Adicionalmente, propõe-se uma análise educacional com o uso de Redes Bayesianas, a fim de investigar a participação e o desempenho dos estudantes no ENEM, considerando o IVS como parâmetro sociodemográfico de referência.

1.5.1 Objetivos Específicos

A partir do objetivo geral, foram definidos os seguintes objetivos específicos que norteiam as fases metodológicas deste estudo:

- Sistematizar as informações bibliográficas de trabalhos relacionados à temática desta tese;
- Compreender como o processo de transmissão durante surtos epidêmicos respiratórios atuou nas diferentes regiões sociodemográficas;

- Analisar o IVS a partir de dados atualizados, como os provenientes da PNADC e do Censo Demográfico de 2022;
- Identificar as dimensões sociais que contribuíram para o aumento do número de casos e mortes por doenças respiratórias;
- Avaliar os impactos da pandemia de COVID-19 sobre a vida da população brasileira no período pós-pandêmico;
- Gerar relatórios científicos e técnicos que possam subsidiar entidades governamentais na preparação e resposta a novos surtos epidêmicos.

1.6 Organização da Tese

Esta tese está estruturada em sete capítulos, além desta introdução, conforme descrito a seguir:

Capítulo 2: Apresenta o referencial teórico e discute o contexto da pandemia de COVID-19 sob duas vertentes principais: a análise educacional, abordando os impactos da pandemia no desempenho dos estudantes, e a análise sociodemográfica, considerando as transformações nas condições sociais e econômicas da população.

Capítulo 3: Introduz conceitos fundamentais sobre doenças infecciosas respiratórias, com ênfase na COVID-19, e descreve o processo de notificação de casos confirmados e óbitos durante surtos epidêmicos.

Capítulo 4: Aborda os fundamentos da ciência de dados, o processo CRISP-DM e os conceitos técnicos aplicados neste trabalho, incluindo o uso de Redes Bayesianas e uma breve descrição do IVS como metodologias de análises.

Capítulo 5: Descreve a metodologia adotada para a coleta, tratamento e análise dos dados, além do ambiente computacional utilizado no desenvolvimento da pesquisa.

Capítulo 6: Apresenta dois estudos de caso: o primeiro, focado na análise sociodemográfica dos impactos da pandemia, e o segundo, na análise educacional baseada nos dados de desempenho dos estudantes participantes do ENEM no estado do Pará.

Capítulo 7: Expõe as conclusões dos estudos realizados, discute as principais dificuldades enfrentadas durante a pesquisa e propõe sugestões para investigações futuras, além de apresentar as publicações decorrentes desta tese.

2 Trabalhos Correlatos

Neste capítulo, serão apresentados os trabalhos recentes e mais relevantes relacionados aos seguintes temas: (i) análise de dados sociodemográficos e (ii) impacto da COVID-19 no contexto escolar. O estudo foi conduzido por meio de uma pesquisa em artigos científicos publicados em periódicos de referência na área, tais como *Health Education & Behavior*, *Nature Medicine*, *Revista Brasileira de Epidemiologia*, *International Journal of Disaster Risk Reduction*, *International Journal of Educational Research Open*, *Infectious Diseases of Poverty*, entre outros.

2.1 Análises Sociodemográficas acerca da COVID-19

Malakar (2022) buscou modelar a vulnerabilidade à COVID-19 utilizando uma abordagem de Tomada de Decisão Multicritério Difusa (MCDM-Fuzzy), por meio do método AHP e da técnica de preferência por similaridade com a solução ideal (TOPSIS), integradas ao Sistema de Informação Geográfica (GIS) aplicado à região de Bengala Ocidental, na Índia. Foram utilizados 15 parâmetros, distribuídos em três critérios: vulnerabilidade social, vulnerabilidade epidemiológica e vulnerabilidade física.

O estudo revelou que aproximadamente 20% da área total de Bengala Ocidental apresenta alta vulnerabilidade à COVID-19; 23,42% é moderadamente vulnerável; e 57,03% está sob baixa vulnerabilidade. As regiões mais vulneráveis incluem Kolkata, South 24 Paraganas e North 24 Paraganas, considerados distritos densamente povoados. Assim, recomenda-se que as agências governamentais priorizem ações de planejamento e proteção nas áreas de maior risco (MALAKAR, 2022).

O estudo de Fortuna Fortuna, Setiawan e Sharifi (2023) analisou como fatores espaciais e sociodemográficos influenciaram os padrões de transmissão da COVID-19 em Surabaya, Indonésia. Os autores destacam a importância da compreensão das dinâmicas locais para o desenvolvimento de estratégias de mitigação eficazes.

Utilizando uma abordagem quantitativa, foram analisados dados de casos confirmados de COVID-19 em diferentes distritos da cidade. As variáveis incluíram densidade populacional, distribuição etária, renda média e acesso a serviços de saúde. Modelos de regressão foram empregados para identificar correlações significativas.

Os resultados mostraram que áreas com maior densidade populacional e menor renda média apresentaram taxas mais altas de transmissão. Distritos com acesso limitado aos serviços de saúde também demonstraram maior vulnerabilidade. Tais achados sugerem que políticas públicas devem considerar as desigualdades espaciais e sociodemográficas para

conter a disseminação do vírus de maneira eficaz (FORTUNA; SETIAWAN; SHARIFI, 2023).

Torre-Luque et al. (2024) investigou as tendências de mortalidade por suicídio na Espanha durante a pandemia, com foco em fatores sociodemográficos. Considerando os impactos globais da pandemia sobre os sistemas de saúde pública e o aumento relatado de suicídios em diversos países, o objetivo foi identificar os grupos mais afetados.

Com base em dados do Índice Nacional de Mortalidade (2000–2021), os autores compararam taxas de suicídio antes e durante a pandemia. Foram utilizados modelos de séries temporais de *Poisson*, com variáveis como sexo, idade, estado civil, status migratório e urbanicidade.

Os resultados indicaram aumento significativo nas taxas de suicídio nos anos pandêmicos ($p < 0{,}01$), com risco relativo (RR) de 1,05 (IC95% = [1,02; 1,08]). Esse aumento foi observado para ambos os sexos, grupos migratórios, adultos de meia-idade, residentes em grandes centros urbanos e pessoas solteiras ($RR > 1,05$), destacando o papel de fatores sociais como ausência de redes de apoio (TORRE-LUQUE et al., 2024).

Yazeedi et al. (2024) analisou os fatores sociodemográficos e de saúde associados a mudanças nos comportamentos de estilo de vida entre adultos omanitas durante a pandemia. Considerando os impactos das medidas de isolamento e *lockdowns*, o estudo avaliou alterações nos hábitos alimentares, atividade física, padrões de sono e saúde mental.

A pesquisa, de delineamento transversal, foi conduzida via questionário online entre janeiro e março de 2022, com 515 participantes (média de idade: 25,7 anos; DP ±6,83). As variáveis incluíram idade, gênero, escolaridade, status acadêmico, estado civil, histórico de infecção por COVID-19 e mudanças de peso.

Os resultados revelaram mudanças negativas nos comportamentos de estilo de vida (média = −2,75, DP ±9,08). Indivíduos mais jovens ($p < 0{,}001$), solteiros ($p = 0{,}012$), universitários ($p = 0{,}004$) e aqueles com ganho ou incerteza sobre o peso ($p < 0{,}001$) apresentaram piores pontuações. Cada aumento de um ano na idade correspondia a incremento de 0,19 na pontuação ($B = 0{,}19$, $p = 0{,}02$). Já o ganho de peso representou queda superior a cinco pontos ($B = -5{,}56$, $p < 0{,}001$). O estudo destaca a importância de intervenções voltadas à promoção da saúde mental e de hábitos saudáveis entre jovens adultos (YAZEEDI et al., 2024).

O estudo de Blackwell et al. (2025) identificou padrões sociodemográficos relacionados ao status vacinal infantil contra a COVID-19 e às razões da hesitação vacinal parental. A compreensão desses padrões é essencial para políticas públicas eficazes de imunização.

Com dados de 5.103 pais participantes do estudo ECHO (*Environmental Influences on Child Health Outcomes*), foram realizados testes qui-quadrado e regressões de *Poisson* ajustadas para investigar associações entre características sociodemográficas e motivos de

hesitação.

Verificou-se que 41,7% das crianças haviam recebido ao menos uma dose da vacina. Apenas 1% dos pais hesitantes citaram dificuldades de acesso como motivo. As principais preocupações incluíram efeitos colaterais (gerais: 53,1%; longo prazo: 56,2%), percepção de baixo risco (33,7%) e infecção prévia (21%). Pais de crianças negras e indígenas demonstraram menor preocupação com efeitos adversos, enquanto os de maior renda e escolaridade atribuíram menor risco à infecção. Tais resultados reforçam a necessidade de estratégias específicas para distintos grupos sociais (BLACKWELL et al., 2025).

Dharmayani e Mihrshahi (2025) investigou a prevalência de sofrimento psicológico entre adultos australianos (18–64 anos) durante a pandemia, e os fatores sociodemográficos associados. Com base em dados da Pesquisa Nacional de Saúde Australiana (2020–2021), o sofrimento foi medido pela escala K10 (*Kessler Psychological Distress Scale*). Foram utilizados modelos de regressão logística multivariada ajustados ao desenho amostral.

A média da escala foi de 16,94, com 21,13% apresentando sofrimento alto/muito alto. Mulheres jovens (18–25 anos) registraram as maiores médias (M = 20,44). Fatores de risco incluíram baixa renda, estar solteiro, viúvo ou separado e ser nativo australiano. Por outro lado, possuir imóvel e ter filhos mostraram-se protetores. Entre as mulheres, ter mais de 56 anos reduziu significativamente o risco. As conclusões sugerem a necessidade de ações voltadas à saúde mental, especialmente para mulheres jovens (DHARMAYANI; MIHRSHAHI, 2025).

Diante desse panorama, o primeiro estudo de caso desta tese visa compreender, de maneira geral, quais setores foram mais impactados pela pandemia de COVID-19 no Brasil. Para isso, utiliza-se a métrica do IVS, com dados da PNADC.

Esse índice, calculado pelo Atlas IBGE, reflete a qualidade de vida da população nas unidades federativas. Este estudo se diferencia por correlacionar dados de 2015 a 2019 nas capitais brasileiras, identificando dimensões sociais mais suscetíveis e que podem atuar como vetores de surtos epidêmicos de doenças respiratórias.

2.2 Impacto da COVID-19 no Nicho Escolar

A pesquisa conduzida por Bartholo et al. (2023) utilizou dados longitudinais de 671 crianças com idades entre 5 e 6 anos, matriculadas em 21 escolas privadas do Rio de Janeiro — tanto com fins lucrativos quanto sem fins lucrativos. A análise comparativa foi realizada entre duas coortes: uma de 2019, anterior à pandemia, e outra de 2020, período em que as escolas estavam fechadas. As crianças foram avaliadas no início e no fim do ano letivo, por meio de uma adaptação do instrumento *Performance Indicators in Primary Schools* (PIPS). Para mensurar as variações no desempenho em linguagem e matemática, os autores

aplicaram modelos estatísticos de valor agregado e análises hierárquicas, permitindo isolar o efeito da pandemia sobre a aprendizagem.

Os resultados indicam perdas significativas no desempenho das crianças avaliadas em 2020, em comparação às de 2019: cerca de 0,23 desvio padrão em linguagem e 0,25 em matemática, correspondendo a um atraso de aproximadamente três a quatro meses de aprendizagem. O estudo também evidenciou o agravamento das desigualdades educacionais: crianças de famílias com menor nível socioeconômico aprenderam apenas 48% do esperado em um ano letivo regular, enquanto aquelas de famílias com maior renda alcançaram cerca de 75%. Os autores concluem que a pandemia comprometeu o aprendizado e aprofundou desigualdades preexistentes no sistema educacional brasileiro (BARTHOLO et al., 2023).

O estudo de Zhuo e Harrigan (2023) investigou a relação entre nível educacional e mortalidade por COVID-19 nos Estados Unidos. Embora várias pesquisas apontem correlações entre desigualdades socioeconômicas e risco de óbito, os autores propuseram que boa parte dessa associação pode ser explicada pelo nível de escolaridade. O objetivo foi avaliar se a baixa escolaridade está diretamente associada ao aumento da mortalidade por COVID-19, independentemente de outros fatores socioeconômicos e comportamentais.

Com dados de 3.108 condados dos EUA, foram aplicados modelos de regressão binomial negativa multinível para analisar mortes ocorridas entre 20 de janeiro de 2020 e 10 de maio de 2022. As análises controlaram variáveis como economia, raça, geografia, vacinação, orientação política, condições de saúde e comportamentos preventivos.

Os resultados revelaram forte correlação entre baixa escolaridade e maior mortalidade por COVID-19, com razão de taxa de incidência de 1,17 (IC95% = [1,15; 1,20]). Esse risco é comparável ao de pessoas com mais de 65 anos. Os autores sugerem que a educação influencia tanto a compreensão individual de informações científicas quanto normas coletivas de saúde, como o uso de máscaras. Assim, a baixa escolaridade deve ser considerada fator de risco essencial nas estratégias de combate à pandemia (ZHUO; HARRIGAN, 2023).

A pesquisa de Csorba e Dabija (2024) teve como objetivo analisar como a pandemia influenciou o comportamento futuro de estudantes em relação à educação online. O estudo destaca a importância de compreender mudanças nas percepções e atitudes diante do ensino remoto em contextos de crise sanitária.

Foi utilizada uma abordagem quantitativa com aplicação de questionários estruturados a uma amostra representativa de estudantes universitários. As variáveis analisadas incluíram experiência prévia com EaD, percepção da eficácia do ensino remoto durante a pandemia e intenção de continuar utilizando plataformas digitais no futuro. Técnicas de regressão foram utilizadas para identificar os fatores mais influentes.

Os resultados apontaram que experiências positivas com o ensino remoto aumenta-

ram a propensão dos estudantes a adotarem a educação online no futuro. Flexibilidade, autonomia e familiaridade com tecnologia foram fatores determinantes para a aceitação. Recomenda-se que as instituições de ensino superior invistam em infraestrutura tecnológica e capacitação docente para aprimorar a qualidade do ensino online (CSORBA; DABIJA, 2024).

O estudo de Kong et al. (2024) analisou os efeitos da transição para o ensino remoto sobre o consumo energético de instituições educacionais e o desempenho acadêmico dos estudantes. Com a adoção generalizada do ensino a distância, tornou-se necessário compreender os impactos sobre a eficiência energética e a aprendizagem.

Foram coletados dados de consumo de energia de diversas instituições antes e durante o ensino remoto, além de indicadores acadêmicos como notas e taxas de aprovação. Técnicas de regressão foram aplicadas para identificar correlações entre o modelo de ensino, o consumo energético e os resultados acadêmicos.

Os achados mostraram redução significativa no consumo de energia dos edifícios educacionais, devido à menor ocupação física. Contudo, observou-se também uma queda no desempenho dos estudantes, sugerindo limitações na eficácia do ensino remoto. Destaca-se a necessidade de equilibrar estratégias de sustentabilidade com métodos pedagógicos eficazes (KONG et al., 2024).

Emara et al. (2025a) investigou os efeitos não lineares e interativos do ensino a distância (EaD) nas taxas de matrícula universitária após a pandemia. A rápida transição para o EaD evidenciou a necessidade de compreender como essa mudança afetou diferentes grupos sociais.

Foram analisados dados de matrículas em várias instituições antes e depois da adoção do EaD. Modelos econométricos consideraram variáveis como renda familiar, localização e acesso à tecnologia, buscando entender relações complexas entre EaD e matrícula.

Os resultados indicaram variações significativas entre grupos. Enquanto alguns se beneficiaram da flexibilidade do EaD, outros enfrentaram barreiras tecnológicas. Políticas educacionais devem ser adaptadas às diferentes realidades para garantir equidade no acesso à educação superior (EMARA et al., 2025a).

Emara et al. (2025b) também analisou impactos do EaD nas matrículas universitárias no período pós-pandemia, utilizando dados de cerca de 5.000 instituições entre 2019 e 2022. Foram aplicados modelos de regressão com efeitos fixos, considerando variáveis como adoção do ensino remoto, contexto pandêmico e características institucionais.

Os resultados mostraram que um aumento de 1% na taxa de EaD corresponde a cerca de 7.964 estudantes a mais na matrícula total. Esse impacto foi ainda maior no pós-pandemia (9.361 estudantes), mas apresentou retornos decrescentes em níveis

elevados de EaD: 91% sem COVID-19 e 67% com COVID-19. Isso evidencia a necessidade de políticas que equilibrem a expansão do EaD com a qualidade e equidade no ensino (EMARA et al., 2025b).

Dessa forma, o segundo estudo de caso desta tese está voltado para a análise de dados educacionais com base no ENEM. Utilizando ciência de dados e Redes Bayesianas, busca-se compreender os impactos da pandemia sobre o setor educacional. Posteriormente, pretende-se aplicar técnicas de aprendizado de máquina nos parâmetros mais propensos a afetar surtos epidêmicos de SARS no Brasil, os quais podem intensificar a transmissão da doença e impactar negativamente o desempenho escolar.

2.3 Considerações Finais

No contexto das investigações sobre os impactos provocados pela pandemia de COVID-19, a análise de dados sociodemográficos mostra-se essencial para a identificação de áreas com maior propensão a surtos epidêmicos. Além disso, o conhecimento gerado por esses estudos pode subsidiar ações governamentais, contribuindo para a formulação de estratégias de contenção e para a elaboração de protocolos de resposta a futuras crises sanitárias.

Apesar dos avanços já registrados na literatura científica, ainda persistem lacunas importantes. Destaca-se, sobretudo, a necessidade de metodologias mais precisas e atualizadas que permitam correlacionar, em tempo real, variáveis sociodemográficas e epidemiológicas, minimizando a defasagem entre a coleta e a análise dos dados.

Adicionalmente, observa-se a escassez de estudos que avaliem de forma sistemática a efetividade das políticas públicas adotadas durante a pandemia, bem como seu impacto na mitigação das desigualdades sociais. Outra limitação significativa refere-se ao acesso restrito a bases de dados abrangentes, detalhadas e atualizadas, o que compromete a robustez das inferências estatísticas e limita a generalização dos resultados obtidos. A seguir, será apresentado o capítulo referente ao referencial teórico, abordando os temas COVID-19, SARS e o IVS.

3 Referencial Teórico

Este capítulo oferece uma breve descrição das bases de dados utilizadas nos estudos de caso apresentados neste documento sobre a propagação da COVID-19. Além disso, introduz-se sucintamente o processo de disseminação da COVID-19 no Brasil.

3.1 *Dados Abertos*

O conceito de dados abertos surgiu em 1995, com o objetivo de garantir o acesso livre a dados ambientais e geofísicos, promovendo a colaboração entre pesquisadores (CHIGNARD, 2013). Esse movimento se consolidou ao longo do tempo, especialmente após o memorando de transparência de Barack Obama, nos Estados Unidos (OBAMA, 2009). Em 2011, oito países, incluindo o Brasil, formalizaram a *Open Government Partnership* (MEDEIROS et al., 2017).

A Lei nº 12.527/2011, sancionada em 18 de novembro de 2011, assegura o acesso à informação pública no Brasil, estabelecendo que os dados governamentais sejam disponibilizados em formato aberto. Tais dados são essenciais para promover a transparência, e podem ser livremente utilizados e reutilizados por qualquer pessoa ou instituição. Eles devem obedecer a três princípios fundamentais: disponibilidade, reuso e participação universal (OPEN DATA COMMONS, 2019).

O Plano de Dados Abertos (PDA) da Enap, aprovado em 2022, orienta a implementação de ações voltadas aos dados abertos, conforme o Decreto nº 8.777/2016 (ENAP, 2022). Além disso, a Lei Complementar nº 101/2000 exige que o poder público utilize meios eletrônicos de gestão financeira e ferramentas de transparência, garantindo o acesso a informações orçamentárias e promovendo maior eficiência e participação social na gestão pública (ENAP, 2022).

No Brasil, existem bases de dados públicas que refletem com a diversidade da população. A PNADC, por exemplo, abrange diferentes níveis de divisão territorial, permitindo análises tanto em escala macro quanto microrregional. Essa base é fundamental para medir diversos aspectos sociais, como características sociodemográficas, condições socioeconômicas, renda da população e infraestrutura habitacional (IBGE, 2023a).

Entre as fontes de dados públicas utilizadas neste estudo, destacam-se:

SIDRA: Plataforma do IBGE que disponibiliza séries históricas e dados estatísticos sobre indicadores econômicos e sociodemográficos, como emprego, renda e PIB (IBGE, 2023b).

DATASUS: Sistema de informação do Ministério da Saúde responsável pela coleta, processamento e divulgação de dados relacionados à saúde pública no Brasil (BRASIL, 2023).

Brasil.IO: Repositório de dados públicos que reúne informações epidemiológicas sobre casos e óbitos por COVID-19, além de dados do Censo Escolar e outros registros administrativos de interesse público[1] (JUSTEN; et all, 2023).

Censo Escolar: Levantamento anual coordenado pelo INEP que reúne informações detalhadas sobre a educação básica no Brasil, incluindo matrícula, infraestrutura, docentes e fluxo escolar.

ENEM: Avaliação de abrangência nacional aplicada aos concluintes do ensino médio, cujos microdados oferecem informações sobre desempenho, perfil dos participantes e variáveis socioeconômicas (INEP, 2023).

3.2 Doenças Infecciosas

De acordo com Youngerman (2008), as doenças infecciosas são caracterizadas por um número elevado de indivíduos contaminados, superior ao esperado. Elas são classificadas como endêmicas quando sua ocorrência é constante dentro de um determinado padrão. Quando há um aumento significativo nos casos confirmados em uma região específica, a infecção passa a ser considerada um surto epidêmico, uma vez que pode facilmente ultrapassar o limiar epidêmico estabelecido (BUTANTAN, 2022).

Ainda segundo Youngerman (2008), a diferença entre epidemia e pandemia está na capacidade de proliferação da doença. Enquanto a epidemia se restringe a uma área geográfica limitada, a pandemia é uma epidemia em maior escala, atingindo múltiplos países ou até continentes.

Doenças endêmicas podem evoluir para surtos, epidemias e até pandemias, frequentemente associadas a fatores sociais, econômicos e, particularmente, à urbanização. Esses fatores podem alterar as condições iniciais de transmissão da doença, além das mutações genéticas sofridas pelo agente infeccioso (SILVA, 2003).

As investigações de surtos epidêmicos devem seguir uma sequência de etapas para serem efetivas (BRASIL, 2018a), a saber: estabelecer a presença do surto; confirmar o diagnóstico; contabilizar o total de infectados; descrever a localização, os indivíduos afetados e o período em que o surto ocorreu; avaliar o risco de novas contaminações; iniciar um estudo analítico investigativo por meio da formulação de hipóteses e análise dos dados disponíveis sobre a doença; implementar medidas preventivas eficazes para desacelerar o contágio; e, por fim, divulgar os resultados da análise investigativa à comunidade.

[1] https://brasil.io/home/

No Brasil, o processo de investigação de surtos é conduzido pelas secretarias municipais, estaduais e federais de saúde, seguindo uma hierarquia de risco voltada à população mais suscetível ao adoecimento. Em outras palavras, quanto maior a incidência da doença na população, maior será o apoio das Secretarias de Vigilância Epidemiológica para conter a sua transmissão (BRASIL, 2018a).

Historicamente as doenças infecciosas desempenharam um papel central na geração de desequilíbrios sociais profundos, como evidenciado pela pandemia de influenza de 1918, também conhecida como gripe espanhola. Causada pelo vírus *influenza* A (H1N1), essa pandemia provocou a morte de aproximadamente 50 milhões de pessoas no mundo, com impactos devastadores sobre a saúde pública, a economia e a organização social (TAUBENBERGER; MORENS, 2006; MORENS; FOLKERS; FAUCI, 2004).

Caracterizada por uma disseminação extremamente rápida e uma alta taxa de mortalidade entre adultos jovens, a influenza de 1918 expôs as fragilidades dos sistemas de saúde da época e impulsionou mudanças duradouras nas políticas sanitárias globais.

Posteriormente, outros surtos importantes também evidenciaram a vulnerabilidade das sociedades a novas ameaças infecciosas. O surgimento do coronavírus MERS-CoV em 2012, associado à Síndrome Respiratória do Oriente Médio, e a recente pandemia de COVID-19, iniciada em 2019 com a disseminação do SARS-CoV-2, reforçam a necessidade contínua de vigilância epidemiológica e fortalecimento dos sistemas de saúde pública (KIM et al., 2021; UDDIN et al., 2021).

3.3 Doenças causada pelo Coronavírus

De acordo com a Organização Pan-Americana da Saúde (OPAS)[2], os coronavírus constituem uma família de vírus capazes de causar sintomas que variam desde resfriados comuns até doenças respiratórias mais graves, afetando o funcionamento normal do sistema respiratório. Segundo (SOUZA et al., 2021; SU et al., 2016), existem sete variantes do coronavírus que podem infectar seres humanos e provocar doenças respiratórias de diferentes gravidades, incluindo pneumonias.

Souza et al. (2021) destaca três das sete variantes conhecidas do coronavírus por estarem associadas ao desencadeamento de surtos epidêmicos e doenças graves, como a SARS. O surto de SARS, ocorrido na China em 2002, inicialmente foi confundido com a infecção bacteriana causada por Chlamydia pneumoniae, devido à semelhança dos sintomas, o que atrasou sua identificação, conforme relatado pela Organização Mundial da Saúde (OMS).

Esse surto, no entanto, não foi um evento isolado. Em 2012, uma nova variante

[2] https://www.paho.org/pt/doenca-causada-pelo-novo-coronavirus-covid-19

do coronavírus surgiu, provocando infecções respiratórias graves e comprometimento da função renal em diversos pacientes. Essa nova variante ficou conhecida como MERS-CoV, causadora da Síndrome Respiratória do Oriente Médio (LIU et al., 2020). Mais recentemente, uma outra variante da mesma família, denominada SARS-CoV-2, deu origem à pandemia de COVID-19.

A SARS representa um grave problema de saúde pública, caracterizado por infecções respiratórias de rápida progressão, provocadas por diferentes variantes do coronavírus, classificadas como SARS-CoV-1, MERS-CoV e SARS-CoV-2 (BARMAN et al., 2022). Assim, o SARS-CoV é considerado uma ameaça global, com variantes que causam infecções respiratórias graves e apresentam diferentes características clínicas e epidemiológicas.

3.4 COVID-19 no Brasil

Em 3 de fevereiro de 2020, o Brasil emitiu uma nota de advertência, declarando o surto do novo coronavírus como Emergência de Saúde Pública de Importância Nacional (ESPIN), mesmo antes da confirmação de casos em seu território (BRASIL, 2020f). Algumas semanas depois, em 26 de fevereiro de 2020, o Ministério da Saúde confirmou o primeiro caso de COVID-19 na cidade de São Paulo. Em 12 de março de 2020, foi registrada a primeira morte associada ao vírus no país (BRITO, 2020).

Após a confirmação do SARS-CoV-2 no Brasil, o Congresso Nacional aprovou um decreto que estabeleceu o estado de calamidade pública, o que proporcionou maior flexibilidade orçamentária às autoridades locais para direcionar recursos ao enfrentamento da COVID-19 (BRASIL, 2020b).

Nesse contexto,OMS como forma de retardar a propagação do vírus, com ênfase na importância do distanciamento social entre as principais medidas recomendadas (BUENO; SOUTO; MATTA, 2021; SOUSA, 2020).

Em junho de 2021, o Brasil atingiu a marca de 500 mil óbitos por COVID-19, com uma taxa de mortalidade aproximada de dois óbitos por milhão de habitantes — 4,7 vezes superior à média global. Esses números colocaram o país entre os principais epicentros da pandemia (PILAR; CASTRO, 2020). Em julho de 2022, o Ministério da Saúde, por meio do painel de monitoramento de casos, reportou cerca de 670 mil mortes e uma taxa de infecção de 319 mil por milhão de habitantes (BRASIL, 2020a).

Delatorre et al. (2020) publicou um estudo pelo Instituto Oswaldo Cruz que revelou discrepâncias entre os dados notificados oficialmente e os reais. A pesquisa apontou que a taxa de infecção era superior à relatada pelos painéis governamentais. O aumento da mobilidade populacional durante o carnaval, ocorrido nas primeiras semanas de fevereiro de 2020, precedeu a primeira confirmação oficial da COVID-19 no país (BRITO, 2020).

Portanto, é possível concluir que, mesmo antes da notificação oficial, o vírus já estava em transmissão ativa, não detectada pelo sistema de saúde vigente. Essa constatação está em consonância com estudos de Cereda et al. (2020), Davis et al. (2020), que, a partir de análises laboratoriais genéticas, evidenciaram a circulação precoce e não monitorada do vírus pelas autoridades sanitárias.

O avanço da COVID-19 no Brasil foi agravado por fatores sociodemográficos, como a desigualdade social, a densidade populacional elevada e a ampla parcela da população vivendo em situação de vulnerabilidade socioeconômica. Estima-se que o país tenha aproximadamente 213,3 milhões de habitantes, dos quais mais de 30

Famílias de baixa renda enfrentam deficiências no acesso a serviços básicos, como saúde, educação e distribuição de renda. Além disso, trabalham, muitas vezes, em condições insalubres (DANA et al., 2020). Nesse cenário,SUS, que atende cerca de 70

Grupos prioritários, como pessoas com comorbidades (imunossupressão, diabetes, hipertensão), gestantes e profissionais da saúde, mostraram-se mais vulneráveis à infecção pelo SARS-CoV-2. Esses indivíduos, por vezes, tiveram de encurtar seus períodos de isolamento para buscar medicação nas unidades de saúde (LOPES; COSTA, 2020). Profissionais que atuam em locais de alta exposição, como hospitais e centros de triagem, enfrentaram grandes fluxos de atendimento, aumentando ainda mais sua vulnerabilidade (CASTRO et al., 2020).

O diagnóstico laboratorial da COVID-19 foi essencial para a compreensão da dinâmica de transmissão. No entanto, no início da pandemia, a escassez de insumos comprometeu a eficácia da testagem e o atendimento à população sintomática (OPAS, 2021).

Com a evolução da pandemia, medidas restritivas passaram a ser intensificadas nas cinco regiões do país. Em 11 de março de 2020, escolas começaram a ser fechadas no estado do Rio de Janeiro, e o isolamento social, assim como a quarentena voluntária, foram instituídos nacionalmente após a identificação da transmissão comunitária (BRASIL, 2020d).

Com o avanço do contágio em cidades como Manaus, surgiram sinais de colapso na gestão hospitalar, com superlotação, escassez de leitos de UTI e aumento no número de óbitos, tornando a capital amazonense um dos epicentros da pandemia (BUENO; SOUTO; MATTA, 2021).

Diante disso, medidas emergenciais foram implementadas, como a construção de hospitais provisórios, conforme orientações da OMS. Entretanto, a falta de profissionais qualificados para operar equipamentos especializados agravou os atrasos no atendimento (CASTRO et al., 2020).

Essa dinâmica contribuiu para a disseminação progressiva da pandemia entre as

regiões do país, iniciando no Norte e se expandindo até o Sul, por volta de junho de 2020 (MOTA; TEIXEIRA, 2020). No final de dezembro do mesmo ano, os indicadores epidemiológicos apresentaram uma aparente estabilização, o que motivou o relaxamento das medidas restritivas (BUENO; SOUTO; MATTA, 2021).

Contudo, esse relaxamento teve consequências negativas, especialmente com a chegada da variante Delta, que causou um novo aumento no número de casos em todo o mundo (PATANÉ et al., 2021). No início de 2021, Manaus voltou a registrar recordes de infecções e óbitos, superando os números do ano anterior.

A alternância de epicentros entre as capitais brasileiras está relacionada à ausência de centros especializados no tratamento de doenças respiratórias. A pandemia evidenciou uma crise generalizada na saúde pública, destacando as desigualdades regionais no acesso aos serviços prestados pelo governo federal. As áreas com menor investimento mostraram-se mais suscetíveis a surtos epidêmicos (SCHMIDT et al., 2021).

Por outro lado, a chegada das vacinas representou um marco na contenção da pandemia. Em abril, a OMS convocou uma reunião emergencial com a coalizão COVAX, visando discutir formas de distribuir as vacinas de forma equitativa entre os países, independentemente de seu poder econômico (BERKLEY, 2020).

No Brasil, a Fundação Oswaldo Cruz (Fiocruz) firmou, em 8 de setembro de 2021, um acordo para importação da vacina AstraZeneca, desenvolvida pela Universidade de Oxford, com o objetivo de atender à demanda nacional e garantir a efetividade do Programa Nacional de Imunizações (PNI) (FIOCRUZ, 2020a).

Além da AstraZeneca, outras vacinas foram incorporadas gradualmente à campanha de vacinação, como Pfizer/BioNTech, Sinovac (Butantan) e Janssen, todas aprovadas pela Agência Nacional de Vigilância Sanitária (Anvisa) (BRASIL, 2022b).

Apesar disso, a eficácia dos imunizantes foi alvo de questionamentos por parte de setores da sociedade e de autoridades públicas, o que atrasou a aquisição de insumos e comprometeu a cobertura vacinal em algumas regiões (BUENO; SOUTO; MATTA, 2021).

Desde o início da pandemia de COVID-19, discutiu-se intensamente as práticas de enfrentamento e controle da doença. Nesse contexto, o SUS mostrou-se fundamental na gestão dos surtos, especialmente por meio do PNI, reafirmando seu papel essencial no atendimento da população brasileira, em especial das camadas mais vulneráveis.

3.4.1 Processo de Notificação de COVID-19

O processo de notificação e registro de casos de Síndrome Gripal no Brasil é atualmente gerenciado pelo Ministério da Saúde, por meio do sistema unificado e-SUS. Esse sistema segue diretrizes específicas para a realização da notificação de pacientes com

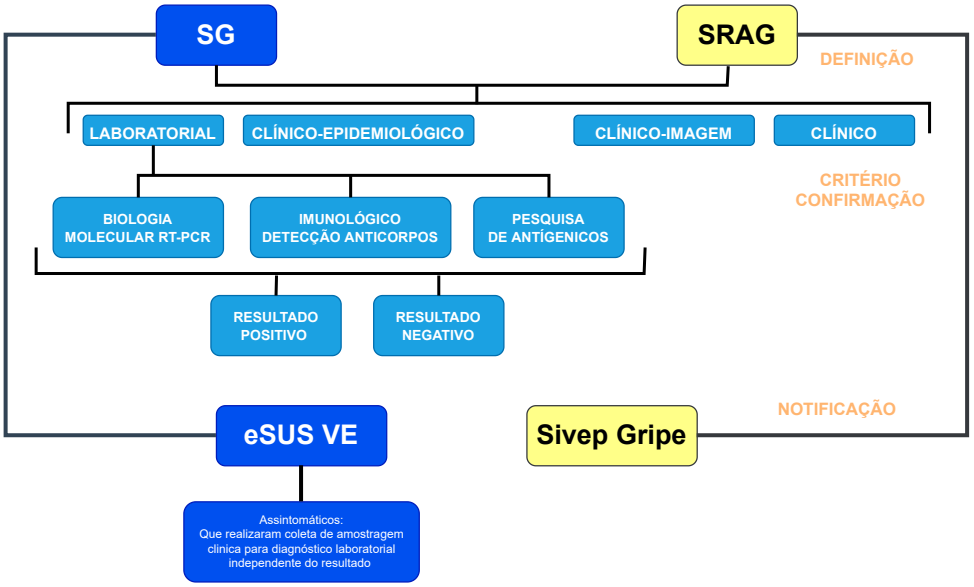
quadros respiratórios atípicos durante a triagem médica, considerando também o histórico clínico pré-sintomático relatado pelo próprio paciente.

Sintomas comuns da Síndrome Respiratória Aguda Grave (SRAG), como febre, calafrios, tosse, dor de garganta e coriza, associados a alterações no paladar e olfato, bem como sinais específicos, como dispneia, desconforto torácico, dor no peito e saturação de oxigênio inferior a 95% no ar ambiente, são indicativos de alto risco de infecção por COVID-19 (BRASIL, 2021a).

Diante da confirmação clínica ou laboratorial, os profissionais de saúde dos setores público e privado devem seguir a legislação nacional vigente, especialmente a Portaria nº 356/GM/MS, de 11 de março de 2020, que define as diretrizes obrigatórias para a notificação de casos de COVID-19. Essa norma estabelece que o registro das informações deve ser realizado em até 24 horas nos sistemas oficiais de vigilância epidemiológica (BRASIL, 2020g).

O procedimento de notificação é efetuado por meio da Ficha Individual de Notificação, apresentada no Anexo A, que reúne dados sobre sintomas, queixas, comorbidades e situação vacinal do paciente (SUS, 2020). Atualmente, essas informações são inseridas no Sistema de Informação de Vigilância Epidemiológica da Gripe (Sivep-Gripe), conforme o fluxo de investigação ilustrado na Figura 2 (BRASIL, 2020c).

Figura 2 – Critérios de confirmação e notificação de casos de COVID-19 no Brasil.



Fonte: Extraído de SUS, (2024).

Os registros relacionados à COVID-19 devem seguir os procedimentos previamente estabelecidos pela vigilância epidemiológica para gripe, influenza e outros vírus respiratórios.

Em caso de óbito, o Ministério da Saúde disponibiliza o fluxo de notificação

específico para COVID-19 (Figura 2) e orienta que o responsável pela notificação confirme as informações referentes à causa da morte por meio de documentação oficial. Esses dados devem ser inseridos no Sistema de Informações sobre Mortalidade (SIM) (BRASIL, 2022a). Informações detalhadas sobre o preenchimento e a emissão da declaração de óbito, bem como diretrizes para a codificação da causa mortis por COVID-19, estão disponíveis no site oficial do SIM[3].

3.4.2 Índice de óbitos de COVID-19

A mortalidade representa a probabilidade de um indivíduo morrer ou desenvolver uma condição adversa com potencial letal. Essa avaliação é fundamentada no número absoluto e na proporção de infecções observadas durante o mesmo período de ocorrência de um surto epidêmico.

Dessa forma, torna-se essencial mensurar a ocorrência de surtos com base no número de óbitos confirmados (YOUNGERMAN, 2008). Para isso, destacam-se duas métricas amplamente utilizadas na análise de doenças infecciosas não sazonais: a taxa de letalidade (TL) e a taxa de mortalidade (TM).

A TL, também denominada coeficiente de letalidade, refere-se à proporção de óbitos entre os indivíduos diagnosticados com a doença. Essa métrica avalia a gravidade do quadro clínico, indicando o percentual de infectados que evoluem para óbito durante um surto epidêmico. A TL é calculada pela razão entre o número de mortes e o total de casos confirmados da doença em um determinado intervalo de tempo, conforme expressa a Equação 3.1:

$$TL^{doença(\%)} = \frac{\sum óbitos^{doença}}{\sum infectados^{doença}} \times 100 \tag{3.1}$$

Por sua vez, TM corresponde ao número de óbitos atribuídos à doença em relação ao total da população exposta. Durante pandemias, esse indicador é amplamente utilizado para mensurar a incidência da doença, considerando as características sociodemográficas das regiões afetadas (YOUNGERMAN, 2008). A TM é geralmente expressa por 100 mil habitantes, conforme definido na Equação 3.2:

$$TM^{doença(\%)} = \frac{Total\ de\ óbitos}{100\ mil\ habitantes} = \frac{\sum óbitos^{doença}}{\sum população} \times 100.000 \tag{3.2}$$

Além dos indicadores de letalidade e mortalidade — fundamentais para compreender o impacto de doenças com alto potencial de disseminação , é igualmente relevante analisar os fatores sociais que favorecem a transmissibilidade. Nesse contexto, a presente análise

[3] http://sim.saude.gov.br/default.asp

incorpora o IVS, que considera variáveis estruturais como o acesso ao saneamento básico, os níveis de escolaridade e a distribuição de renda da população.

3.5 Considerações Finais

Neste capítulo, explora-se os conceitos fundamentais da Síndrome Respiratória Aguda Grave, abordando aspectos como biologia, genética, formas de transmissão e a gravidade do coronavírus. Compreender esses elementos é essencial para uma melhor compreensão do comportamento do vírus, especialmente durante surtos e epidemias descontroladas, como a pandemia de COVID-19, que representa uma séria ameaça à saúde global.

No próximo capítulo, discuti-se as ferramentas e técnicas utilizadas nos estudos de caso, com ênfase nos métodos de ciência de dados e na metodologia CRISP-DM. Essas abordagens são fundamentais para a análise dos dados considerados, incluindo o IVS, e para a obtenção de *insights* relevantes, que permitam formular conclusões significativas e estratégias de mitigação mais eficazes.

4 Análise de Dados

Este capítulo apresenta os conceitos, aplicações e técnicas que fundamentam o desenvolvimento deste trabalho. São abordadas a análise exploratória de dados, o *Big Data* e a ciência de dados, com destaque para sua relevância no cenário científico e tecnológico atual. Além disso, discute-se a técnica CRISP-DM, empregada na elaboração e estruturação dos conjuntos de dados utilizados neste estudo, bem como o uso de Redes Bayesianas.

4.1 Análise Exploratória de Dados

A análise exploratória de dados possibilita a identificação de padrões e tendências que fornecem informações valiosas para a compreensão da problemática da pesquisa, aplicando-se a conjuntos de dados transacionais ou não (GAMMA et al., 2017). Os resultados obtidos nessa exploração auxiliam na escolha da técnica mais adequada para o problema, além de fornecer subsídios essenciais para o pré-processamento dos dados e para a criação de modelos preditivos mais condizentes com a realidade do estudo (DEVORE, 2015; MEDEIROS, 2007).

Inicialmente, realiza-se uma análise estatística descritiva, na qual os dados brutos são resumidos com base em suas principais características, utilizando métricas como média, tendências, frequências e percentis (SILVA, 2018a; DEVORE, 2015). Essas métricas fornecem aos cientistas de dados informações essenciais sobre a base analisada, permitindo melhor classificação e entendimento dos dados (GAMMA et al., 2017).

O processo inicial de compreensão dos dados fundamenta-se na estatística e na teoria da probabilidade, ambas relacionadas ao conceito de *entropia* — que mede o grau de dispersão de uma variável e a quantidade de informação compartilhada entre variáveis aleatórias (KLOSTERMAN, 2020). Em síntese, a *entropia* é utilizada para reduzir as incertezas associadas às relações entre valores, representadas por $X \rightarrow Y$.

Essa análise exploratória depende da organização estrutural dos dados. Em casos de bancos de dados descentralizados, o cientista de dados realiza a integração, identificando coleções de objetos que devem ser combinadas. Esse procedimento é conhecido como problema de reconhecimento de entidade (HURWITZ et al., 2013a; DEGAN, 2005).

Entretanto, a integração de dados enfrenta desafios como divergências nas nomenclaturas das propriedades entre bases distintas, dificultando a atualização das informações — fenômeno conhecido como diacronia temporal (GAMMA et al., 2017). Uma estratégia para mitigar esses problemas é o uso de metadados, que descrevem as principais características dos dados.

Outro desafio frequente é o desbalanceamento de dados, especialmente em bases reais. Isso ocorre quando a distribuição entre classes é desigual, favorecendo a predição das classes majoritárias e prejudicando o desempenho para as minoritárias (YU; ZHOU, 2021; GHASEMIAN; HOSSEINMARDI; CLAUSET, 2020). Técnicas como a geração de dados artificiais e o uso de métricas alternativas são aplicadas para contornar esse problema.

No contexto da estrutura e organização dos dados, o uso de *Big Data* tornou-se fundamental na última década. O volume crescente de informações provenientes de dispositivos móveis, *smart cities*, sensores RFID e outros meios impõe novos desafios às estruturas tradicionais de armazenamento e processamento de dados (GAMMA et al., 2017; MARZ; WARREN, 2015).

Além disso, a análise exploratória desempenha papel essencial no modelo CRISP-DM, permitindo uma compreensão aprofundada dos dados antes da modelagem preditiva. Estatísticas como média, variância e mediana, e visualizações como histogramas, gráficos de dispersão e *boxplots*, são amplamente utilizadas para detectar padrões, tendências e anomalias.

4.2 Big Data

O termo *Big Data* refere-se a um conceito abstrato que ganhou destaque a partir de meados de 2010, em razão do crescimento exponencial da geração de dados por instituições públicas e privadas (Philip Chen; ZHANG, 2014). Autores como Deepa et al. (2022), Davenport (2014) e Green-Jr e Chow-White (2013) destacam o *Big Data* como uma ferramenta emergente para apoiar a tomada de decisão corporativa. Em 2012, estimava-se que essas instituições geravam mais de 2,8 trilhões de gigabytes de dados (SILVEIRA; MARCOLIN; FREITAS, 2015; DAVENPORT, 2014). Desde então, a produção diária de dados é estimada em aproximadamente 2,5 *exabytes* ($2{,}5 \times 10^{18}$ bytes) (SOUZA, 2023).

O *Big Data* pode ser definido como um conjunto de dados de grande volume e alta complexidade, que não pode ser processado de maneira eficiente por técnicas tradicionais (KRISHNAN, 2019; DAVENPORT, 2014). Além do crescimento no volume de dados, o conceito envolve desafios operacionais relacionados ao armazenamento e ao processamento dessas informações. Nesse contexto, destacam-se cinco características fundamentais do *Big Data*: volume, velocidade, variedade, veracidade e valor (DEEPA et al., 2022; KLAUS-DIETER, 2020; SOUZA, 2023; HURWITZ et al., 2013b).

A seguir, essas características são detalhadas:

- **Volume**: refere-se à enorme quantidade de dados gerados continuamente por diversas fontes, como transações comerciais, redes sociais e projetos acadêmicos.

- **Velocidade**: diz respeito à rapidez com que os dados são gerados e processados, exigindo sistemas capazes de lidar com grandes fluxos de informação em tempo real (SOUZA, 2023).
- **Variedade**: refere-se à diversidade dos dados, que podem assumir diferentes formatos, como textos, imagens, vídeos, dados estruturados e não estruturados (SOUZA, 2023).
- **Veracidade**: relaciona-se à confiabilidade e à qualidade dos dados coletados. Em ambientes como as redes sociais, onde informações podem ser manipuladas, a veracidade se torna um desafio (GARCÍA LOZANO et al., 2020).
- **Valor**: refere-se à capacidade dos dados de gerar informações úteis e insights relevantes para a tomada de decisão (GUNTHER et al., 2017).

Diante disso, o investimento em infraestrutura de *Big Data* é essencial. A expressão *garbage in, garbage out* — entrada ruim, saída ruim — resume a importância de inserir dados de alta qualidade para obter resultados analíticos confiáveis (SOUZA, 2023).

Com o avanço tecnológico e a digitalização de processos, setores como saúde, educação, engenharia e política passaram a gerar e a depender cada vez mais de dados em grande escala (DAVENPORT, 2014). Assim, instituições públicas e privadas que utilizam soluções tradicionais precisam buscar novas estratégias para integrar o *Big Data* em seus processos organizacionais, otimizando a análise e a geração de valor a partir das informações disponíveis (KLAUS-DIETER, 2020).

4.3 Ciência de Dados

Diante da crescente geração de dados e dos desafios associados ao seu processamento e análise, torna-se fundamental compreender as abordagens e métodos que possibilitam a extração de conhecimento a partir desses grandes volumes de informações. Nesse contexto, a ciência de dados emerge como uma disciplina interdisciplinar que integra estatística, computação e conhecimento de domínio, oferecendo ferramentas e metodologias para transformar dados brutos em informações relevantes e acionáveis.

A ciência de dados está relacionada ao processo de aquisição e estruturação do conhecimento, sendo considerada um campo relativamente recente em comparação às abordagens tradicionais de análise de dados. Essa área ganhou destaque devido ao grande volume de dados gerados em diversos setores, como comércio, indústria, agronegócio, telecomunicações, engenharia e medicina (LIM; KWANG-JAE; MAGLIO, 2018).

O avanço dos meios de comunicação e o surgimento de ambientes como as *Smart Cities*, capazes de gerar enormes quantidades de dados diariamente, reforçaram a necessidade

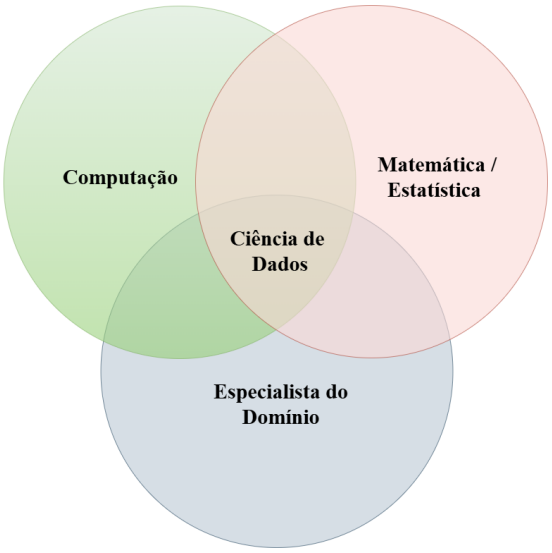
de técnicas mais robustas para análise e interpretação desses dados (FRANK; WITTEN, 2000).

Assim, a ciência de dados é compreendida como a disciplina que busca extrair informações relevantes e novos saberes a partir de grandes volumes de dados (CAO, 2017). Para tanto, ela envolve uma série de competências, tais como álgebra linear, estatística, modelagem matemática, técnicas de *machine learning* e visualização de dados (BOSCHETTI; MASSARON, 2015). Ou seja, a ciência de dados não se limita à descoberta do conhecimento, exigindo também uma sólida formação técnica e prática.

Autores como Cady (2017) e VanderPlas (2016) ressaltam que a ciência de dados deve integrar três elementos fundamentais: o domínio computacional, as ciências exatas (matemática e estatística) e o conhecimento especializado na área de aplicação. Essa interação entre diferentes campos do saber configura a ciência de dados como uma disciplina de natureza interdisciplinar, adaptável aos novos desafios da sociedade da informação (CADY, 2017).

A Figura 3 ilustra, por meio de um diagrama de Venn, a relação entre essas três competências essenciais para a formação e prática do cientista de dados.

Figura 3 – Competências essenciais para a aplicação da ciência de dados.



Fonte: Elaborado pelo autor, segundo o modelo Cady (2017).

Quando o objetivo do estudo é a descoberta de padrões, podem ser aplicadas técnicas de mineração de dados, utilizando metodologias como o (*Knowledge Discovery in Databases*) KDD ou o CRISP-DM . Este último será abordado na próxima seção, por sua importância no desenvolvimento desta pesquisa.

4.4 CRISP-DM

O CRISP-DM é uma abordagem amplamente utilizada para a resolução de problemas em ciência de dados, com foco na descoberta de padrões, identificação de relações e desenvolvimento de modelos preditivos. Trata-se de um ciclo de vida estruturado em fases, mas concebido de forma flexível. Desenvolvido no início dos anos 2000 pelas empresas Daimler Chrysler, SPSS e NCR (CHAPMAN et al., 2000), o CRISP-DM destaca-se por permitir que suas etapas não sejam executadas de maneira estritamente sequencial. Ao contrário, o modelo é cíclico, possibilitando o retorno a fases anteriores sempre que necessário para aprimorar os resultados e manter o alinhamento com os objetivos do estudo.

O CRISP-DM é composto por seis fases principais:

- **Entendimento do Negócio**: Fase inicial fundamentada na definição clara dos objetivos do problema a ser solucionado. A partir desse entendimento, constrói-se um plano preliminar de análise de dados que orientará todas as fases subsequentes. Essa etapa contribui para o caráter cíclico do CRISP-DM, permitindo revisões contínuas.
- **Entendimento dos Dados**: Refere-se à coleta inicial, descrição e exploração dos dados, com o objetivo de compreender suas propriedades básicas, identificar problemas de qualidade e formular hipóteses iniciais sobre possíveis relações relevantes.
- **Preparação dos Dados**: Nesta etapa, os dados são processados e transformados no conjunto final a ser utilizado na modelagem. Isso inclui seleção de atributos relevantes, tratamento de valores faltantes, normalização, transformação de variáveis e criação de novas variáveis, assegurando a integridade da informação.
- **Modelagem**: Aplicação de técnicas de mineração de dados, como classificação, regressão, associação ou agrupamento. A seleção da técnica apropriada depende do tipo de problema, dos dados disponíveis e dos objetivos da análise.
- **Avaliação**: Nesta etapa, avaliam-se os resultados obtidos na modelagem, verificando se o modelo atende aos objetivos definidos inicialmente. Utilizam-se métricas de desempenho como acurácia, precisão, recall e F1-score, além da análise crítica dos resultados em relação ao problema de negócio.
- **Implantação**: Implementação prática do modelo em ambiente operacional ou aplicação dos resultados obtidos para suportar a tomada de decisão. É importante ressaltar que, devido ao caráter dinâmico dos dados, o modelo implantado pode exigir reavaliações e ajustes periódicos.

No contexto desta tese, o CRISP-DM orientou a estruturação do processo de análise dos dados sociodemográficos e educacionais. Especificamente, foram utilizadas

Redes Bayesianas na etapa de modelagem, dada sua capacidade de representar relações causais e lidar com dados incompletos, como será detalhado na seção seguinte.

4.5 Redes Bayesianas

As RBs são modelos gráficos probabilísticos baseados no Teorema de Bayes, proposto por Thomas Bayes em 1763. Seu principal objetivo é representar e inferir dependências condicionais entre um conjunto de variáveis aleatórias observadas (ULAK; YAZICI; ZHANG, 2020). A estrutura dessas redes é composta por um grafo direcionado acíclico (DAG — *Directed Acyclic Graph*), em que os nós representam variáveis e as arestas indicam relações de dependência entre elas.

A Equação 4.1 expressa o Teorema de Bayes, aplicado para calcular a probabilidade condicional de um evento A dado que outro evento B tenha ocorrido:

$$P(A|B) = \frac{P(B|A) \cdot P(A)}{P(B)} \tag{4.1}$$

- $P(A|B)$ representa a probabilidade condicional de A dado B;
- $P(B|A)$ é a probabilidade de B dado A;
- $P(A)$ é a probabilidade a priori de A;
- $P(B)$ é a probabilidade total de B.

As RBs possibilitam modelar relações causais e realizar inferências probabilísticas com base em evidências observadas, mesmo em contextos com dados incompletos. Essas características tornam as RBs especialmente adequadas para estudos que envolvem incerteza e que demandam explicabilidade nos modelos — como ocorre na análise de fenômenos educacionais e sociais.

4.5.1 Justificativa para o uso das Redes Bayesianas

A escolha das Redes Bayesianas nesta pesquisa justifica-se pela natureza interpretável e causal da técnica. Diferentemente de outras abordagens em inteligência artificial, como redes neurais profundas, máquinas de vetores de suporte ou árvores de decisão, as RBs permitem representar explicitamente relações de dependência entre variáveis e possibilitam inferências causais, desde que respeitados os pressupostos do modelo (KORB; NICHOLSON, 2010a; PEARL, 2009).

Outra vantagem relevante é a capacidade das RBs de trabalhar com dados incompletos — uma característica comum em bases públicas utilizadas nesta tese, como

o ENEM, a PNADC e o Censo Escolar. Além disso, os modelos gerados pelas RBs são transparentes, permitindo que analistas e gestores compreendam as implicações das variáveis no comportamento do sistema analisado, o que é essencial para subsidiar políticas públicas baseadas em evidências.

4.5.2 Limitações das Redes Bayesianas

Apesar das vantagens, as RBs também apresentam limitações. A principal delas está relacionada ao pressuposto de independência condicional entre variáveis, que pode não ser plenamente observado em situações reais. Modelos mal estruturados podem gerar inferências incorretas, comprometendo a validade das análises (DARWICHE, 2009).

Outro ponto importante refere-se à distinção entre causalidade e correlação. Embora as RBs permitam inferências causais sob certos pressupostos, sua aplicação em bases observacionais exige cautela, uma vez que relações estatísticas nem sempre implicam causa e efeito. Assim, a interpretação dos resultados deve considerar o contexto dos dados e as limitações do modelo.

Por fim, a construção e o treinamento de RBs com muitos nós e múltiplas dependências podem se tornar computacionalmente custosos, exigindo técnicas específicas de aprendizado estrutural e de parametrização para garantir eficiência e escalabilidade.

4.5.3 Estrutura de Redes Bayesianas

As RBs podem ser representadas por um DAG. Nesse tipo de grafo, as arestas indicam conexões unidirecionais entre os nós, sem a formação de ciclos. Cada nó representa uma variável aleatória do modelo, e as arestas indicam a existência de relações de dependência condicional entre essas variáveis.

Segundo a teoria dos grafos, um DAG garante que, ao percorrer os arcos a partir de qualquer nó inicial, não se retorna a um nó previamente visitado. Essa propriedade é essencial para que as inferências probabilísticas possam ser realizadas de forma coerente.

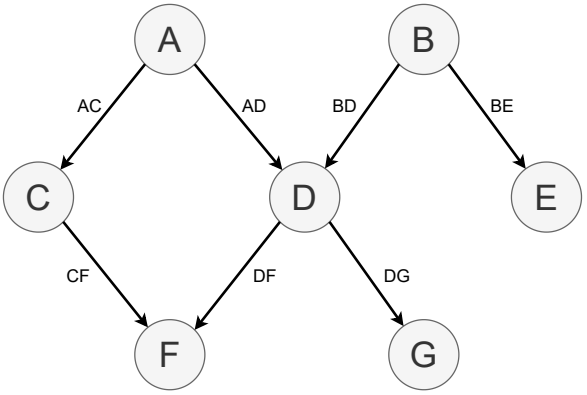
A representação formal de uma RB considera um conjunto de variáveis aleatórias $A = \{A_1, A_2, A_3, \ldots, A_n\}$ associadas a um conjunto de nós n_j. A distribuição conjunta das variáveis pode ser decomposta segundo a Equação 4.2:

$$P(A_1, A_2, \ldots, A_n) = \prod_{j=1}^{n} P(A_j \mid \text{Pais}(A_j)) \tag{4.2}$$

Essa equação expressa a decomposição da distribuição conjunta com base nas dependências condicionais entre cada variável A_j e seus respectivos nós pais.

A Figura 4 ilustra uma estrutura hierárquica de RB simplificada, composta por oito nós e sete arestas. Os nós A e B são pais de C; A também é pai de D, e B, de E. Já os nós F e G são filhos de C e D, conectados por meio das arestas $\overline{CF}$, $\overline{DF}$ e $\overline{DG}$.

Figura 4 – Representação de um grafo DAG em uma RB.

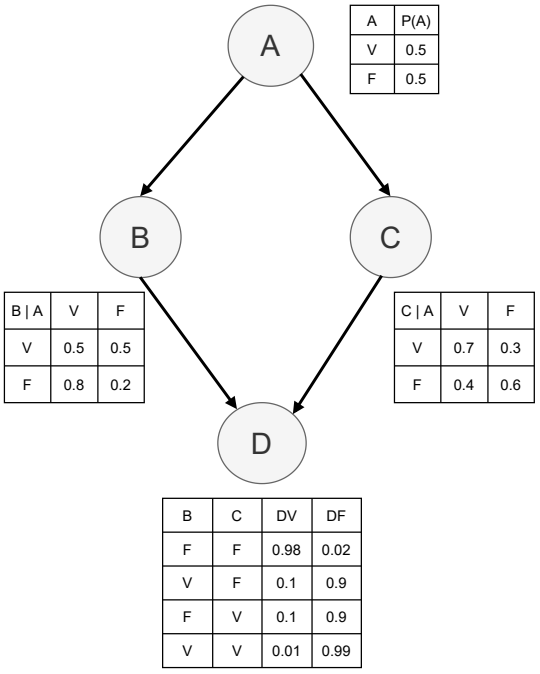


Fonte: Adaptado de Tan, Steinbach e Kumar (2009).

Outro componente essencial nas RBs é a Tabela de Probabilidade Associativa (TPA), que representa a distribuição condicional de probabilidade de uma variável em relação a seus pais. Essas tabelas devem ser construídas de forma que a soma das probabilidades de cada linha seja igual a 1.

A Figura 5 apresenta uma RB hipotética com quatro nós, onde A é pai de B e C, e ambos são pais de D — que se comporta como um nó folha.

Figura 5 – Exemplo de RB hipotética com TPA.



A	P(A)
V	0.5
F	0.5

B | A	V	F
V	0.5	0.5
F	0.8	0.2

C | A	V	F
V	0.7	0.3
F	0.4	0.6

B	C	DV	DF
F	F	0.98	0.02
V	F	0.1	0.9
F	V	0.1	0.9
V	V	0.01	0.99

Fonte: Adaptado de Russell e Norvig (2020).

A construção das TPAs pode ser computacionalmente intensiva, especialmente em redes com muitos nós e combinações de dependências. Contudo, elas são fundamentais para viabilizar as inferências probabilísticas e tornar o modelo capaz de refletir o comportamento real das variáveis representadas.

4.5.4 Inferências Bayesianas

Nas RBs, o termo inferência refere-se ao processo de atualização das distribuições de probabilidade a posteriori, em resposta à incorporação de novas evidências. Em outras palavras, à medida que novos dados são observados, as crenças sobre o estado das variáveis na rede são recalculadas para refletir essas novas informações (KORB; NICHOLSON, 2010b).

Esse processo é fundamental para o funcionamento das RBs e baseia-se no Teorema de Bayes, aplicado sistematicamente à estrutura probabilística do modelo. Independentemente da complexidade da rede, a inferência probabilística busca calcular a probabilidade de interesse condicionada às evidências disponíveis.

A atualização da distribuição de probabilidade pode ser expressa pela Equação 4.3:

$$P(A) = \alpha \cdot P(B_n) \cdot P(B_n) \tag{4.3}$$

Em que α representa o fator de normalização, garantindo que a soma das probabilidades seja igual a 1.

De forma mais geral, a distribuição de probabilidade conjunta para um conjunto de variáveis observadas e não observadas pode ser expressa pela Equação 4.4:

$$P(\overline{X}_n) = \sum_{i=1}^{n} P(X_i \mid \text{pais}(X_i)) \tag{4.4}$$

Nesse contexto:

- X_i são as variáveis da rede; - pais(X_i) indica o conjunto de nós pais de X_i.

A complexidade do processo de inferência é altamente dependente da estrutura da RB. Redes com muitos nós, múltiplas dependências ou estruturas densamente conectadas tendem a exigir algoritmos mais sofisticados e técnicas de otimização para viabilizar a inferência em tempo hábil.

Segundo Marques e Dutra (1999), a ausência de uma estrutura eficiente pode resultar em modelos excessivamente complexos, dificultando tanto a interpretação quanto a capacidade preditiva da rede. Assim, a definição adequada das relações de dependência entre as variáveis é crucial para o desempenho do processo inferencial.

Nesta tese, a inferência foi utilizada para estimar a probabilidade de eventos educacionais e de saúde com base em variáveis observadas em bases públicas. Essa abordagem permitiu uma análise probabilística mais robusta dos fatores associados à vulnerabilidade social e aos impactos da pandemia de COVID-19.

4.5.5 Aprendizagem de Estrutura

O processo de aprendizagem em RBs visa encontrar a melhor configuração estrutural da rede a partir de um conjunto de dados, de forma que a distribuição de probabilidade modelada represente adequadamente as relações de dependência ou independência entre as variáveis envolvidas.

De acordo com Hruschka (2003), a estimação da estrutura de uma RB pode ser abordada sob duas óticas principais: (i) métodos de busca heurística orientados por pontuação, e (ii) métodos baseados na análise de independência condicional. Adicionalmente, Magalhães (2007) propõem uma terceira abordagem, baseada em métodos híbridos que combinam características das duas estratégias anteriores.

Na busca heurística, o processo é guiado pela maximização de uma função de pontuação (*score*). Normalmente, parte-se de uma rede inicial sem conexões (ou com uma configuração simples), adicionando ou removendo arcos entre variáveis para maximizar o *score* associado à estrutura gerada. No entanto, essa abordagem é sensível à ordem das variáveis e pode ser influenciada por armadilhas de ótimos locais.

Por outro lado, a análise de independência condicional fundamenta-se nos conceitos de separação d-separação em grafos direcionados. Segundo Neapolitan (2004), em uma rede $\xi = (V, A)$ com variáveis X, Y e Z, são definidos três tipos principais de conexões:

1. *Head-to-tail*: $X \longrightarrow Y \longrightarrow Z$;
2. *Tail-to-tail*: $X \longleftarrow Y \longrightarrow Z$;
3. *Head-to-head*: $X \longrightarrow Y \longleftarrow Z$.

Nos dois primeiros casos, a variável intermediária Y bloqueia o caminho se for observada; no terceiro caso (*head-to-head*), o caminho só é bloqueado se nem Y nem seus descendentes forem observados. A identificação dessas relações permite reduzir o espaço de busca estrutural, otimizando o aprendizado da rede.

Entre os métodos híbridos disponíveis, destaca-se o pacote PGMPY da linguagem Python, utilizado neste trabalho para a construção de modelos com aprendizagem estrutural eficiente.

4.5.5.1 *Hill-Climb Search*

O algoritmo *Hill-Climb Search* é uma técnica de busca heurística que explora o espaço de possíveis estruturas da RB, movendo-se na direção que aumenta o valor da função objetivo (*score*). Seu funcionamento pode ser resumido em etapas:

- **Inicialização**: define-se uma solução inicial x_{atual};
- **Geração de Vizinhança**: pequenas modificações em x_{atual} geram novas soluções candidatas;
- **Avaliação**: cada solução candidata é avaliada segundo a função objetivo $f(x)$;
- **Movimentação**: seleciona-se a solução que otimiza (maximiza ou minimiza) $f(x)$, conforme as Equações 4.5 e 4.6:

$$x_{\text{proximo}} = \arg \max_{x \in N(x_{\text{atual}})} f(x) \tag{4.5}$$

$$x_{\text{proximo}} = \arg \min_{x \in N(x_{\text{atual}})} f(x) \tag{4.6}$$

- **Atualização**: se $f(x_{\text{proximo}})$ supera $f(x_{\text{atual}})$, então atualiza-se x_{atual}:

$$x_{\text{atual}} = x_{\text{proximo}} \tag{4.7}$$

Caso não haja melhora, o algoritmo termina.

4.5.5.2 Algoritmo K2

O algoritmo K2 é uma estratégia baseada em ordenação e busca orientada por pontuação para estimar a estrutura de RBs. Para cada variável, busca-se a melhor combinação de nós pais que maximize a probabilidade do conjunto de dados observado, dada a estrutura.

A função de pontuação utilizada no K2 é expressa pela Equação 4.8:

$$P(\xi|X) = \prod_{i=1}^{m} \prod_{j=1}^{q_i} \frac{(r_i - 1)!}{(N_{ij} + r_i - 1)!} \prod_{k=1}^{r_i} N_{ijk}! \tag{4.8}$$

Onde:

- X representa o conjunto de observações, - m é o número de variáveis, - r_i é o número de estados possíveis da variável X_i, - q_i é o número de configurações possíveis dos pais de X_i, - N_{ijk} representa o número de vezes em que X_i assume o estado k, condicionado à configuração j dos pais.

4.5.5.3 *Variable Elimination*

O algoritmo *Variable Elimination* é utilizado para calcular distribuições marginais em RBs de forma eficiente. Dado um conjunto de variáveis $X = \{X_1, X_2, \ldots, X_n\}$, a distribuição conjunta é expressa como:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid \text{pais}(X_i)) \tag{4.9}$$

Para calcular a distribuição marginal de uma variável de interesse $X_{\text{interesse}}$, marginaliza-se sobre as variáveis restantes:

$$P(X_{\text{interesse}}) = \sum_{\text{restante}} P(X_1, X_2, \ldots, X_n) \tag{4.10}$$

Essa técnica é fundamental para inferência eficiente em grandes redes, reduzindo o custo computacional associado à propagação de probabilidades.

A seguir, será apresentado o IVS, utilizado nesta tese como parâmetro de análise sociodemográfica associado aos efeitos da pandemia e à estrutura de desigualdades nos territórios analisados.

4.6 IVS

O IVS é um indicador sintético composto por três subíndices principais: infraestrutura urbana, capital humano e renda e trabalho. O subíndice de infraestrutura urbana abrange variáveis como acesso à água tratada, coleta de lixo, esgotamento sanitário e fornecimento de energia elétrica (ATLAS IVS, 2025). Esses elementos são fundamentais para garantir condições adequadas de vida e reduzir riscos à saúde pública. A carência desses serviços pode agravar os impactos de surtos epidêmicos em áreas vulneráveis. Por isso, sua mensuração é essencial para orientar ações públicas de prevenção e mitigação.

O subíndice de capital humano considera aspectos relacionados à educação e à saúde da população. Entre os indicadores analisados estão a taxa de analfabetismo, a frequência escolar e a expectativa de vida ao nascer (ATLAS IVS, 2025). Tais variáveis influenciam diretamente a capacidade das pessoas de compreender e adotar medidas protetivas em situações de crise sanitária. Além disso, refletem o grau de desenvolvimento social de diferentes regiões do país. Assim, esse subíndice contribui para identificar territórios com maiores fragilidades sociais.

O subíndice de renda e trabalho mede a inserção econômica da população em relação ao mercado formal e ao acesso a renda. São considerados indicadores como taxa de desemprego, informalidade no trabalho e renda domiciliar per capita (ATLAS IVS,

2025). A vulnerabilidade econômica afeta diretamente a capacidade de acesso a serviços essenciais e a resiliência das famílias frente a crises. Ao integrar os três subíndices, o IVS permite diagnosticar desigualdades estruturais com base empírica. Dessa forma, torna-se uma ferramenta estratégica para a formulação de políticas públicas mais equitativas.

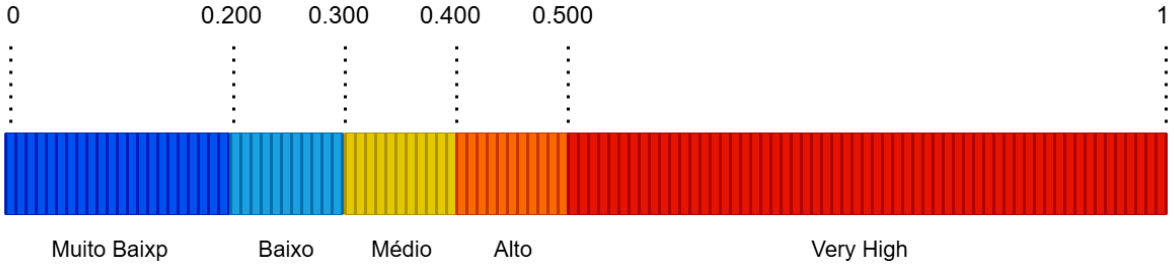
Figura 6 – Dimensões sociais utilizadas na composição do IVS.

Fonte: Adaptado de ATLAS IVS (2025).

Cada subíndice que compõe o IVS é ponderado com base em pesos específicos, definidos pelo Instituto de Pesquisa Econômica Aplicada (IPEA), segundo diretrizes metodológicas do Atlas do Desenvolvimento Humano (ADH/IBGE). Esses pesos refletem a importância relativa de cada dimensão e são ajustados de acordo com o nível de desenvolvimento humano observado no ano de referência. Após o ajuste, todos os valores são normalizados em uma escala de 0 a 1 (ATLAS IVS, 2025).

A classificação do IVS é realizada por meio de faixas de avaliação que variam de "muito baixa"a "muito alta"vulnerabilidade social, conforme ilustrado na Figura 7.

Figura 7 – Faixas de avaliação do IVS.



Fonte: Adaptado de Atlas IVS (2024).

O cálculo do IVS parte da obtenção do percentual de ocorrência de uma variável ($valor\%$) no tempo t, considerando a média ($mean$) e o desvio padrão (std) da série histórica

até o período $t-1$. A normalização é realizada conforme a Equação 4.11:

$$D = \frac{valor\%}{mean + 2 \cdot std} \tag{4.11}$$

Em seguida, os valores normalizados de cada dimensão são ponderados pelos respectivos pesos, originando os subíndices (SBI), de acordo com a Equação 4.12:

$$SBI_i = \prod_{i=1}^{n} (D_i \cdot peso_i) \tag{4.12}$$

O IVS final é calculado pela média aritmética dos três subíndices, como indicado na Equação 4.13:

$$IVS = \frac{\sum_{i=1}^{3} SBI_i}{3} \tag{4.13}$$

O IVS desempenha papel essencial na identificação de carências estruturais no acesso a serviços públicos fundamentais, como saúde, educação, saneamento básico e segurança — garantidos pela Constituição Federal de 1988. Dessa forma, trata-se de um indicador estratégico para a formulação de políticas públicas voltadas à redução das desigualdades e à promoção da qualidade de vida nas regiões mais vulneráveis do país.

No contexto desta tese, o IVS é utilizado como variável central na análise da relação entre infraestrutura urbana e os impactos da pandemia sobre os indicadores educacionais e de saúde. Sua adoção permite mensurar, de forma integrada, a influência das dimensões sociais sobre os desfechos observados durante o período analisado.

4.7 Considerações Finais

Neste capítulo, foi apresentada a fundamentação teórica que sustenta o desenvolvimento desta pesquisa. Inicialmente, foram contextualizados os conceitos de *Big Data* e ciência de dados, com ênfase em sua natureza interdisciplinar e na importância desses campos para a obtenção de informações em grandes volumes de dados. Em seguida, exploraram-se metodologias voltadas para a descoberta de conhecimento, com destaque para o modelo CRISP-DM, cujas etapas operacionais foram detalhadas, além da introdução à técnica das RBs.

Além disso, considera-se os fatores socioeconômicos que influenciam a vulnerabilidade das populações, com destaque para o IVS. Esse indicador, ao avaliar condições de vida relacionadas à infraestrutura urbana, à educação, à renda e ao acesso a serviços essenciais, contribui para identificar as áreas mais afetadas pela pandemia e compreender as desigualdades no impacto da doença. No próximo capítulo, será descrita a metodologia adotada neste trabalho, considerando os estudos de caso analisados.

5 Materiais e Métodos

Neste capítulo, apresentam-se as ferramentas computacionais utilizadas no desenvolvimento deste trabalho, amplamente adotadas para otimizar o tempo de processamento em atividades que envolvem determinados tipos de análise de dados. Essas ferramentas também oferecem recursos que facilitam a visualização e a manipulação das informações de forma mais intuitiva. Além disso, descreve-se o método aplicado, abrangendo desde o pré-processamento até o pós-processamento dos dados.

5.1 Ambiente Computacional

Ao direcionarmos a atenção para a análise de dados, é fundamental que o ambiente computacional utilizado seja projetado para minimizar o esforço do pesquisador. Tal necessidade se justifica pelo fato de que a etapa de pré-processamento pode representar até 90% do esforço total envolvido na aplicação proposta. Para otimizar esse processo, diversas tecnologias têm se destacado por sua facilidade de uso e eficiência, como *Jupyter Lab*, *Matplotlib*, *Miniconda*, *Pandas*, *Python*, entre outras. A seguir, apresentamos uma breve descrição do ambiente computacional utilizado neste estudo:

- O *Jupyter Lab V- 3.6.8* é uma plataforma de desenvolvimento que oferece aos usuários maior interatividade e flexibilidade, além de melhor organização de arquivos em um ambiente unificado para visualização e manipulação de dados. Outra característica relevante é o suporte a diferentes tipos de arquivos, como *.csv*, *.json*, *.md* (*Markdown*), *.pdf*, *.vega*, *.vega-lite*, entre outros (KLUYVER et al., 2016).

- *Matplotlib V- 3.5.3* é uma biblioteca de visualização bidimensional para *Python*, capaz de gerar gráficos de alta qualidade em diversos formatos, operando em um ambiente interativo e independente da plataforma de desenvolvimento. Está disponível para uso em *scripts Python, Shell Python, IPython, Jupyter Notebook* e suas variantes, além de servidores de aplicações *web* (HUNTER, 2007).

- O *Miniconda V- 4.11.0* é uma distribuição mínima e gratuita do Anaconda, que contém apenas o interpretador *Python* e o gerenciador de pacotes *conda*. Por ser mais leve, oferece aos usuários maior controle sobre os ambientes e os aplicativos instalados, permitindo a utilização apenas dos pacotes necessários para a análise. Além disso, possibilita o gerenciamento eficiente de pacotes por meio do próprio *conda* (ANACONDA INC, 2020).

- *Pandas V- 1.3.6* é uma biblioteca da linguagem *Python* que fornece estruturas de dados rápidas, flexíveis e expressivas, projetadas para tornar o trabalho com dados "relacionais" ou "rotulados" mais intuitivo e eficiente. Seu principal objetivo é servir como um componente essencial para a análise prática de dados. Além disso, busca ser uma ferramenta de manipulação de dados poderosa, de código aberto, flexível e acessível (MCKINNEY, 2010).

- *Python V- 3.7.16* é uma linguagem de programação de código aberto, interpretada, imperativa, orientada a objetos e de tipagem dinâmica. Introduzida em 1991 pelo matemático Guido van Rossum, tem sido amplamente utilizada em tarefas como processamento de texto, análise de dados científicos e desenvolvimento de aplicações *web* dinâmicas via *Common Gateway Interface* (CGI) (VAN ROSSUM; DRAKE JR, 1995). No contexto deste trabalho, *Python* foi escolhida como linguagem principal devido à sua simplicidade, facilidade de codificação e ampla gama de bibliotecas voltadas à análise de dados.

- *GeoPandas V-0.14.2* é uma biblioteca de código aberto desenvolvida para facilitar a manipulação de dados geoespaciais em *Python*. Ela estende a biblioteca *pandas* ao incorporar tipos geométricos, como pontos, polígonos e linhas, baseando-se na integração com bibliotecas como *shapely*, *fiona* e *pyproj*. Com isso, torna-se possível executar operações espaciais, como interseções, sobreposições e projeções cartográficas, diretamente sobre estruturas tabulares (JORDAHL et al., 2020). No contexto deste trabalho, *GeoPandas* foi empregada na organização, filtragem e visualização de dados geográficos associados aos municípios e regiões do estado do Pará.

- *PGMPY V-0.1.23* (*Python Library for Probabilistic Graphical Models*) é uma biblioteca especializada na construção, treinamento e inferência de Modelos Gráficos Probabilísticos, incluindo Redes Bayesianas e Redes de Markov. Desenvolvida com foco em pesquisa e aplicações práticas, a *pgmpy* permite tanto a definição manual de estruturas quanto o aprendizado a partir de dados, além de oferecer suporte a algoritmos de inferência exata e aproximada (ANKAN; TEXTOR, 2024).

- *pyAgrum V-1.1.0* é uma biblioteca em *Python* voltada à modelagem, inferência e visualização de Redes Bayesianas e Diagramas de Influência. Baseada no *framework aGrUM* (implementado em C++), a *pyAgrum* oferece uma interface poderosa para construção de modelos probabilísticos, aprendizado estrutural, eliminação de variáveis e simulação de intervenções causais (DUCAMP; GONZALES; WUILLEMIN, 2020).

- *GitHub* é uma plataforma online de hospedagem e versionamento de código, amplamente utilizada em projetos científicos e tecnológicos. Baseada no sistema *Git*, ela facilita a colaboração entre pesquisadores, assegura o controle de versões e promove a transparência e a reprodutibilidade de experimentos. Seu uso tem se destacado na

ciência de dados, permitindo o compartilhamento aberto de *scripts*, bases de dados e modelos computacionais.

Os links disponibilizados no *GitHub* correspondem aos projetos desenvolvidos para os dois estudos de caso. O Estudo de Caso I, voltado para a análise da educação [1], e o Estudo de Caso II [2], focado na análise sociodemográfica, contam com repositórios que reúnem os códigos, bases de dados e documentações utilizados para a realização das respectivas análises.

5.2 Abordagem Metodológica

A metodologia adotada nesta qualificação envolve a aplicação do processo CRISP-DM como abordagem principal para a extração de conhecimento a partir de bases de dados, tendo como um de seus objetivos o uso das informações obtidas para subsidiar a tomada de decisão. Os dados utilizados como objeto de estudo nesta pesquisa são provenientes de instituições governamentais brasileiras, tais como Atlas IVS, Brasil.IO, Censo Demográfico, DATASUS, ENEM, IBGE, PNAD Contínua, SIDRA, entre outras, representando bases públicas de livre acesso. Diante do contexto analisado, o processo de investigação dos dados sociodemográficos é estruturado em duas frentes empíricas principais.

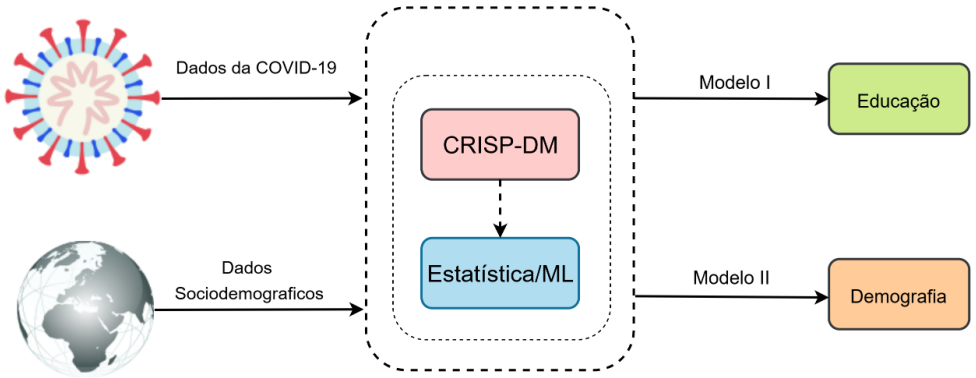
A primeira análise tem como objetivo explorar a relação entre fatores sociais e o aumento dos casos confirmados de COVID-19 no Brasil. Para isso, utilizam-se como referência as dimensões do IVS (ATLAS IVS, 2025), permitindo avaliar de que forma desigualdades socioeconômicas influenciaram a propagação do vírus. Essa abordagem está alinhada à hipótese de que a desigualdade e a vulnerabilidade social desempenham um papel central no agravamento dos surtos epidêmicos. Ademais, a análise de dados públicos sobre esses fatores pode servir de base para a formulação de políticas públicas mais eficazes, ajustadas às diferentes realidades sociais.

A segunda análise concentra-se no impacto da pandemia sobre a educação, com foco nos estudantes do último ano do ensino médio e concluintes. O objetivo é compreender os efeitos da COVID-19 sobre a formação desses alunos, com base no desempenho obtido no ENEM. Este estudo empírico busca testar a hipótese de que as classes sociais mais baixas foram as mais afetadas durante a pandemia, especialmente em razão da falta de acesso a serviços essenciais, como saúde e educação. A desigualdade na infraestrutura educacional e as dificuldades de adaptação ao ensino remoto agravaram as condições de aprendizagem, contribuindo para um impacto negativo no desempenho acadêmico dos estudantes mais vulneráveis.

[1] https://github.com/sandioms/open-enem/tree/main
[2] https://github.com/sandioms/ivs-capitais

Embora os estudos de caso apresentem abordagens distintas, ambos percorreram, inicialmente, todas as etapas da fase de entendimento do problema, destacando as informações mais relevantes discutidas na literatura sobre a transmissão da COVID-19 e os fatores sociodemográficos associados. Consequentemente, é definida uma metodologia de análise de dados para o processamento das informações, e, por fim, são gerados modelos representativos das dimensões sociais mais suscetíveis a surtos epidêmicos. Uma visualização mais clara o fluxo de analise de dados para os modelos adotado o qual pode ser observada na Figura 8.

Figura 8 – Metodologia e estratégia de investigação adotadas no estudo.



Fonte: Elaborado pelo autor, (2024).

A Figura 8 ilustra o fluxo de execução da análise de dados adotado neste trabalho. A metodologia proposta permite a investigação de dimensões sociais relacionadas ao aumento de casos de COVID-19 no Brasil, por meio da aplicação de técnicas de ciência de dados. Essa abordagem é utilizada como processo central na busca por conhecimento em bases de dados, orientando as etapas de pré-processamento, análise e geração de modelos.

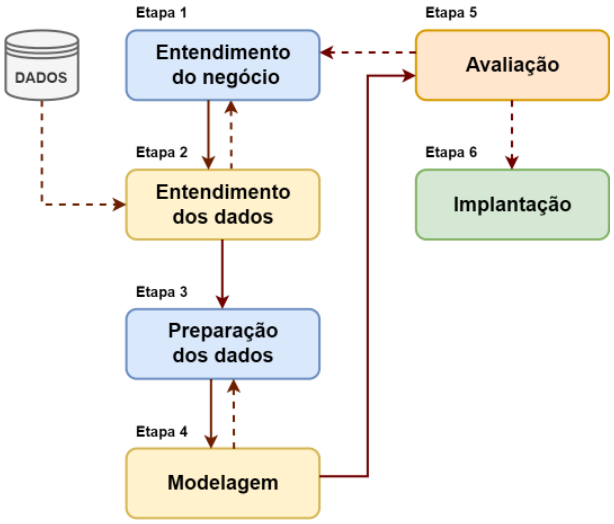
5.2.1 Entendimento do estudo

O método de pesquisa adotado neste trabalho utiliza técnicas de ciência de dados com base no processo CRISP-DM, por se tratar de uma abordagem que fornece uma visão abrangente do ciclo de vida da análise de dados (CHAPMAN et al., 2000). Atualmente, esse modelo é amplamente utilizado com esse propósito. Vale destacar que o CRISP-DM é composto por seis fases operacionais de mineração de dados: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e implantação (PIATETSKY, 2014).

As fases de aplicação estão representadas, de forma resumida, na Figura 9.

9 seguir:

Figura 9 – Metodologia aplicada no desenvolvimento dos estudo de caso.



Fonte: Elaborado pelo autor, (2024).

- **I. Entendimento do negócio:** Consiste em reconhecer as características essenciais da COVID-19, como os mecanismos de transmissão e propagação do vírus, formas de infecção e possibilidades de tratamento. Com base nesse entendimento, busca-se mensurar o impacto social da pandemia e desenvolver estratégias para mitigar novas formas de disseminação de surtos epidêmicos não sazonais.

- **II. Entendimento dos dados:** Envolve a identificação e exploração de bases de dados relacionadas a indicadores sociais e registros de infecção e óbitos por COVID-19. Para isso, foram utilizados o IVS e o IDH com o intuito de classificar aspectos sociodemográficos e econômicos da população. As fontes incluem bases públicas como o Censo Demográfico e a PNAD Contínua. Considerando a defasagem temporal dos dados do Censo em relação ao período da pandemia, os dados da PNAD Contínua foram incorporados com o objetivo de reduzir possíveis vieses.

- **III. Preparação dos dados:** Corresponde à seleção, ao processamento e à organização dos dados em formatos apropriados, de modo a viabilizar a aplicação eficiente das técnicas de mineração de dados (TAN; STEINBACH; KUMAR, 2009; GOLDSCHMIDT; PASSOS; BEZERRA, 2015). Nesta etapa, foram selecionadas variáveis pertencentes tanto às dimensões do IVS quanto do IDH, com o objetivo de compreender o comportamento da transmissão da COVID-19. Adicionalmente, as informações foram agrupadas em formatos compatíveis com o modelo de análise adotado, a fim de garantir maior coerência e compreensão na etapa de modelagem.

- **IV. Modelagem:** Refere-se à análise de dados sociais da população brasileira com o objetivo de identificar anomalias, padrões e correlações durante os períodos de maior

incidência de casos confirmados e óbitos. Essa etapa busca avaliar a interferência de tais variáveis nas taxas de mortalidade e letalidade.

- **V. Avaliação:** Tem como finalidade interpretar os resultados obtidos por meio de modelos computacionais, como mineração de dados, inteligência computacional e análises qualitativas, a fim de verificar se as saídas produzidas são válidas para o problema investigado (TAN; STEINBACH; KUMAR, 2009).
- **VI. Implantação:** Consiste na geração de relatórios e propostas que possam subsidiar planos de contingência para novos surtos epidêmicos — sazonais ou não — em diferentes localidades do Brasil.

Dessa forma, realiza-se uma análise qualitativa e quantitativa sobre dados sociodemográficos, com base em um fluxo de trabalho estruturado em ciência de dados, que integra métodos analíticos estatísticos e o CRISP-DM. O objetivo é rastrear possíveis fatores sociodemográficos relacionados ao aumento de casos de COVID-19 no Brasil, com ênfase em aspectos sociais e educacionais da população.

5.2.2 Seleção dos dados Sociodemográficos

A etapa de seleção abrange todo o processo de catalogação das bases de dados utilizadas, destacando-se, entre elas, PNADC, disponibilizada trimestralmente em formato de microdados com informações sociodemográficas da população, por meio do site do IBGE. Dessa forma, para cada ano são obtidas quatro bases distintas, as quais são triadas e posteriormente consolidadas em uma única base, a fim de facilitar as manipulações de dados nas etapas subsequentes.

Em seguida, foi realizado um processo de filtragem das variáveis consideradas essenciais para a análise. Foram descartadas aquelas que apresentavam menor relevância para o objetivo proposto, conforme descrito na Tabela 1. A base final resultante é composta por 16 variáveis, selecionadas por sua importância metodológica, uma vez que essas variáveis são as mesmas utilizadas na construção do IVS. Dessa forma, a seleção preserva a capacidade analítica necessária para o cálculo do índice e para a avaliação das condições de infraestrutura urbana, capital humano e renda e trabalho.

Tabela 1 – Variáveis selecionadas da PNADC e suas respectivas descrições.

Variáveis	**Descrição**
Trimestre	Coleta trimestral
UF	Unidade Federativa
Capital	Capital
UPA	Unidade Primária de Amostragem
V1008	Número de seleção do domicílio
V1014	Painel
V2001	Número de pessoas no domicílio
V1028	Peso do domicílio e das pessoas
V2007	Sexo
V2009	Idade do morador na data de referência
V3002	frequenta escola?
V3009A	Qual foi o curso mais elevado que frequentou anteriormente?
VD4020	Rendimento Mensal
VD2002	Condição no domicílio
V3001	Sabe ler e escrever?
VD4002	Situação de ocupação (Trabalho)
V4012	Tipo de vinculo empregatício

Fonte: Elaborado pelo autor, (2024).

Em seguida, foram identificados valores ausentes e nulos no banco de dados, os quais poderiam comprometer a qualidade da análise, gerando viés nos resultados ou dificultando a aplicação adequada das técnicas de mineração de dados. Diante desse cenário, realizou-se a etapa de limpeza dos dados, com a exclusão dos registros incompletos. Essa ação visou garantir maior precisão, consistência e robustez ao conjunto final, que passou a conter exclusivamente informações completas e compatíveis com os critérios definidos para a série temporal analisada.

PNADC desempenha um papel fundamental neste estudo, uma vez que o cálculo do IVS tem como premissa o uso de informações provenientes do Censo Demográfico, por este apresentar maior detalhamento sobre a população brasileira. No entanto, considerando que a catalogação do último Censo ainda está em andamento, a PNADC assume a função de fornecer dados mais atualizados, permitindo obter resultados mais condizentes com a realidade atual da população.

O Brasil.IO é responsável por disponibilizar dados sobre casos confirmados e óbitos por COVID-19. A base fornecida por essa plataforma apresenta informações detalhadas por município, permitindo análises em nível territorial. Para este estudo, foram selecionadas variáveis como o código do município, código da unidade federativa, data da ocorrência e taxa de mortalidade por 100 mil habitantes, conforme apresentado na Tabela 2.

Tabela 2 – Variáveis selecionadas da base de dados sobre COVID-19 do repositório Brasil.IO.

Variáveis	**Descrição**
date	Data do registro
state	Unidade Federativa (UF)
city	Município
place_type	Tipo de local (estado ou cidade)
confirmed	Número total de casos confirmados
deaths	Número total de óbitos
order_for_place	Ordem cronológica do registro para o local
estimated_population	População estimada do local
is_last	Indica se é o último registro disponível
city_ibge_code	Código IBGE do município
confirmed_per_100k_inhabitants	Casos confirmados por 100 mil habitantes
death_rate	Taxa de letalidade (óbitos / casos confirmados)

Fonte: Elaborado pelo autor, (2024).

5.2.3 Tratamento

A partir da etapa de seleção dos dados, o processo de preparação da base iniciou-se com a padronização das informações por meio da conversão dos valores absolutos em percentuais. Essa conversão foi aplicada a cada Unidade da Federação (UF) e suas respectivas capitais, abrangendo as variáveis selecionadas nas dimensões do IVS e do IDH. O uso de valores percentuais teve como objetivo facilitar a comparação entre diferentes territórios, independentemente do tamanho populacional, proporcionando uma análise proporcional mais precisa e equitativa.

A escolha por utilizar indicadores provenientes tanto do IVS quanto do IDH fundamenta-se na necessidade de uma análise integrada das condições socioeconômicas e estruturais que influenciam a propagação e os efeitos da pandemia. Enquanto o IVS fornece subsídios sobre carências sociais e infraestrutura precária, o IDH oferece uma visão consolidada sobre os níveis de desenvolvimento humano nos territórios, considerando educação, longevidade e renda.

Em consonância com a metodologia de agrupamento sociodemográfico proposta pelo Atlas do IVS e por estudos correlatos com o IDH, as UFs e capitais foram segmentadas com base nos dados de casos confirmados de COVID-19 e óbitos registrados nos anos de 2020 e 2021. Esse agrupamento permitiu identificar padrões de associação entre os níveis de vulnerabilidade/desenvolvimento e os desfechos epidemiológicos da pandemia.

A segmentação territorial teve como principal finalidade estabelecer relações entre as taxas de infecção e mortalidade pela COVID-19 e as particularidades sociais e regionais

de cada localidade analisada. Tal abordagem permitiu não apenas a visualização das disparidades existentes no território nacional, como também forneceu subsídios para a análise explicativa e preditiva dos impactos da pandemia, à luz de modelos de ciência de dados e redes bayesianas.

5.2.4 Processamento de dados

Foi realizado um levantamento bibliográfico sobre a temática da análise sociodemográfica durante a pandemia de COVID-19 nos anos de 2019 e 2022, com o objetivo de identificar o método mais adequado para a análise dos dados coletados. Nesse contexto, optou-se pela aplicação de métodos estatísticos descritivos, por estarem diretamente relacionados ao entendimento preliminar do comportamento dos dados no cenário social.

Entre os métodos descritivos, destacam-se aqueles voltados para a representação gráfica, tais como histogramas, gráficos de barras, gráficos de setores *pie charts*, *boxplots*, entre outros. Essas representações visuais têm como principal finalidade facilitar a exploração e a compreensão dos dados, contribuindo para a identificação e redução de possíveis inconsistências presentes nos conjuntos analisados (SILVA, 2018b).

5.2.5 Seleção de dados educacionais

Os dados utilizados neste estudo são mantidos pelo governo federal brasileiro e disponibilizados por meio dos sites do Ministério da Educação (MEC) e do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP).

Para a análise, foram selecionadas bases de dados do ENEM referentes aos anos de 2019 (pré-pandemia), 2020 e 2021 (durante a pandemia) e 2022 (pós-pandemia), além dos respectivos censos escolares desses períodos. Essas bases abrangem momentos críticos da crise sanitária, permitindo avaliar os impactos da pandemia sobre o desempenho dos estudantes do ensino médio. O objetivo é compreender de que forma as paralisações nas instituições de ensino, diante da emergência de surtos epidêmicos desconhecidos, influenciaram a formação educacional dos alunos.

5.2.5.1 Censo Escolar

A base de dados dos censos escolares de 2019 a 2022 contém 370 variáveis e possui mais de 228 megabytes compactados, abrangendo diferentes níveis da educação básica — desde a educação infantil até o ensino médio — e suas diversas modalidades (educação especial, regular e educação de jovens e adultos), tanto em escolas públicas quanto privadas.

A análise realizada neste estudo concentra-se nas escolas públicas e particulares que ofertam o último ano do ensino médio no estado do Pará. Vale destacar que a base analisada contém 37.920 registros. Após a eliminação de dados redundantes e de todas as

informações relacionadas ao atendimento de necessidades educacionais especiais, foram identificados 11 parâmetros considerados potencialmente mais influentes no desempenho dos estudantes, conforme apresentado na Tabela 3.

Tabela 3 – Parâmetros selecionados dos microdados dos Censos Escolares de 2019 a 2022 e suas respectivas descrições.

Parâmetros	**Descrição**
CO_ENTIDADE	Código da Escola
TP_DEPENDENCIA	Dependência Administrativa
TP_LOCALIZACAO	Localização da Escola
IN_AGUA_FILTRADA	Água consumida pelos alunos
IN_ESGOTO_INEXISTENTE	Esgoto sanitário - Não há esgotamento sanitário
IN_LIXO_SERVICO_COLETA	Destinação do lixo - Serviço de coleta
IN_PATIO_COBERTO	Na escola o patio e aberto
QT_SALAS_EXISTENTES	Quantidade de salas na escola
QT_DESKTOP_ALUNO	Quantidade de computadores em uso pelos alunos
QT_COMP_PORTATIL_ALUNO	Quantidade de computadores em uso pelos alunos
IN_INTERNET_ALUNOS	Acesso à Internet - Para uso dos alunos

Fonte: Elaborado pelo autor, (2023).

Além disso, outras informações são extraídas dos censos escolares e pré-processadas, sendo disponibilizadas por meio das sinopses estatísticas anuais, que consolidam dados referentes aos microdados educacionais. Essas sinopses abrangem informações sobre o rendimento dos alunos, bem como dados sociodemográficos, como sexo, idade, região, entre outros.

5.2.5.2 Microdados do ENEM

Os microdados do ENEM, referentes aos anos de 2019 a 2022, somam mais de 5 gigabytes e compreendem um conjunto de 76 variáveis. No total, a base reúne mais de 14 milhões de instâncias, representando o número de inscritos nos exames em todo o país.

Após uma análise inicial da base, constatou-se a necessidade de eliminar algumas variáveis. Foram descartadas aquelas consideradas redundantes, como o código da UF e o nome da cidade onde o candidato realizou a prova. Também foram excluídas variáveis irrelevantes para o escopo deste estudo, como o número de matrícula do participante, e variáveis relacionadas a casos específicos que não foram analisados, como a obrigatoriedade do uso de Braille e outras adaptações voltadas a candidatos com deficiência.

Com essas exclusões, o número total de variáveis foi reduzido para 44, conforme apresentado na Tabela 4.

Tabela 4 – Variáveis selecionadas dos microdados do ENEM, referentes aos anos de 2019 a 2022, e suas respectivas descrições.

Variáveis	**Descrição**
CO_MUNICIPIO_ESC	Código do município da escola
TP_SIT_FUNC_ESC	Situação de funcionamento (Escola)
TP_LOCALIZACAO_ESC	Localização (Escola)
TP_ESTADO_CIVIL	Estado Civil
TP_SEXO	Gênero
TP_COR_RACA	Cor/raça
TP_ST_CONCLUSAO	Situação de conclusão do Ensino Médio
TP_ENSINO	Tipo de instituição do Ensino Médio
TP_DEPENDENCIA_ADM_ESC	Dependência administrativa (Escola)
TP_PRESENCA_CN	Presença na prova objetiva de Ciências da Natureza
TP_PRESENCA_CH	Presença na prova objetiva de Ciências Humanas
TP_PRESENCA_LC	Presença na prova objetiva de Linguagens e Códigos
TP_PRESENCA_MT	Presença na prova objetiva de Matemática
NU_NOTA_CN	Nota da prova de Ciências da Natureza
NU_NOTA_CH	Nota da prova de Ciências Humanas
NU_NOTA_LC	Nota da prova de Linguagens e Códigos
NU_NOTA_MT	Nota da prova de Matemática
Q001	Grau de estudo do Pai
Q002	Grau de estudo do Mãe
Q003	Ocupação do pai
Q004	Ocupação do mãe
Q005	Quantidade de pessoas na residência
Q006	Renda mensal familiar
Q007	Se família possui empregado doméstico
Q008	Quantidade de banheiros na residência
Q009	Quantidade de cômodos dormitórios
Q010	Se família têm carros
Q011	Se família têm motocicletas
Q012	Na casa há geladeira
Q013	Na casa há freezer
Q014	Na casa tem máquina de lavar roupa
Q015	Na casa tem máquina de secar
Q016	Na casa tem forno micro-ondas
Q017	Na casa tem máquina de lavar louça
Q018	Na casa tem aspirador de pó
Q019	Na casa tem televisão em cores
Q020	Na casa tem aparelho de DVD
Q021	Na casa tem TV por assinatura
Q022	Na casa tem telefone celular
Q023	Na casa tem telefone fixo
Q024	Na casa tem computador
Q025	Na casa há acesso à *Internet*

Fonte: Elaborado pelo autor, (2024).

Consequentemente, foi realizado um recorte na base de dados para selecionar informações referentes ao estado do Pará, com o objetivo de mensurar os impactos ocasionados pela pandemia de COVID-19, especificamente entre os participantes das edições do ENEM de 2019 a 2022. A seleção dos dados foi conduzida por meio de uma análise estatística descritiva aplicada às variáveis sociodemográficas dos alunos.

5.2.6 Tratamento

Nessa etapa, foram eliminados os dados duplicados, inconsistentes ou incompletos (como valores nulos e vazios), os quais poderiam comprometer o desempenho dos algoritmos e inviabilizar a realização de análises mais realistas e representativas. Esse processo de pré-processamento foi aplicado de forma sistemática tanto aos dados do ENEM quanto aos dados do Censo Escolar, assegurando a qualidade e a integridade das informações utilizadas nas etapas subsequentes da pesquisa.

Além disso, foi realizada a discretização de variáveis contínuas (numéricas) e discretas que apresentavam múltiplas categorias, com o objetivo de reduzir o espaço amostral de possíveis valores e agrupá-los em intervalos. Essa abordagem facilita a aplicação de técnicas analíticas e a compreensão do comportamento dos estudantes em cada edição do exame analisada.

O estudo classifica o desempenho dos participantes com base na divisão das notas por quartis, considerando os valores mínimo e máximo em cada área de conhecimento (Equação 5.1), bem como o número de evasões em cada edição do ENEM, representada por ξ. Nessa equação, $\overrightarrow{u}$ representa um vetor de valores ordenados de forma crescente; Q indica o percentil analisado; e P, que varia de 1 a 3, determina os percentis de cada quartil K.

$$K_Q = \frac{P(\overrightarrow{u} + 1)}{4} \tag{5.1}$$

Ao contrário de outros estudos, que até então utilizam apenas a média das notas como critério de desempenho (CARMO; HECKLER; CARVALHO, 2020; ROCHA et al., 2022; NETO et al., 2022), a abordagem adotada neste trabalho proporciona um detalhamento mais preciso do rendimento dos alunos no ENEM, ao classificá-los em quatro grupos distintos, conforme apresentado na Tabela 6.

Tabela 6 – Distribuição dos participantes do ENEM por edição (2019–2022), segundo os quartis de rendimento.

Edição	ξ (ausentes)	$K_{Q_{\leq 25\%}}$	$25\% < K_Q < 75\%$	$K_{Q_{\geq 75\%}}$
2019	%	$-\infty$ – 443	444 – 546	557 – $+\infty$
2020	%	$-\infty$ – 439	440 – 544	545 – $+\infty$
2021	%	$-\infty$ – 443	444 – 572	573 – $+\infty$
2022	%	$-\infty$ – 484	484 – 602	603 – $+\infty$

Legenda: ξ: participantes ausentes; $-\infty$: notas abaixo do 1º quartil (25%); $+\infty$: notas acima do 3º quartil (75%).

Fonte: Elaborado pelo autor (2024).

A Tabela 6 apresenta a discretização das notas em três grupos, com base no método de quartis, definidos da seguinte forma: $K_{Q_{\leq 25\%}}$ para notas com rendimento inferior ou igual a 25%, $K_{Q_{26\%-74\%}}$ para notas entre 26% e 74%, $K_{Q_{\geq 75\%}}$ para notas superiores ou iguais a 75%, e ξ representa a quantidade de desistências. Essa categorização é fundamental para compreender como as dimensões sociodemográficas impactam o desempenho dos alunos diante de epidemias respiratórias iminentes ou de outros eventos sanitários que possam paralisar instituições de ensino.

Dentre as variáveis analisadas, 22 foram selecionadas por apresentarem maior correlação com o rendimento das notas. Essa seleção teve como objetivo reduzir o número de variáveis que influenciam o desempenho final, otimizando a construção da RB representativa do problema em questão, conforme apresentado na Tabela 7.

Tabela 7 – Variáveis com maior influência no desempenho dos alunos no ENEM, de 2019 a 2022.

Variáveis	**Tipo**
TP_DEPENDENCIA_ADM_ESC	Numérica
CO_MUNICIPIO_ESC	Numérica
TP_PRESENCA_CN	Numérica
TP_PRESENCA_CH	Numérica
TP_PRESENCA_LC	Numérica
TP_PRESENCA_MT	Numérica
TP_COR_RACA	Numérica
NU_NOTA_CN	Numérica
NU_NOTA_CH	Numérica
NU_NOTA_LC	Numérica
NU_NOTA_MT	Numérica
TP_SEXO	Numérica
Q001	Alfanumérica
Q002	Alfanumérica
Q003	Alfanumérica
Q004	Alfanumérica
Q005	Alfanumérica
Q006	Alfanumérica
Q008	Alfanumérica
Q009	Alfanumérica
Q024	Alfanumérica
Q025	Alfanumérica

Fonte: Elaborado pelo autor, (2023).

5.2.6.1 Transformação de dados

Para que os dados presentes na base possam ser utilizados como entrada para algoritmos de mineração de dados, é imprescindível submetê-los a etapas de transformação, com o objetivo de adequá-los aos requisitos técnicos desses algoritmos e otimizar seu desempenho. Assim, após o tratamento inicial, procede-se à etapa de transformação dos dados.

Essa fase compreende diversas operações, dentre as quais se destaca a discretização — processo responsável por reduzir a amplitude dos valores contínuos por meio de sua subdivisão em categorias ou intervalos. A discretização é especialmente útil em algoritmos que operam de forma mais eficiente com variáveis categóricas, contribuindo para a simplificação do espaço de atributos, a facilitação da análise e o aumento da eficácia das inferências geradas. Além disso, esse procedimento permite a identificação de padrões e tendências de forma mais intuitiva, favorecendo a extração de conhecimento relevante a partir dos dados transformados.

Conforme apresentado na Tabela 6, as variáveis por área de conhecimento, que compõem as notas dos participantes do ENEM, foram agrupadas em quatro faixas de desempenho. Posteriormente, o mesmo método utilizado para o cálculo da pontuação média geral de cada participante foi aplicado ao agrupamento por quartis e ao número de evasões.

No caso da variável **Renda Mensal Familiar** (Q006), cujos valores são expressos em faixas (por exemplo: "de R0, 00aR 998,00"), optou-se por convertê-la para faixas baseadas em múltiplos do salário mínimo, com o objetivo de reduzir a complexidade das categorias. Da mesma forma, as variáveis relacionadas à **quantidade de banheiros** (Q008) e de **dormitórios** (Q009) na residência, que apresentavam ampla variação, foram transformadas com base no número total de pessoas por domicílio, conforme as Tabelas 8 e 9.

Tabela 8 – Categorização da renda familiar (Q006) dos estudantes participantes do ENEM.

Classificação	**Faixas de Renda**	**Descrição**
Classe 1	[0,00 - 998,00]	Até 1 Salário
Classe 2	[998,01 - 1.497,00]	1,5 Salário
Classe 3	[1.497,01 - 1.996,00]	2 Salários
Classe 4	[1.996,01 - 2.495,00]	2,5 Salário
Classe 5	[2.495,01 - 2.994,00]	3 Salários
Classe 6	[2.994,01 - 3.992,00]	Mais de 3 Salários

Fonte: Elaborado pelo autor, (2024).

Tabela 9 – Categorização da quantidade de banheiros e dormitórios (Q008 e Q009) informada pelos estudantes participantes do ENEM.

Classificação	**Faixas utilizadas**
Classe 1	Nenhum
Classe 2	Sim, um
Classe 3	Sim, dois
Classe 4	Sim, três ou mais

Fonte: Elaborado pelo autor, (2024).

Com o intuito de facilitar a análise e otimizar o desempenho dos algoritmos de mineração de dados, as variáveis originalmente contínuas ou textuais foram transformadas em categorias bem definidas. Essa estratégia visa reduzir a complexidade dos dados e tornar sua estrutura mais adequada aos modelos analíticos adotados. As Tabelas 8 e 9 apresentam exemplos dessa abordagem, aplicando a categorização a variáveis socioeconômicas como a renda familiar (Q006) e a quantidade de banheiros e dormitórios (Q008 e Q009),

conforme informado pelos participantes do ENEM. A adoção de faixas padronizadas permite não apenas uma leitura mais clara e objetiva das informações, como também facilita a assimilação dos dados por parte dos algoritmos, promovendo maior precisão na identificação de padrões e correlações relevantes entre as variáveis.

A exemplo da variável de renda familiar, outras variáveis do questionário socioeconômico do ENEM também apresentam um número elevado de categorias, o que pode dificultar a visualização de padrões e comprometer a eficácia de modelos preditivos. Para lidar com esse desafio, foi realizada a reclassificação de variáveis como o número de moradores na residência (Q005), a ocupação do responsável (Q003) e a escolaridade dos pais (Q004), por meio de agrupamentos que tornam os dados mais sintéticos e analiticamente manejáveis. As Tabelas 10 e 11 ilustram esse processo, apresentando as novas categorias organizadas de forma hierárquica e funcional. Essa abordagem contribui para uma melhor assimilação das informações pelos algoritmos de mineração de dados, além de favorecer a identificação de correlações entre as variáveis explicativas e os desfechos de interesse da pesquisa.

Tabela 10 – Categorização da quantidade de moradores na residência (Q005) informada pelos estudantes participantes do ENEM.

Classificação	**Descrição**
Classe 1	1 morador
Classe 2	2 moradores
Classe 3	3 moradores
Classe 4	4 moradores
Classe 5	5 moradores
Classe 6	6 ou mais moradores

Fonte: Elaborado pelo autor, (2024).

Tabela 11 – Categorização da atividade laboral dos responsáveis (Q003 e Q004) informada pelos estudantes participantes do ENEM.

Classificação	**Descrição**
Grupo 1	Lavrador, agricultor sem empregados, bóia fria, apicultor. *etc.*
Grupo 2	Diarista, empregado doméstico, cuidador de idosos, cozinheiro. *etc.*
Grupo 3	Padeiro, cozinheiro industrial ou em restaurantes, costureiro. *etc.*
Grupo 4	Professor, técnico em geral, policial quadro praças até sargento. *etc.*
Grupo 5	Médico, engenheiro, dentista, psicólogo, economista, juiz. *etc.*

Fonte: Elaborado pelo autor, (2024).

Os dados disponíveis na base do Censo Escolar são binários, ou seja, suas respostas estão limitadas aos valores 0 (não) e 1 (sim). Esses dados foram correlacionados aos

parâmetros de desempenho nas competências avaliadas no ENEM, a saber: Redação, Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, e Matemática, representadas, respectivamente, pelas variáveis NU_NOTA_REDACAO, NU_NOTA_CN, NU_NOTA_CH, NU_NOTA_LC e NU_NOTA_MT, conforme classificação apresentada na Tabela 6.

A seleção dos parâmetros extraídos do Censo Escolar (Tabela 3) está diretamente relacionada às classificações das notas, com o objetivo de identificar os fatores com maior grau de suscetibilidade que possam comprometer o desempenho dos participantes do ENEM.

5.2.7 Mineração de Dados

A fase de mineração de dados foi conduzida por meio da aplicação da técnica de RBs, utilizando-se a biblioteca *PGMPY* . Essa ferramenta foi selecionada devido à sua flexibilidade, facilidade de configuração e capacidade de representar, de maneira intuitiva e eficiente, as relações probabilísticas entre variáveis. Além disso, a *PGMPY* permite a construção e o treinamento de modelos probabilísticos gráficos com suporte à geração automática das TPCs para cada nó da rede, de acordo com as dependências definidas e com os dados observados. Tais características a tornam especialmente adequada para o contexto deste estudo, no qual a compreensão das interações entre múltiplas variáveis socioeducacionais e seus impactos sobre os desfechos analisados demanda um modelo explicativo robusto e interpretável.

As inferências estatísticas e probabilísticas realizadas sobre os microdados do ENEM e do Censo Escolar tiveram como principal objetivo analisar os efeitos sociodemográficos de surtos epidêmicos sobre o desempenho discente. Procurou-se identificar os fatores que tornam determinados estudantes mais suscetíveis a impactos em contextos de crise sanitária e alertas de saúde pública.

Após a modelagem inicial, foi necessário realizar um processo de validação da estrutura da RB, a fim de garantir a robustez e a confiabilidade do modelo proposto, conforme descrito a seguir.

5.2.7.1 Validação de Estrutura e Limitações das RBs

Embora as RBs ofereçam uma vantagem significativa na captura de relações complexas entre variáveis, apresentam limitações inerentes, como a necessidade de assumir independência condicional ao empregar o algoritmo *Hill-Climb Search* (KOLLER, 2009). Para mitigar essas limitações, foi conduzida uma análise de validação da estrutura utilizando métricas de avaliação, incluindo o *K2Score*, *BicScore* e o *BDeuScore*, com o objetivo de assegurar a robustez e a confiabilidade dos resultados, conforme apresentado na Ta-

bela 12. Essas métricas fornecem medidas quantitativas da qualidade estrutural da rede, equilibrando o grau de ajuste do modelo com sua complexidade.

- **K2Score**: valores mais altos indicam melhor ajuste do modelo segundo a métrica K2, refletindo a adequação da estrutura da rede aos dados observados.
- **BicScore** e **BDeuScore**: valores negativos indicam a penalização pela complexidade do modelo, contribuindo para evitar o sobreajuste. Essas métricas desencorajam estruturas excessivamente complexas que não resultam em melhorias significativas no desempenho do modelo.

Tabela 12 – Comparação dos escores para diferentes estruturas de Redes Bayesianas.

Ano	**K2Score**	**BicScore**	**BdeuScore**
2019	1.32×10^7	4.99×10^5	-4.98×10^5
2020	7.34×10^6	-3.68×10^5	-3.66×10^5
2021	1.06×10^7	-3.14×10^5	-3.11×10^5
2022	1.54×10^7	-3.95×10^5	-3.94×10^5

Fonte: Elaborado pelo autor (2024).

A Tabela 12 apresenta a comparação dos scores obtidos para diferentes estruturas de RBs (RBs) aplicadas a edições sucessivas do ENEM, fornecendo uma base quantitativa sólida para avaliar a robustez dos modelos propostos. *O K2Score*, por exemplo, reflete diretamente o grau de aderência do modelo aos dados, com valores mais elevados indicando melhor ajuste. Por outro lado, os escores *BicScore* e *BDeuScore* incorporam critérios de penalização para evitar sobreajuste, buscando o equilíbrio entre a fidelidade ao conjunto de dados e a simplicidade da estrutura inferida.

Essas métricas desempenham papel crucial na validação estrutural das RBs, pois ajudam a garantir que as dependências probabilísticas sejam capturadas de maneira precisa, sem incorrer em complexidade excessiva. No entanto, a ausência de uma discussão interpretativa mais aprofundada sobre esses escores pode limitar a compreensão sobre sua real importância no processo de modelagem. Para que se possa avançar no uso dessas técnicas em contextos educacionais e sociais, é necessário que esses indicadores sejam analisados não apenas sob a ótica numérica, mas também em termos de significado e coerência com o domínio estudado.

Assim, recomenda-se que futuras pesquisas complementem essa avaliação quantitativa com análises qualitativas e validações empíricas, de modo a explorar com mais profundidade os impactos das escolhas estruturais. A integração entre métricas formais e julgamento contextual pode ampliar a aplicabilidade dos modelos em cenários reais, especialmente em sistemas educacionais complexos como o brasileiro. Essa abordagem

híbrida permitirá maior precisão na identificação de relações causais e maior confiabilidade na geração de inferências baseadas em dados.

5.3 Dificuldades encontradas

A utilização de bases de dados públicas, embora fundamental para o desenvolvimento de análises sociodemográficas e epidemiológicas, impõe uma série de desafios metodológicos. Em primeiro lugar, a necessidade de tratamentos complexos — como limpeza de inconsistências, normalização de variáveis e adequação de formatos — exige um esforço técnico considerável.

Além disso, é importante reconhecer que essas bases, embora amplas, nem sempre capturam de maneira precisa toda a diversidade e complexidade da realidade social brasileira. Fatores como subnotificações, defasagens temporais, falhas de cobertura amostral e heterogeneidade na coleta de dados entre diferentes regiões podem comprometer a representatividade dos indicadores. Assim, qualquer inferência ou modelagem baseada nesses dados deve considerar essas limitações de maneira crítica, para evitar interpretações simplistas e para reforçar a necessidade de análises complementares e triangulações metodológicas.

5.4 Considerações Finais

Neste capítulo, foi apresentado o ambiente computacional utilizado para a aplicação da análise de dados desenvolvida nesta dissertação, com o uso de ferramentas de código aberto, como *Miniconda* e *Python*, visando à melhoria da eficiência do processo analítico. Também foram destacadas as vantagens computacionais associadas ao uso dessas tecnologias. Por fim, foram descritas as linhas metodológicas adotadas nos dois estudos de caso.

A seguir, são apresentados e discutidos os resultados obtidos a partir da aplicação das metodologias descritas nos estudos de caso. Esses resultados têm como objetivo evidenciar, de forma empírica, os impactos de fatores sociodemográficos e educacionais em contextos de crise sanitária, com ênfase nas edições do ENEM realizadas durante a pandemia de COVID-19.

6 Resultados e Discussão

Neste capítulo, são apresentados dois estudos de caso que envolvem a aplicação de ciência de dados a bases de dados abertas e públicas, utilizando microdados sociodemográficos provenientes de instituições como o IBGE, INEP, DATASUS e Brasil.IO. A análise é conduzida por meio de uma abordagem de classificação paramétrica, focada no aumento de casos de COVID-19, além de investigar o impacto da pandemia sobre os estudantes participantes do ENEM, com ênfase no estado do Pará. Por fim, conforme descrito no capítulo anterior, são apresentados os resultados obtidos a partir da aplicação das metodologias adotadas em cada estudo de caso.

6.1 Estudo de Caso I

O Estudo de Caso I analisa o impacto da pandemia de COVID-19 no desempenho de alunos do ensino médio no estado do Pará, com base nos resultados do ENEM. Para essa análise, foram utilizados os microdados das edições do exame realizadas entre 2019 e 2022, além dos dados do Censo Escolar correspondentes ao mesmo período.

A avaliação do desempenho considerou variáveis como o tipo de unidade administrativa da escola, o período de paralisação das instituições em decorrência dos surtos epidêmicos de *SARS-CoV-2* e diversos fatores sociodemográficos. Entre estes, destacam-se: o número de quartos e banheiros na residência, o nível de escolaridade dos responsáveis, suas ocupações, a renda familiar e a disponibilidade de equipamentos tecnológicos para o acompanhamento das atividades remotas, entre outros.

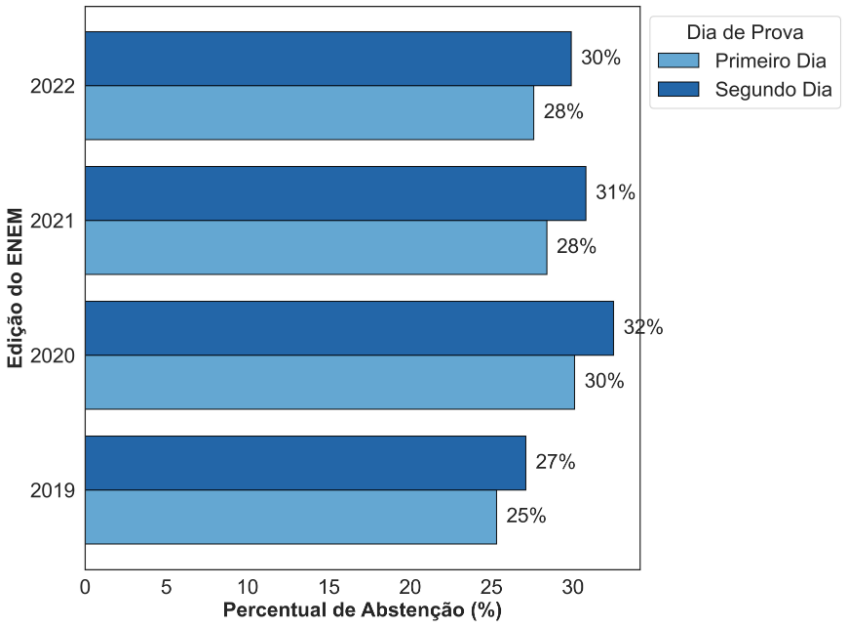
6.1.1 Resultados

Nesta seção, apresenta-se uma análise exploratória inicial dos dados, sem a aplicação do processo de categorização das variáveis previamente identificadas como mais relevantes, conforme descrito na Seção 5 O objetivo dessa análise é identificar quais variáveis sociodemográficas exerceram maior impacto sobre a participação dos alunos nas diferentes áreas do conhecimento, durante os surtos epidêmicos respiratórios não sazonais provocados pela COVID-19.

Na sequência, são apresentadas as RBs geradas com o uso da biblioteca *PGMPY*, por meio do algoritmo descrito na Seção 5 Destaca-se que, com base na topologia dessas redes, é possível formular interpretações contextualizadas aos objetivos do estudo. Para cada edição do ENEM analisada, foram construídas três RBs distintas, considerando exclusivamente os participantes do estado do Pará, localizado na região Norte do Brasil.

Inicialmente, foi realizado um levantamento geral sobre o número de desistentes do ENEM entre os anos de 2019 e 2022, com o objetivo de avaliar os impactos das paralisações nas instituições de educação básica sobre o acesso dos estudantes ao ensino superior, cuja seleção se baseia no desempenho obtido no exame. A Figura 10 apresenta o percentual de desistentes em relação ao número de inscritos no ENEM no período analisado.

Figura 10 – Percentual de participantes ausentes por edição do ENEM.



Fonte: Elaborado pelo autor, (2023).

A Figura 10 evidencia que o número de desistentes do ENEM em 2019 foi inferior ao registrado nos anos subsequentes. O próprio INEP tem observado essa tendência, com a taxa de evasão em queda desde 2016, passando de 32,2% em 2019 para 27,2%, o que representa, segundo o instituto, o menor índice já registrado até então (BRASIL, 2019). Uma possível explicação para esse fenômeno é a redução no número de matrículas no ensino médio, associada a um menor número de candidatos realizando o exame — fator atribuído a mudanças demográficas e ao aumento das taxas de aprovação (BRASIL, 2019).

Em 2020, o Exame Nacional do Ensino Médio (ENEM) foi fortemente impactado pela incerteza social provocada pela pandemia de COVID-19, elevando a taxa de abstenção a 53% do total de inscritos — um recorde histórico (OLIVEIRA, 2021). Entre os principais fatores que contribuíram para o abandono das provas, destacam-se as condições da infraestrutura das escolas onde o exame foi aplicado e a adequação às normas sanitárias definidas pelo Ministério da Saúde, como o distanciamento social e os riscos de contágio. Para mitigar esse impacto, o INEP abriu um segundo edital destinado a candidatos afetados por essas condições, recebendo cerca de 18 mil solicitações de reaplicação, das quais aproximadamente 13 mil foram aprovadas (BRASIL, 2021b).

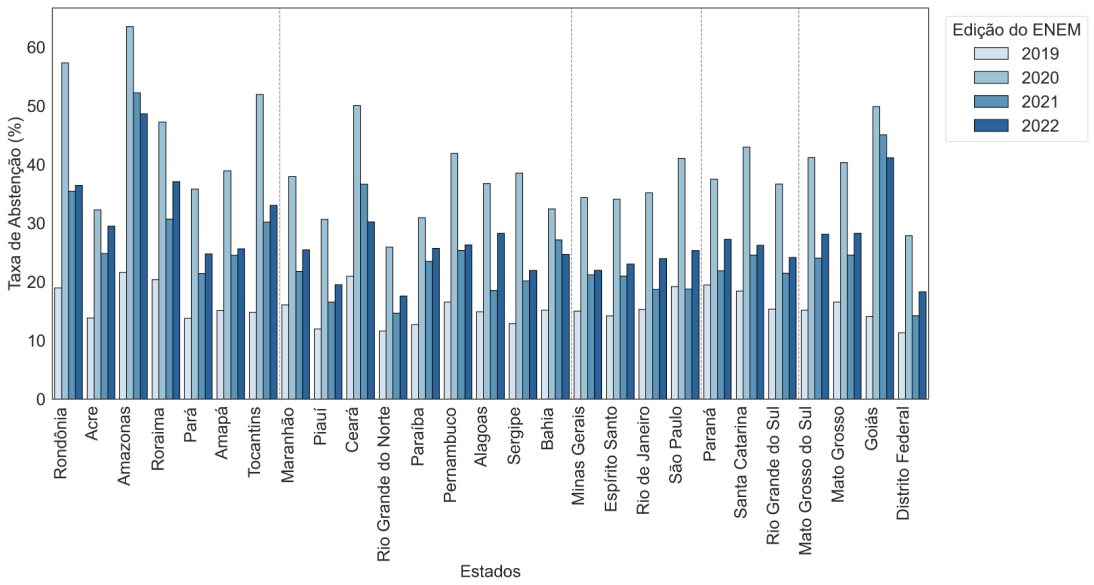
No ENEM 2021, observou-se uma redução na taxa de abstenção em comparação ao

ano anterior, embora o número de casos e óbitos por COVID-19 ainda fosse expressivo. Em 2020, por exemplo, 26% dos inscritos não compareceram a nenhuma das provas aplicadas aos domingos (BRASIL, 2021b). Esse declínio pode estar relacionado a uma maior clareza quanto às medidas de segurança individual e a um planejamento mais eficiente para a realização do exame no segundo ano da pandemia.

Em 2022, a taxa de abstenção manteve a tendência de queda observada em 2021, com 27,1% de ausência no primeiro dia de provas e 22,09% no segundo. Esses dados indicam uma recuperação gradual da participação dos candidatos, após os impactos significativos provocados pela pandemia. Em perspectiva comparativa, a taxa de abstenção foi de 23,2% em 2019 (pré-pandemia), subiu para 53% em 2020 (no auge da crise sanitária), caiu para 34% em 2021 e continuou em declínio em 2022. Essa redução sugere um cenário de maior estabilidade, com crescente confiança dos participantes nas condições de aplicação do exame, além de uma melhor adaptação das instituições e dos estudantes ao contexto pós-pandêmico.

Adicionalmente, foram analisados os dados de desistência por unidade federativa, com o objetivo de verificar se o aumento dos casos de abstenção ocorreu de maneira homogênea em todo o território nacional. A análise revelou que a Região Norte apresentou os maiores percentuais de abstenção, com destaque para os estados do Amazonas e de Rondônia, conforme ilustrado na Figura 11.

Figura 11 – Percentual de participantes ausentes por edição do ENEM por estados e o Distrito Federal.



Fonte: Elaborado pelo autor, (2023).

Os microdados referentes ao estado do Pará abrangem, em média, cerca de 180 mil participantes por edição do ENEM. Conforme apresentado nas Tabelas 13 e 14, a análise identificou 12 parâmetros relevantes, selecionados ao longo do processo de

tratamento e análise dos dados. Essas variáveis refletem alterações significativas nas condições socioeconômicas e comportamentais dos participantes durante a realização do exame, no contexto da pandemia de COVID-19 no Brasil, impactando diretamente o desempenho geral nas diferentes áreas do conhecimento avaliadas.

Tabela 13 – Percentual de participantes oriundos de escolas públicas que desistiram do ENEM.

Variável	2019		2020		2021		2022	
	Categoria	%	Categoria	%	Categoria	%	Categoria	%
Dependência ADM	Pública	15,45	Pública	41,49	Pública	26,78	Pública	33,52
Cor e Raça	Parda	69,16	Parda	66,84	Parda	64,25	Parda	67,71
Escolaridade Mãe	Ens. Fund.	39,85	Ens. Fund.	35,51	Ens. Fund.	30,60	Ens. Fund.	32,72
Escolaridade Pai	Ens. Fund.	45,44	Ens. Fund.	42,37	Ens. Fund.	38,53	Ens. Fund.	24,43
Ocupação Mãe	Grupo 2	46,07	Grupo 2	46,98	Grupo 2	47,11	Grupo 2	46,07
Ocupação Pai	Grupo 1	37,24	Grupo 1	32,31	Grupo 1	28,77	Grupo 2	31,21
Nº de Pessoas	+4	45,10	+4	41,31	+4	41,07	+4	49,45
Renda Familiar	1 Salário	63,73	1 Salário	70,00	1 Salário	62,41	1 Salário	67,36
Nº de banheiros	1	87,35	1	83,74	1	82,15	1	81,51
Nº de Dormitórios	2	51,39	2	51,89	1	82,15	2	51,02
Computador	Não	87,80	Não	84,10	Não	80,88	Não	83,90
Internet	Não	64,79	Sim	52,46	Sim	69,42	Sim	71,31

Legenda: **ADM:** Administrativa; **Ens. Fund.**: Ensino Fundamental; **Ens. Méd.**: Ensino Médio; **Grupo:** Ocupação (detalhes na Tabela 11).

Fonte: Elaborado pelo autor (2024).

Tabela 14 – Percentual de participantes oriundos de escolas privadas que desistiram do ENEM.

Variável	2019		2020		2021		2022	
	Categoria	%	Categoria	%	Categoria	%	Categoria	%
Dependência ADM	Privada	67,02	Privada	48,85	Privada	51.08	Privada	43.05
Cor e Raça	Parda	57,43	Parda	49,39	Branca	47,68	Parda	49,60
Escolaridade Mãe	Ens. Méd.	28,09	Ens. Méd.	37,58	Ens. Méd.	34,27	Ens. Méd.	37,50
Escolaridade Pai	Ens. Fund.	28,09	Ens. Méd.	29,24	Ens. Méd.	30,41	Ens. Méd.	31,74
Ocupação Mãe	Grupo 2	30,99	Grupo 4	35,97	Grupo 4	36,76	Grupo 4	36,11
Ocupação Pai	Grupo 4	23,14	Grupo 4	30,55	Grupo 4	29,89	Grupo 4	29,76
Nº de Pessoas	+4	35,00	+4	28,76	+4	31,20	+4	35,51
Renda Familiar	1 Salário	32,23	1 Salário	20,20	1 Salário	15,20	1 Salário	19,64
Nº de banheiros	1	55,78	1	39,89	1	36,85	1	43,84
Nº de Dormitórios	2	40,08	3	41,90	1	36,85	1	42,46
Computador	Não	53,30	Sim	39,49	Sim	34,53	Sim	57,00
Internet	Sim	69,83	Sim	88,24	Sim	95,87	Sim	92,06

Legenda: **ADM:** Administrativa; **Ens. Fund.**: Ensino Fundamental; **Ens. Méd.**: Ensino Médio; e **Grupo:** Ocupação (detalhes na Tabela 11).

Fonte: Elaborado pelo autor, (2023).

Conforme evidenciado nas Tabelas 13 e 14, os participantes cuja renda familiar não ultrapassa um salário-mínimo apresentaram as maiores taxas de evasão no ENEM

ao longo dos anos analisados — especialmente em 2020 e 2021, quando a desistência entre estudantes de escolas públicas atingiu 41,49% e 26,78%, respectivamente. Em 2022, essa taxa voltou a subir, alcançando 33,52%, o que sugere que os efeitos da pandemia continuaram a impactar negativamente a participação desses alunos no exame.

A análise revela que, em todos os anos considerados, a evasão entre estudantes da rede pública foi substancialmente maior do que entre os da rede privada. No entanto, em 2022, a taxa de desistência nas escolas privadas apresentou um aumento expressivo, alcançando 43,05%, enquanto na rede pública foi de 33% — o maior percentual do período. Esse crescimento pode estar relacionado a fatores socioeconômicos mais amplos, como a crise econômica e seus reflexos na continuidade dos estudos, mesmo entre estudantes de instituições particulares.

Outro fator determinante para a evasão no ENEM é o acesso a recursos tecnológicos. Em 2019, 87,80% dos alunos de escolas públicas que abandonaram o exame não possuíam computador — percentual que, embora tenha apresentado leve melhora em 2021 (80,88%) e 2022 (83,90%), permaneceu elevado. O acesso à internet também apresentou variações relevantes: entre os desistentes da rede pública, a proporção de estudantes sem acesso caiu de 64,79% em 2019 para 52,46% em 2020, mas voltou a crescer em 2022, alcançando 71,31%. Esses dados sugerem que, apesar de avanços pontuais no acesso digital, outros fatores estruturais continuaram a comprometer a permanência desses estudantes.

A escolaridade dos responsáveis também se mostrou relevante. Em 2022, 24,43% dos estudantes evadidos da rede pública tinham responsáveis com, no máximo, o ensino fundamental — um percentual inferior ao observado em 2019 (45,44%) e 2021 (38,53%). Essa redução pode indicar uma mudança no perfil socioeconômico dos estudantes que deixaram de participar do exame, sugerindo que, progressivamente, mesmo alunos com responsáveis mais escolarizados passaram a enfrentar dificuldades para se manter no processo avaliativo.

No que diz respeito à renda familiar, os dados de 2022 apontam que 67,36% dos participantes da rede pública que desistiram do ENEM pertenciam a famílias com renda de até um salário-mínimo, evidenciando a vulnerabilidade socioeconômica desse grupo. Na rede privada, essa proporção foi significativamente menor, alcançando 19,64%, o que reforça que, apesar do aumento da evasão também nesse setor, a desigualdade de renda continua impactando mais intensamente os estudantes da rede pública.

Dessa forma, os dados indicam que o tempo de paralisação das instituições públicas afetou de maneira desproporcional os estudantes de baixa renda, ampliando as desigualdades no acesso à educação e comprometendo diretamente a participação desses alunos no ENEM. Em consonância com Oliveira, Gomes e Barcellos (2020), esta tese defende que, embora 2022 tenha sido um período de transição, marcado pelo retorno gradual das atividades presenciais, os efeitos da pandemia ainda se refletiram nas taxas de desistência

— sobretudo entre aqueles com menor acesso a recursos educacionais e tecnológicos.

Seguindo a metodologia analítica adotada, a Tabela 15 apresenta informações separadas entre participantes de escolas públicas e privadas, com o objetivo de comparar o impacto da COVID-19 em diferentes contextos educacionais. Os dados corroboram análises empíricas de especialistas ao indicar que os estudantes da rede pública, com renda familiar de até um salário-mínimo e sem acesso à internet, foram os mais afetados pela transição do ensino presencial para o remoto.

Tabela 15 – Percentual geral de participantes que desistiram do ENEM.

Escola	Variável	K_ξ							
		2019	%	2020	%	2021	%	2022	%
Pública	Renda Familiar	[0-1] Salário	63.73	[0-1] Salário	70,00	[0-1] Salário	62,42	[0-1] Salário	67,36
	Computador	Não	87,80	Não	84,11	Não	80,89	Não	83,90
Privada	Renda Familiar	[0-1] Salário	32.23	1 Salário	20,20	[0-1] Salário	15,21	[0-1] Salário	19,64
	Computador	Não	53,31	Sim	37,19	Sim	32,27	Sim	57,00

Fonte: Elaborado pelo autor, (2024).

Os dados indicam que os participantes oriundos de escolas privadas apresentaram desempenho superior em comparação àqueles da rede pública, mantendo esse padrão desde os anos que antecederam a pandemia. Em 2019, 66,36% dos alunos com nota superior ao terceiro quartil (acima de 75%) no ENEM eram provenientes da rede pública; no entanto, esse percentual caiu para 43,05% em 2022, evidenciando uma redução significativa no desempenho desse grupo ao longo do período analisado, conforme apresentado na Tabela 16.

Fatores socioeconômicos também exerceram influência substancial sobre os resultados. Os alunos cujas mães possuíam ensino médio completo representaram a maior parcela entre os de desempenho elevado, reforçando a relação entre o nível educacional dos responsáveis e o rendimento acadêmico dos estudantes. Ademais, a proporção de mães inseridas no "Grupo 4"— que abrange ocupações profissionalmente mais qualificadas — aumentou de 28,11% em 2019 para 48,47% em 2022 entre os alunos de alto desempenho, indicando uma possível vantagem para estudantes provenientes de famílias com maior estabilidade econômica.

O acesso a recursos tecnológicos também se destacou como um fator determinante. Em 2019, apenas 47,91% dos alunos com notas acima de 75% no ENEM possuíam computador em casa; esse percentual cresceu para 80,64% em 2022. A disponibilidade de acesso à internet apresentou evolução semelhante, passando de 70,78% para 97,91% no mesmo intervalo. Esses dados reforçam a hipótese de que o ensino remoto contribuiu para o aprofundamento das desigualdades educacionais, favorecendo especialmente os estudantes que dispunham de estrutura tecnológica adequada para manter os estudos fora do ambiente escolar tradicional.

Tabela 16 – Comparação dos participantes com desempenho igual ou superior a 75% da nota total no ENEM.

Escola	Variável	2019		2020		2021		2022	
		Categoria	%	Categoria	%	Categoria	%	Categoria	%
Pública	Escolaridade Mãe	Ens. Méd.	39,22	Ens. Méd.	37,37	Ens. Méd.	34,84	Ens. Méd.	32,60
	Ocupação Mãe	Grupo 4	43,48	Grupo 4	46,40	Grupo 4	48,53	Grupo 4	48,47
	Computador	Sim	45,83	Sim	47,27	Sim	44,43	Sim	80,64
	Internet	Sim	88,19	Sim	94,27	Sim	98,05	Sim	97,91
Privada	Escolaridade Mãe	Ens. Méd.	38,76	Ens. Méd.	37,46	Ens. Méd.	34,49	Ens. Méd.	32,60
	Ocupação Mãe	Grupo 4	43,48	Grupo 4	46,40	Grupo 4	48,53	Grupo 4	48,47
	Computador	Sim	45,83	Sim	47,27	Sim	44,43	Sim	80,64
	Internet	Sim	88,19	Sim	94,27	Sim	98,05	Sim	97,91

Legenda: **Ens. Méd.**: Ensino Médio; **Grupo:** Ocupação (detalhes na Tabela 11).

Fonte: Elaborado pelo autor (2024).

Antes da pandemia, os alunos da rede pública — especialmente aqueles cujas mães possuíam apenas o ensino fundamental e exerciam ocupações com rendimento de até um salário-mínimo — já apresentavam desempenho acadêmico inferior em relação aos estudantes da rede privada. Com a chegada da pandemia, esse contraste tornou-se ainda mais acentuado. A escolaridade dos responsáveis mostrou-se um fator decisivo: entre os participantes oriundos de escolas privadas, mais de 40% tinham mães com ensino médio completo ou superior, enquanto, entre os alunos da rede pública, esse percentual era significativamente menor, conforme ilustrado na Tabela 17.

Com o avanço da pandemia, as desigualdades de oportunidades educacionais se intensificaram. A interrupção das atividades presenciais impactou de maneira mais severa os estudantes com menor poder aquisitivo, limitando o acesso a conteúdos escolares e comprometendo a preparação para o ENEM. Como evidenciado nos dados, os alunos da rede privada — que já dispunham de vantagens estruturais nos anos anteriores — foram menos afetados pelas dificuldades impostas pelo ensino remoto, preservando melhores condições para estudar e se preparar para a prova.

Adicionalmente, os dados reforçam a correlação entre escolaridade materna, acesso a recursos educacionais e desempenho acadêmico. Estudantes cujas mães exerciam profissões mais qualificadas e possuíam maior nível de instrução continuaram a apresentar melhores resultados. Por outro lado, aqueles cujas mães apresentavam baixa escolaridade e ocupações menos remuneradas enfrentaram maiores obstáculos para manter um bom desempenho escolar ao longo do período pandêmico.

Dessa forma, os resultados ressaltam a importância de políticas públicas voltadas à promoção da equidade educacional, assegurando que estudantes de diferentes contextos socioeconômicos tenham acesso a recursos, ferramentas e suporte adequados para uma aprendizagem efetiva — independentemente de adversidades externas, como as impostas pela pandemia de COVID-19.

Tabela 17 – Percentual de participantes que obtiveram nota superior a 75% no ENEM

Variável	2019		2020		2021		2022	
	Categoria	%	Categoria	%	Categoria	%	Categoria	%
Dependência ADM	Pública	66,36	Privada	52,05	Privada	53,70	Privada	43,05
Cor e Raça	Parda	59,55	Parda	54,76	Parda	50,69	Branca	47,48
Escolaridade Mãe	Ens. Méd.	37,26	Ens. Méd.	37,93	Ens. Méd.	35,89	Ens. Méd.	32,60
Escolaridade Pai	Ens. Méd.	43,39	Ens. Méd.	40,58	Ens. Méd.	38,03	Ens. Méd.	35,04
Ocupação Mãe	Grupo 4	28,11	Grupo 4	37,91	Grupo 4	38,58	Grupo 4	48,47
Ocupação Pai	Grupo 2	38,79	Grupo 4	41,22	Grupo 4	43,61	Grupo 4	42,16
Nº de Pessoas	4	35,58	4	39,40	4	40,20	4	39,85
Renda Familiar	1 Salário	13,48	3 Salários	10,70	1 Salário	10,12	3 Salários	13,11
Nº de Banheiros	1	61,56	1	48,60	1	44,24	1	37,15
Nº de Cômodos	2	48,85	2	44,84	2	43,67	2	43,53
Computador	Sim	47,91	Sim	65,83	Sim	64,42	Sim	80,64
Internet	Não	70,78	Sim	87,04	Sim	71,64	Sim	97,91

Legenda: **ADM:** Administrativa; **Ens. Fund.**: Ensino Fundamental; **Ens. Méd.**: Ensino Médio; e **Grupo:** Ocupação (detalhes na Tabela 11).

Fonte: Elaborado pelo autor (2024).

Os dados sugerem que o acesso a recursos tecnológicos teve um impacto significativo no desempenho acadêmico durante a pandemia. Estudantes com acesso domiciliar a computador e internet obtiveram pontuações mais altas no ENEM, o que evidencia a importância de garantir infraestrutura adequada para o ensino remoto — sobretudo em contextos de interrupção prolongada das atividades escolares presenciais.

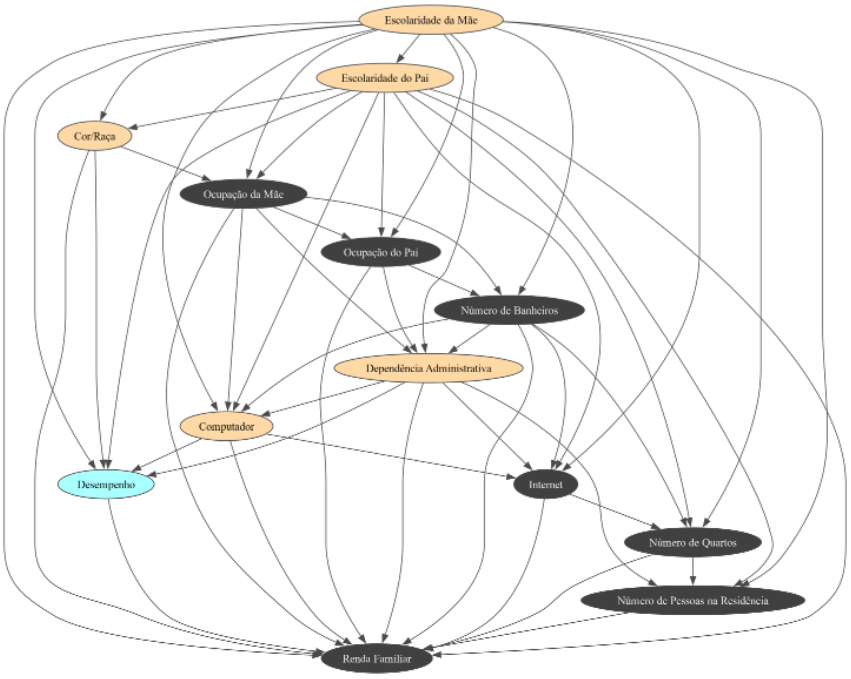
Além disso, observou-se que níveis mais elevados de escolaridade e emprego materno correlacionaram-se com um melhor desempenho estudantil. Esse achado reforça a hipótese de que o ambiente familiar exerce influência substancial sobre os resultados acadêmicos, para além do acesso direto a recursos materiais. O envolvimento parental e a formação educacional dos responsáveis parecem oferecer suporte adicional, seja por meio da organização de um ambiente de estudo estruturado, seja pelo incentivo ao valor da aprendizagem contínua (OLIVEIRA; GOMES; BARCELLOS, 2020; DIAS; RAMOS, 2022).

Após a análise exploratória descrita anteriormente, foi conduzida uma análise probabilística bayesiana com o objetivo de investigar de que forma o aumento dos casos de síndrome respiratória durante a pandemia de COVID-19 impactou o desempenho dos estudantes. Para isso, foram empregadas técnicas como *Hill-Climb Search*, *K2Score* e *Variable Elimination*, com o suporte da biblioteca pyAgrum, a fim de modelar e visualizar Redes Bayesianas referentes ao período de 2019 a 2022.

Inicialmente, foi gerada uma Rede Bayesiana Geral (do inglês, *Bayesian Belief Network* (BBN)) para modelar os fatores relacionados à abstenção dos estudantes paraenses no ENEM durante os anos críticos da pandemia de COVID-19. A escolha por essa abordagem se deu em virtude de sua capacidade de representar relações complexas de

dependência entre variáveis sociodemográficas, geográficas e institucionais, permitindo inferências mais realistas e simulações de cenários de intervenção. A modelagem revelou, por exemplo, que a combinação entre baixa renda, grau de escolaridade, cor ou raça e ausência de acesso à internet e a computadores teve impacto direto na elevação das taxas de abstenção, conforme ilustrado nas Figuras 12, 13 e 14.

Figura 12 – BBN com dados do ENEM de 2019, pré-pandemia de COVID-19



Nota: Nós em **laranja claro** representam variáveis com relação de causa direta com o nó de desempenho. O nó em **azul** indica o próprio desempenho, e os nós em **preto** possuem relação secundária ou nenhuma relação direta com o nó de desempenho.

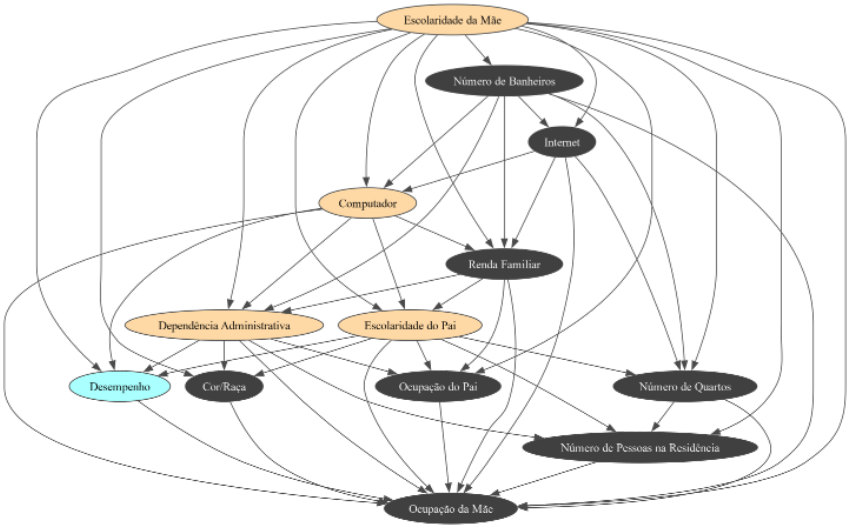
Fonte: Elaborado pelo autor (2024).

A Figura 12, referente ao período anterior à pandemia de COVID-19, apresenta cinco variáveis com relação de causalidade direta com o desempenho dos alunos: escolaridade da mãe, escolaridade do pai, cor/raça, dependência administrativa da escola e acesso a computador. Esses fatores evidenciam a forte influência de aspectos socioeconômicos e institucionais sobre os resultados educacionais em contextos de normalidade, ou seja, sem os impactos de uma crise sanitária. Além dessas, outras variáveis foram submetidas ao processo de refino da estrutura da rede, com o objetivo de identificar relações secundárias ou latentes que contribuíssem para uma compreensão mais abrangente dos determinantes do desempenho escolar.

Por outro lado, durante a pandemia de COVID-19, evidencia uma alteração nas relações de causalidade direta com o desempenho dos alunos. Quatro variáveis mantiveram vínculos diretos: escolaridade da mãe, escolaridade do pai, dependência administrativa da escola e acesso a computador Conforme a Figura 13. Notavelmente, a variável cor/raça,

anteriormente com ligação direta ao desempenho escolar, deixou de ocupar essa posição na rede. Tal alteração pode refletir o agravamento das desigualdades estruturais durante a pandemia, nas quais fatores relacionados ao capital escolar familiar e à infraestrutura tecnológica assumiram papel ainda mais central, especialmente diante da adoção emergencial do ensino remoto. A nova configuração da rede sugere que, em contextos de crise, o impacto de variáveis sociodemográficas pode ser mediado ou até mesmo suprimido por condições materiais mais imediatas, como o acesso a dispositivos tecnológicos e a qualidade da gestão escolar.

Figura 13 – BBN com dados do ENEM de 2021, durante a pandemia de COVID-19

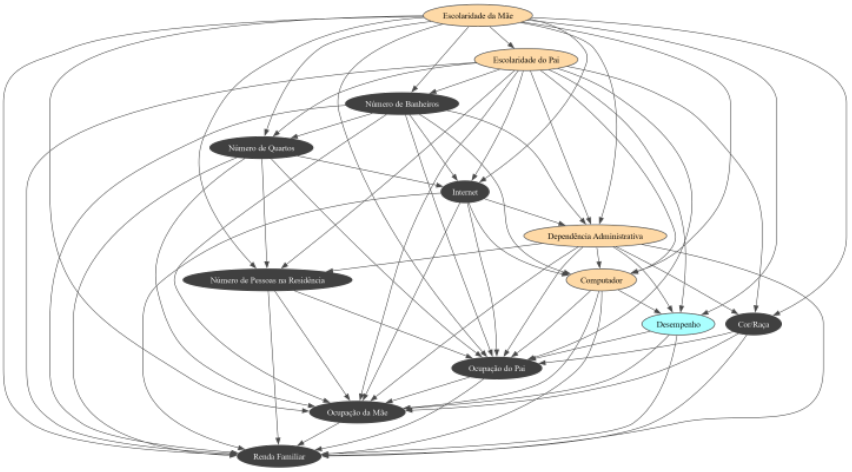


Nota: Nós em **laranja claro** representam variáveis com relação de causa direta com o nó de desempenho. O nó em **azul** indica o próprio desempenho, e os nós em **preto** possuem relação secundária ou nenhuma relação direta com o nó de desempenho.

Fonte: Elaborado pelo autor (2024).

A RB representada na Figura 14, que retrata o período pós-pandemia de COVID-19, apresenta a manutenção das quatro variáveis com relação de causalidade direta com o desempenho dos alunos: escolaridade da mãe, escolaridade do pai, dependência administrativa da escola e acesso a computador. Essa configuração sugere que os efeitos da pandemia reforçaram a centralidade dos fatores estruturais e socioeconômicos no desempenho educacional. A permanência dessas relações evidencia que os impactos da pandemia não apenas agravaram desigualdades já existentes, como também consolidaram o papel da infraestrutura tecnológica e do capital educacional familiar como determinantes do aproveitamento escolar. O cenário pós-pandêmico, portanto, revela uma resiliência limitada das redes educacionais em reverter os danos causados, especialmente entre estudantes em contextos de maior vulnerabilidade.

Figura 14 – BBN com dados do ENEM de 2022, pós-pandemia de COVID-19



Nota: Nós em **laranja claro** representam variáveis com relação de causa direta com o nó de desempenho. O nó em **azul** indica o próprio desempenho, e os nós em **preto** possuem relação secundária ou nenhuma relação direta com o nó de desempenho.

Fonte: Elaborado pelo autor (2024).

Em um segundo momento, foi implementado o modelo *Naive Bayes*, com o objetivo de comparar desempenho, simplicidade estrutural e capacidade preditiva. O *Naive Bayes* apresentou menor tempo de processamento e resultados satisfatórios nas tarefas básicas de classificação (como prever presença ou ausência no exame), mas teve desempenho inferior em cenários mais complexos. A suposição de independência condicional entre as variáveis preditoras comprometeu a capacidade do modelo de capturar interações relevantes, como aquelas entre tipo de escola, renda e região, reduzindo a sensibilidade contextual.

A Tabela 18 resume a comparação entre os dois modelos, considerando características estruturais, computacionais e aplicacionais. Os resultados reforçam que, embora o *Naive Bayes* seja eficiente e útil para tarefas com dados simples e independentes, a RB Geral se mostra mais apropriada para contextos em que múltiplas variáveis estão interligadas, como nas análises de desigualdade educacional agravadas pela pandemia.

Tabela 18 – Comparação entre RBs *Naive* e *Belief*

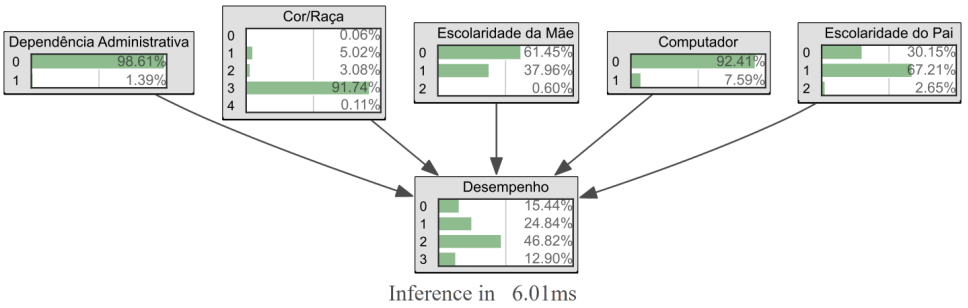
Característica	**Naive Bayes**	**BBN**
Independência entre variáveis preditoras	Sim (condicional à classe)	Não necessariamente
Estrutura da rede	Fixa, tipo estrela	Flexível, qualquer DAG
Complexidade computacional	Baixa	Maior (estrutura + inferência)
Aprendizado de estrutura	Não há (estrutura fixa)	Sim, pode ser automatizado
Aplicações	Classificação simples	Diagnóstico, previsão, simulação

Fonte: Elaborado pelo autor, com base em Koller (2009), Jensen e Nielsen (2007).

Essas evidências corroboram estudos prévios que apontam as limitações do *Naive Bayes* em contextos complexos (KOLLER, 2009; JENSEN; NIELSEN, 2007) e reforçam

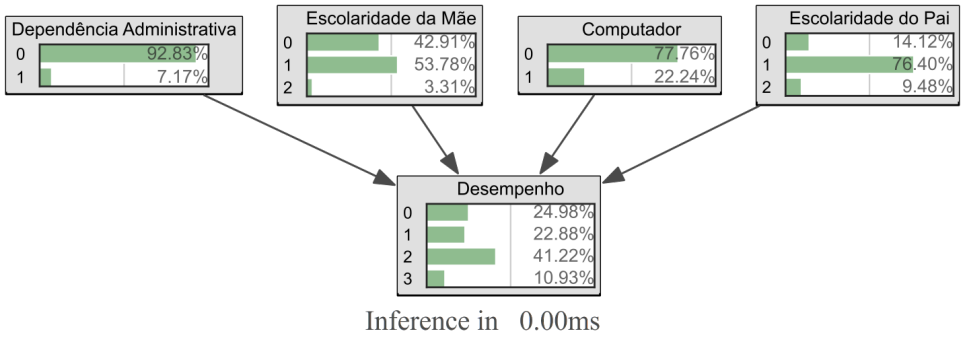
a vantagem das Redes Bayesianas quando se busca inferência causal e compreensão aprofundada dos dados (PEARL, 2009). Dessa forma, optou-se pela manutenção da RB *Belief* como modelo principal de análise da pesquisa, tanto pela robustez analítica quanto pela utilidade prática em cenários de alta interdependência entre variáveis. As estruturas geradas permitiram identificar relações probabilísticas entre variáveis sociodemográficas e desempenho acadêmico, conforme ilustrado nas Figuras 15, 16 e 17.

Figura 15 – RB *Naive* com dados do ENEM de 2019, pré-pandemia de COVID-19



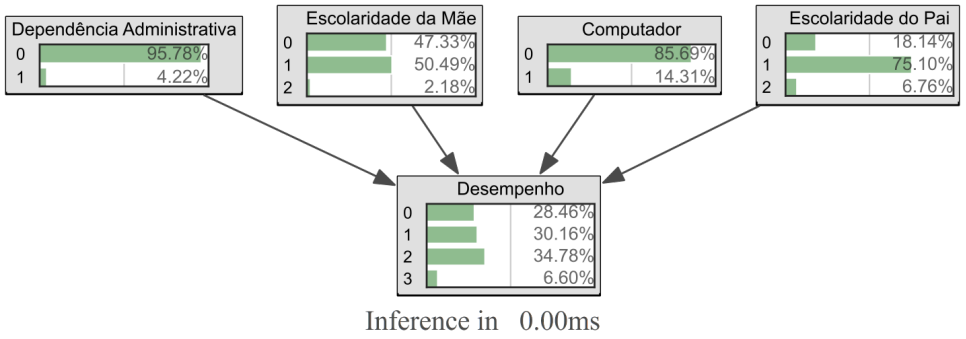
Fonte: Elaborado pelo autor (2024).

Figura 16 – RB *Naive* com dados do ENEM de 2021, durante a pandemia de COVID-19



Fonte: Elaborado pelo autor (2024).

Figura 17 – RB *Naive* com dados do ENEM de 2022, pós-pandemia de COVID-19



Fonte: Elaborado pelo autor (2024).

As RB derivadas dos dados de 2019 evidenciam um conjunto de variáveis-chave com forte influência sobre o desempenho dos participantes do ENEM. Dentre os principais fatores identificados, destacam-se o nível de escolaridade dos pais, a renda familiar, o acesso a computadores e a situação administrativa do domicílio, conforme representado na Figura 15. A estrutura da rede modela as relações de dependência probabilística entre essas variáveis, fornecendo uma base sólida para a análise do desempenho educacional. Esse arranjo permite não apenas a visualização das interações entre os fatores, mas também a inferência de padrões subjacentes que contribuem para o rendimento acadêmico dos estudantes.

A análise dos dados referentes a 2020 apresenta desafios singulares, uma vez que a pandemia alterou de maneira abrupta e imprevisível as relações tradicionalmente estabelecidas entre variáveis socioeconômicas e o desempenho acadêmico. No contexto do ensino remoto emergencial, marcado por acentuada desigualdade no acesso a recursos educacionais, os padrões observados tornam-se atípicos (QEDU, 2023; CASTRO; SOARES, 2021). Variáveis como renda familiar e escolaridade dos pais, que usualmente exibem correlações estáveis com o rendimento dos estudantes, revelaram-se mais voláteis e menos preditivas, refletindo os efeitos desproporcionais da crise sanitária sobre grupos socialmente vulneráveis (QEDU, 2023; CASTRO; SOARES, 2021; IEDE, 2021; (UFOP), 2021).

A aplicação da mesma metodologia para a construção de RB com os dados educacionais de 2021 e 2022 (Figuras 16 e 17) evidencia mudanças significativas no padrão de dependência entre as variáveis, fortemente influenciadas pelo contexto da pandemia de COVID-19. Um dos principais destaques é a redução da centralidade da variável presença de computador no domicílio como fator determinante de desempenho. Esse deslocamento pode ser atribuído à ampliação do uso de dispositivos móveis, especialmente celulares, como principal meio de acesso ao conteúdo educacional remoto. Tal fenômeno ganha relevância à medida que o ENEM 2020 foi realizado em um cenário de interrupções substanciais nas atividades escolares presenciais, com muitos estudantes enfrentando barreiras de conectividade e infraestrutura tecnológica adequadas para acompanhar o ensino a distância (ABUQUERQUE, 2020)

As RB demonstram elevada capacidade para modelar interdependências complexas entre variáveis educacionais e sociodemográficas, possibilitando inferências causais e a identificação de fatores latentes que influenciam o desempenho estudantil (ABUQUERQUE, 2020). Contudo, ao serem aplicadas aos dados do ENEM de 2020, essas redes encontram limitações relevantes. O impacto desproporcional da pandemia sobre estudantes em situação de vulnerabilidade socioeconômica provocou um recorde de abstenções e ampliou de forma significativa as disparidades de desempenho entre diferentes grupos, comprometendo a estabilidade dos padrões estatísticos usualmente observados (BRITO; PEDROSO, 2023).

O contexto pandêmico ressalta a importância de uma avaliação crítica dos modelos

probabilísticos, como as RB. Embora sejam ferramentas robustas, essas redes dependem de suposições de independência condicional, que podem ser comprometidas em cenários extremos, como o imposto pela pandemia. Assim, a interpretação dos resultados exige cautela, considerando as limitações do modelo e os possíveis vieses nos dados analisados (BRITO; PEDROSO, 2023; ROCHA; CASTRO; OLIVEIRA, 2023).

A análise das dependências condicionais identificadas pelas RB *Naive* revela que níveis mais elevados de escolaridade dos pais estão fortemente associados a um melhor desempenho dos participantes, conforme apresentado na Tabela 19 (BIENER; LANDMANN; SANTANA, 2019). No entanto, observa-se um aumento significativo nas taxas de evasão no ENEM (ξ) no período pós-pandemia, especialmente entre estudantes cujos responsáveis possuíam apenas o ensino fundamental — nesse grupo, a evasão cresceu 19% em relação ao período anterior à pandemia. Entre os alunos com pais com ensino superior, o aumento foi de 8%. Em 2020, auge da crise sanitária, as taxas de desistência chegaram a aproximadamente 31% para estudantes com menor escolaridade parental e a 10% para aqueles com maior escolaridade. Em 2021, esses índices recuaram para cerca de 14% e 9%, respectivamente, indicando uma modesta recuperação das condições educacionais.

Tabela 19 – Relação entre a Escolaridade do Pai e o Desempenho dos Participantes do ENEM segundo modelo de RB *Naive*.

ENEM	Grupo	Escolaridade do Pai		
		Ensino Fundamental	Ensino Médio	Ensino Superior
2019	ξ	20,68	13,58	10,88
	$K_{Q_{<=25\%}}$	30,55	23,05	19,25
	$K_{Q_{<75\%}}$	41,95	49,55	47,39
	$K_{Q_{>=75\%}}$	6,81	13,82	22,48
2020	ξ	51,36	38,01	29,60
	$K_{Q_{<=25\%}}$	21,83	19,01	13,24
	$K_{Q_{<75\%}}$	24,05	35,35	36,36
	$K_{Q_{>=75\%}}$	2,76	7,64	20,80
2021	ξ	34,73	25,30	17,75
	$K_{Q_{<=25\%}}$	29,34	24,08	19,20
	$K_{Q_{<75\%}}$	32,11	41,82	40,86
	$K_{Q_{>=75\%}}$	3,82	8,80	20,19
2022	ξ	39,76	27,90	18,92
	$K_{Q_{<=25\%}}$	35,51	32,06	25,72
	$K_{Q_{<75\%}}$	23,20	34,92	39,88
	$K_{Q_{>=75\%}}$	1,54	5,12	15,98

Fonte: Elaborado pelo autor, (2024).

Um aspecto importante a ser destacado é a probabilidade condicional entre a dependência administrativa e a disponibilidade de computador no domicílio para fins

educacionais. As inferências revelam uma correlação significativa, especialmente entre os estudantes de escolas públicas com acesso a computador, demonstrando uma forte associação com seus desempenhos no ENEM. Ao analisar as notas dos estudantes classificados no grupo $K_{Q_{<75\%}}$, observa-se uma disparidade acentuada entre aqueles que possuem e os que não possuem acesso a computador, com um aumento significativo no desempenho dos primeiros. Especificamente, foi registrado um aumento de 13% no desempenho dos estudantes de escolas privadas, conforme apresentado na Tabela 20.

Tabela 20 – Relação entre a Dependência Administrativa da Escola e a Presença de Computador no Domicílio, segundo RB *Naive* (ENEM 2019).

Grupo	Dependência Administrativa		
	Pública	Privada	Computador
ξ	16,09	10,79	Nenhum
$K_{Q_{<=25\%}}$	25,92	17,26	Nenhum
$K_{Q_{<75\%}}$	47,30	48,25	Nenhum
$K_{Q_{>=75\%}}$	10,68	23,70	Nenhum
ξ	7,21	3,51	Pelo menos um
$K_{Q_{<=25\%}}$	11,34	5,29	Pelo menos um
$K_{Q_{<75\%}}$	34,57	24,25	Pelo menos um
$K_{Q_{>=75\%}}$	46,87	66,95	Pelo menos um

Fonte: Elaborado pelo autor, (2024).

Outro aspecto crucial a ser considerado é a probabilidade condicional entre a dependência administrativa e a disponibilidade de computador no domicílio para atividades educacionais. As inferências indicam uma correlação significativa, especialmente entre estudantes de escolas públicas com acesso a computador, demonstrando melhorias notáveis nas notas do ENEM em comparação àqueles sem acesso. Entre os estudantes do grupo $K_{Q_{<75\%}}$, observa-se um aumento considerável nas pontuações daqueles que possuem acesso ao equipamento. Especificamente, os estudantes de escolas privadas apresentaram um aumento de 20%, conforme detalhado na Tabela 20.

Além disso, a análise da renda familiar informada pelos participantes revela uma forte relação entre níveis de renda mais elevados (C6) e as notas dos estudantes, como ilustrado na Tabela 20. Em consonância com essa inferência, ao examinar as faixas de renda familiar preestabelecidas, verifica-se uma queda no desempenho entre os estudantes que declararam renda de até um salário-mínimo (C1). Dentre aqueles classificados no grupo $K_{Q_{>=75\%}}$, houve uma redução expressiva de aproximadamente 6,5% no número de participantes pertencentes a essa faixa de renda.

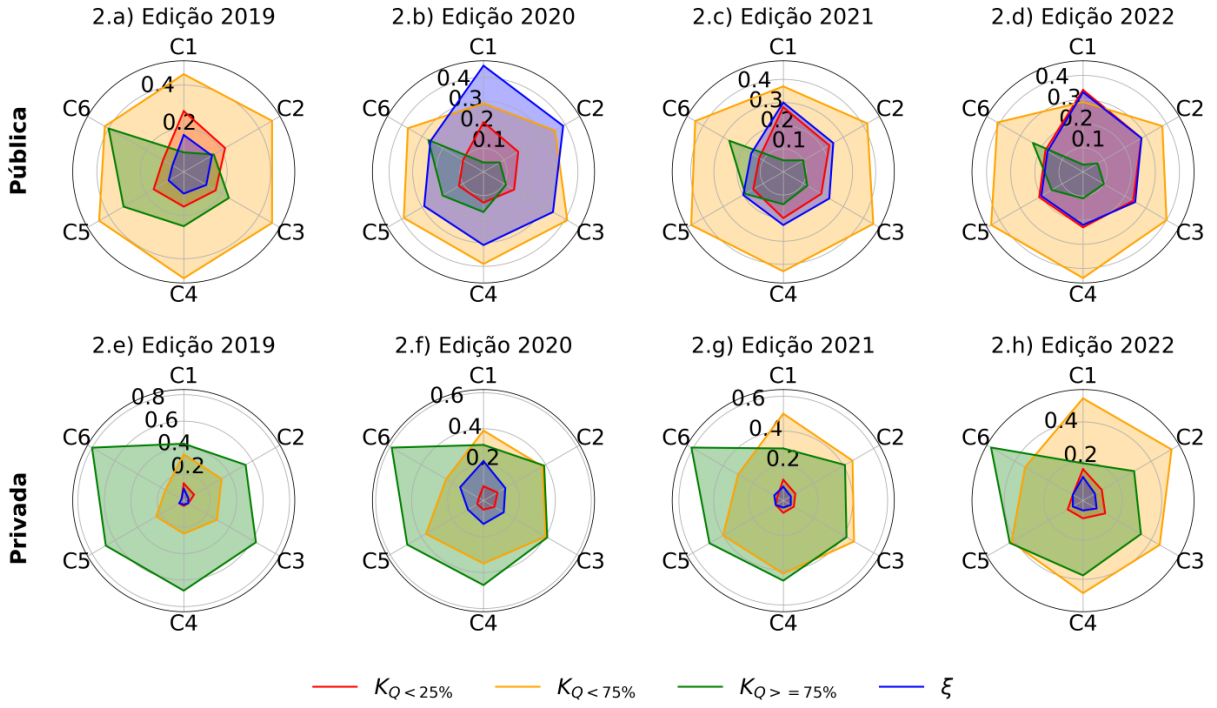
Tabela 21 – Percentual de Participantes com Nota Superior a 75% no ENEM, por Faixa de Renda Familiar, segundo RB *Naive*

Grupo	**2019**		**2020**		**2021**		**2022**	
	Classe 1	Classe 6	Classe 1	Classe 6	Classe 1	Classe 6	Classe 1	Classe 6
ξ	16,22	6,39	42,32	22,53	27,84	12,85	30,96	15,79
$K_{Q_{\leq 25\%}}$	22,51	10,41	20,35	8,07	26,49	9,50	33,96	14,71
$K_{Q_{<75\%}}$	47,01	41,06	31,94	41,59	39,59	39,85	31,61	39,59
$K_{Q_{\geq 75\%}}$	10,25	41,14	5,39	37,81	6,08	37,80	3,47	29,90

Legenda: **Classe 1**: Sem renda até 1 salário mínimo; **Classe 6**: Mais de 3 salários mínimos.
Fonte: Elaborado pelo autor (2024).

De forma mais detalhada, o impacto no desempenho dos participantes é analisado a partir da relação entre a dependência administrativa e a renda familiar. Observa-se que a proporção de alunos matriculados em escolas públicas é menor no grupo com desempenho elevado ($K_Q \geq 75\%$), especialmente entre aqueles cuja renda familiar é de até um salário-mínimo. Por outro lado, verifica-se um aumento de 30% no número de estudantes que abandonaram os estudos nessa mesma faixa de renda. A Figura 18 apresenta a distribuição do rendimento dos participantes de acordo com a renda familiar.

Figura 18 – Radar de Desempenho dos Estudantes no ENEM segundo a Renda Familiar e a Dependência Administrativa da Escola, com Base no Modelo RB *Naive*.



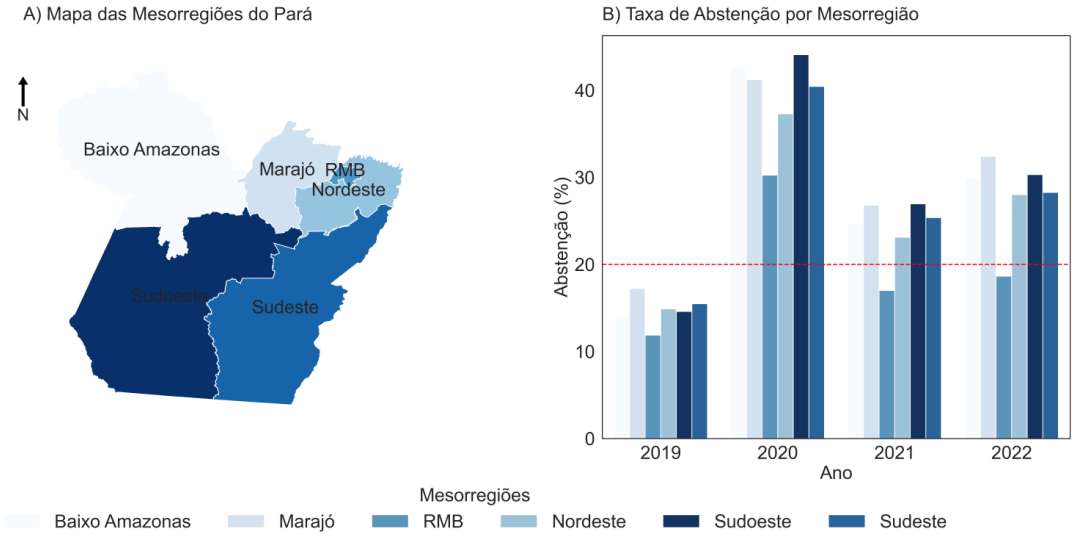
Legenda: **A**: Até 1 Salário; **B**: 1,5 salário; **C**: 2 salário; **D**: 2,5 salários; **E**: 3 salários e **F**: Mais de 3 salários; As cores representam percentual de desempenho: **Azul**: Desistentes; **Laranja**: Nota entre [0-25]; **Verde**: Nota entre [26-74]; **Vermelho**: Nota entre [75-100];

Fonte: Elaborado pelo autor, (2024).

Na Figura 18, observa-se um aumento no número de participantes oriundos de escolas privadas no grupo $K_{Q<75\%}$ entre os anos de 2020 e 2021. Essa mudança pode estar relacionada aos desafios impostos pelo ensino remoto durante os períodos de maior incidência de casos confirmados de COVID-19 no Brasil. Paralelamente, os estudantes da rede pública apresentaram os maiores percentuais de desistência na realização do ENEM no mesmo intervalo de tempo.

Outra análise realizada no Estudo de Caso I considerou os dados do estado do Pará, relacionando as seis mesorregiões paraenses com informações provenientes do Censo Escolar. O objetivo foi verificar se o impacto da pandemia de COVID-19 se manifestou de forma homogênea no número de abstenções e no desempenho geral dos participantes. Para facilitar a compreensão dos leitores, a Figura 19 apresenta a divisão territorial das mesorregiões paraenses, juntamente com os percentuais de abstenção registrados nas edições do ENEM de 2019 há 2022.

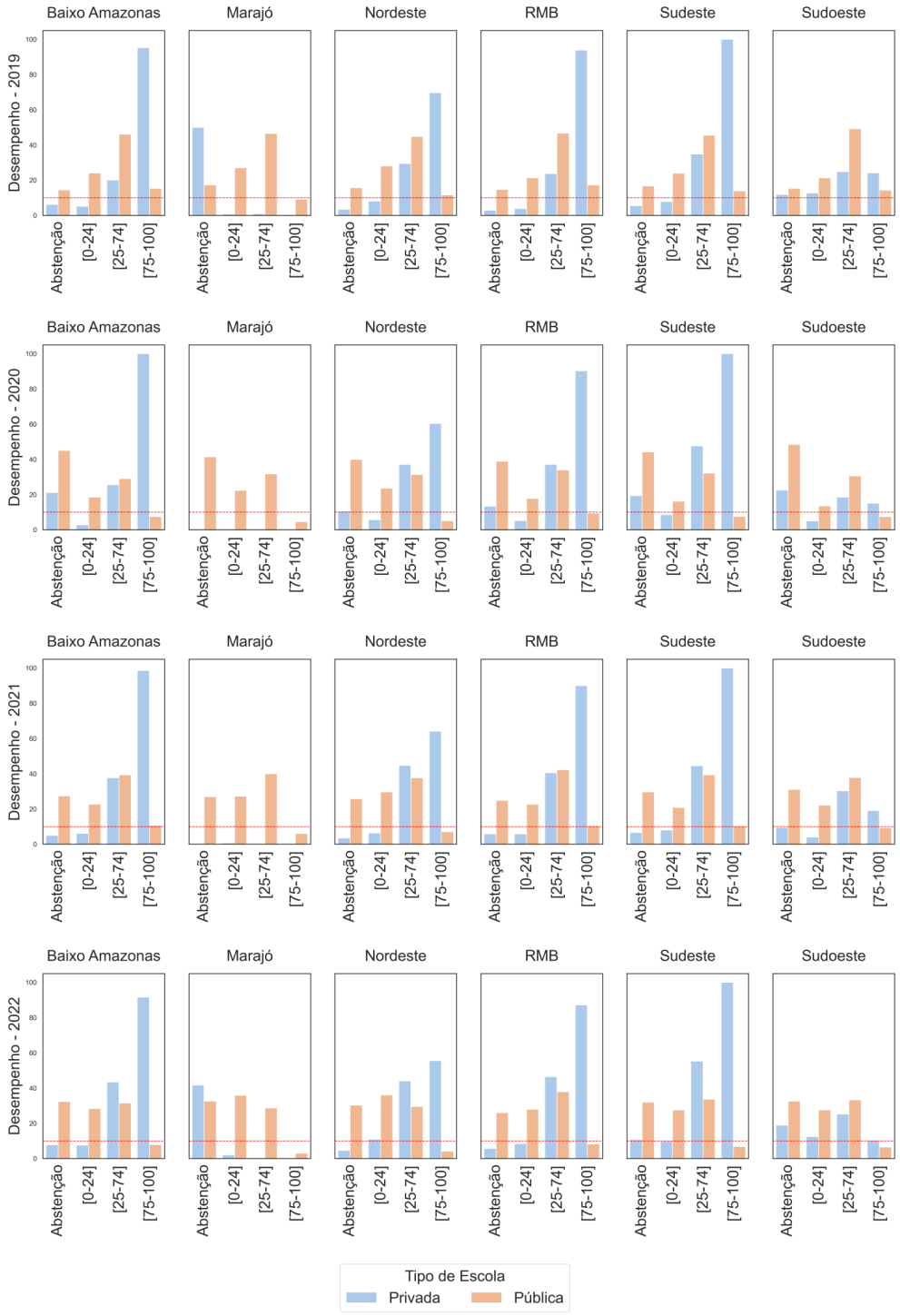
Figura 19 – Comparação do desempenho dos estudantes no estado do Pará por mesorregião no ENEM (2019–2022).



Fonte: Elaborado pelo autor, (2024).

A Figura 19 sugere que as diferenças regionais desempenharam um papel crucial no impacto da pandemia sobre a educação. Enquanto algumas regiões implementaram estratégias que ajudaram a mitigar as taxas de abstenção, outras enfrentaram desafios significativos, como elevados índices de desistência. Entre as mesorregiões do Pará apresentadas na Figura 4, destacam-se apenas a Região Metropolitana de Belém e o Nordeste Paraense, que conseguiram reduzir de forma significativa as taxas de abstenção no ENEM durante a pandemia de COVID-19, entre 2020 e 2021. Em contraste, as demais regiões mantiveram taxas persistentemente elevadas, com percentuais superiores a 20% no mesmo período.

Figura 20 – Abstenção dos alunos do estado do Pará por mesorregião no ENEM.



Fonte: Elaborado pelo autor, (2024).

A Figura 20 evidencia que as diferenças regionais exerceram um papel determinante no impacto da pandemia de COVID-19 sobre a educação básica no estado do Pará. Observa-se que, enquanto algumas regiões conseguiram adotar estratégias eficazes para mitigar os efeitos negativos do ensino remoto emergencial, outras enfrentaram sérias dificuldades estruturais e organizacionais, o que contribuiu para o aumento das taxas de evasão escolar.

A região do Marajó foi uma das mais impactadas após o início da pandemia de COVID-19. Entre 2020 e 2022, menos de 10% dos estudantes de escolas públicas da região alcançaram notas superiores a 75% nas avaliações. Além disso, observou-se um aumento expressivo na abstenção de estudantes de escolas privadas no ENEM, possivelmente em função das restrições de mobilidade e do fechamento das escolas na ilha, conforme ilustrado na Figura 20.

Os resultados indicam que a pandemia agravou as desigualdades socioeconômicas, afetando tanto o acesso às provas quanto o desempenho dos estudantes. A transição para o ensino remoto expôs fragilidades estruturais e reforçou a necessidade de políticas públicas que promovam o acesso igualitário à educação, independentemente das condições econômicas ou da localização geográfica dos alunos. O acesso a recursos tecnológicos e um ambiente familiar adequado foram fatores decisivos para o sucesso acadêmico durante esse período.

O estudo revela que a pandemia impactou fortemente a participação e o desempenho no ENEM, especialmente nas regiões mais vulneráveis do Pará. Alunos de baixa renda e com acesso limitado à tecnologia apresentaram as maiores taxas de evasão entre 2020 e 2021. A interrupção das aulas presenciais e a dificuldade de adaptação ao ensino remoto prejudicaram, sobretudo, os estudantes da rede pública.

O acesso a computadores e à internet foi essencial para o bom desempenho dos alunos. Estudantes de escolas privadas, com maior acesso a esses recursos, obtiveram melhores resultados. Além disso, o nível de escolaridade e a ocupação dos pais influenciaram positivamente o rendimento acadêmico dos alunos, destacando a importância de um ambiente familiar estruturado para o sucesso educacional.

A análise por RBs e a avaliação regional apontaram impactos desiguais da pandemia nas mesorregiões do Pará. A Região Metropolitana de Belém e o Nordeste Paraense conseguiram reduzir as taxas de evasão, enquanto a Ilha do Marajó enfrentou maiores dificuldades, com queda no desempenho e aumento nas abstenções.

6.2 Estudo de Caso II

Neste estudo, foi realizada uma análise com base em métodos estatísticos descritivos, com o objetivo de identificar quais variáveis sociais brasileiras — como infraestrutura

urbana, saneamento básico, educação, renda e trabalho — exercem maior influência sobre a qualidade de vida da população durante surtos de doenças epidêmicas não controladas. A pesquisa evidencia os principais desafios sociais enfrentados em contextos de crise sanitária, como a pandemia de COVID-19, oferecendo subsídios relevantes para que gestores públicos planejem ações mais eficazes de mitigação dos impactos em eventuais emergências futuras.

Os dados utilizados referem-se a indicadores de saúde, trabalho e renda da população brasileira, obtidos em repositórios oficiais amplamente reconhecidos, como o IBGE, por meio da PNADC, o Departamento de Informática do SUS (DATASUS) e a plataforma Brasil.IO, que disponibiliza registros atualizados sobre a COVID-19 no Brasil. Essas bases fornecem informações sociodemográficas essenciais para a realização de estudos sobre as desigualdades sociais persistentes no país, permitindo análises robustas e contextualizadas dos efeitos da pandemia em diferentes territórios.

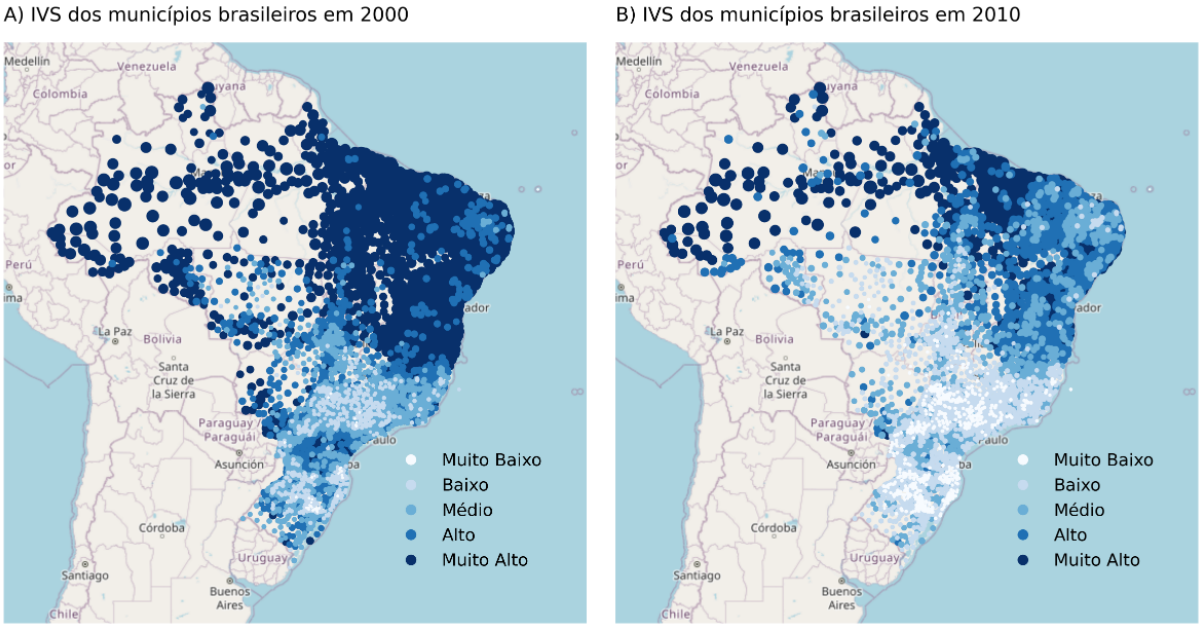
A análise das informações socioeconômicas foi conduzida com base no IVS, o qual mensura o grau de exposição de um território a riscos sociais e sanitários a partir de 16 dimensões associadas à qualidade de vida (MARí-DELL'OLMO et al., 2021). O uso do IVS possibilita a identificação de desigualdades em saúde e nas condições de vida, contribuindo para a compreensão de como fatores como o acesso a serviços essenciais — incluindo saneamento, saúde e educação — influenciam a propagação de doenças respiratórias, como a COVID-19. Além disso, a pesquisa investiga a relação entre as características socioeconômicas das habitações e a adoção de medidas protetivas, com o intuito de ampliar a resiliência das populações mais vulneráveis diante de emergências sanitárias.

6.2.1 Resultados

Inicialmente, foram utilizados dados dos Censos Demográficos de 2000 e 2010 de todos os municípios brasileiros para analisar a evolução das dimensões sociais da população (Figura 21). No entanto, durante boa parte do período da análise, os dados mais recentes eram ainda os de 2010, o que limitou temporariamente a realização de estudos municipais detalhados e atualizados sobre o tema abordado.

Com a posterior disponibilização dos dados do Censo de 2022, foi possível realizar uma atualização mais precisa das informações socioeconômicas da população brasileira. Apesar disso, para garantir a comparabilidade histórica, a análise do IVS foi inicialmente conduzida com base nas edições de 2000 e 2010, considerando as 16 dimensões sociais em diferentes regiões do país. A incorporação dos dados de 2022, embora relevante, exigiu adequações metodológicas para compatibilizar as séries censitárias.

Figura 21 – Distribuição do IVS nos municípios brasileiros nos anos de 2000 e 2010, conforme dados dos censos demográficos.



Nota: A distribuição territorial dos municípios brasileiros não segue uma padronização na delimitação das áreas. Isso afeta o sistema de georreferenciamento adotado neste trabalho, uma vez que cada ponto representa a localização da sede do município, e não a extensão de seu território.

Fonte: Elaborado pelo autor, (2024).

De modo geral, observou-se uma redução significativa da vulnerabilidade social nas regiões Centro-Oeste, Nordeste e parte da região central do Brasil. As regiões Sul e Sudeste apresentaram indicadores relativamente estáveis em ambos os períodos analisados. Por outro lado, a região Norte manteve os piores indicadores sociais, conforme ilustrado na Figura 21.

A distribuição territorial dos municípios brasileiros carece de padronização quanto à delimitação de suas áreas, o que impacta diretamente os sistemas de georreferenciamento utilizados neste trabalho. Como cada ponto geográfico analisado representa a sede do município — e não sua área total —, esse aspecto pode limitar a precisão espacial das análises desenvolvidas.

Diante disso, o segundo estudo de caso baseou-se no IVS, utilizando dados consolidados referentes aos anos de 2000, 2010 (Censos Demográficos) e 2015, disponibilizados pelo Atlas IVS, com atualização parcial a partir da PNAD Contínua. O objetivo foi avaliar a evolução das desigualdades sociais ao longo do tempo. A análise considerou como principal referência o Censo Demográfico de 2010, que, até a realização do Censo de 2022, representava o último levantamento populacional completo conduzido pelo IBGE. Complementarmente, foi realizado um estudo comparativo entre os valores do IVS de 2010

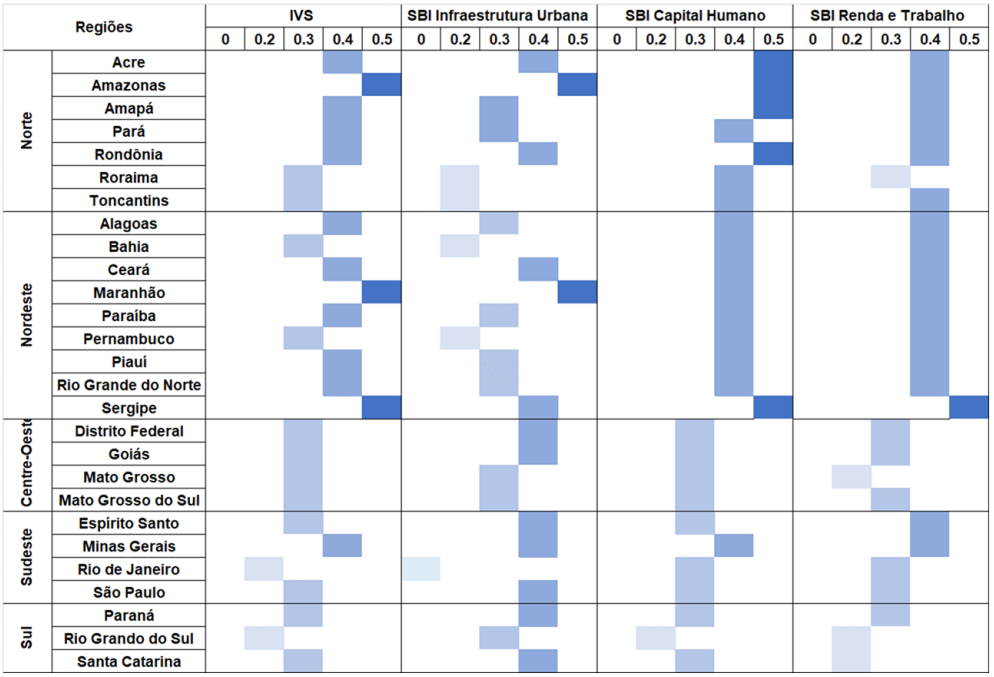
e os índices de mortalidade por COVID-19 nas capitais brasileiras e no Distrito Federal, com o intuito de identificar possíveis correlações entre níveis de vulnerabilidade social e os impactos da pandemia.

Ao considerar apenas as capitais brasileiras e o Distrito Federal (Figura 22 e 23), observa-se que as regiões Sul e Sudeste apresentam, de maneira geral, os menores índices de IVS, indicando uma menor vulnerabilidade social em comparação às demais regiões do país. Vale destacar que o índice varia de 1 a 5, sendo que valores mais altos representam níveis mais elevados de vulnerabilidade social.

Outro aspecto relevante é que a infraestrutura urbana apresentou a menor redução na vulnerabilidade ao longo do período analisado, entre as três dimensões que compõem o IVS. Esse fenômeno pode ser explicado pelo crescimento populacional das capitais, que não foi acompanhado por avanços proporcionais na infraestrutura.

Por fim, é fundamental destacar que os indicadores de infraestrutura urbana — como coleta de lixo, acesso adequado à água e esgoto e tempo de deslocamento até o trabalho — dependem, em grande parte, de políticas locais e regionais. Em contrapartida, fatores como distribuição de renda, geração de empregos e cotas estudantis são fortemente influenciados por políticas do Governo Federal.

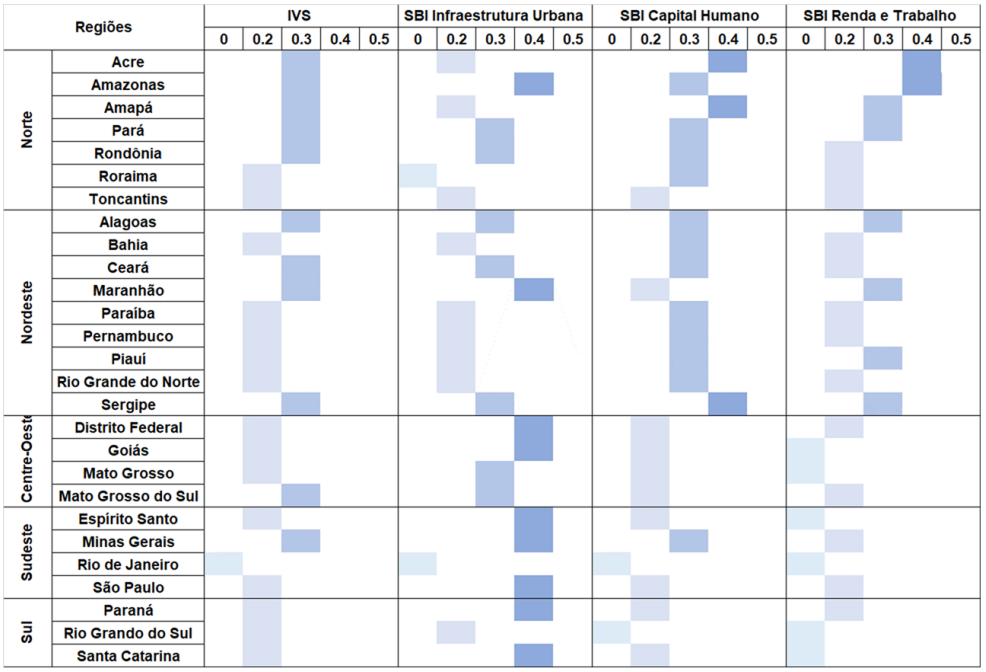
Figura 22 – Análise do IVS nas capitais brasileiras com base nos dados do Censo Demográfico de 2000.



Nota: Os números representam as faixas de vulnerabilidade social: **0**: Muito baixa; **0.2**: Baixa; **0.3**: Média; **0.4**: Alta; **0.5**: Muito alta.

Fonte: Elaborado pelo autor (2024).

Figura 23 – Análise do IVS nas capitais brasileiras com base nos dados do Censo Demográfico de 2010.



Nota: Os números representam as faixas de vulnerabilidade social: **0**: Muito baixa; **0.2**: Baixa; **0.3**: Média; **0.4**: Alta; **0.5**: Muito alta.

Fonte: Elaborado pelo autor (2024).

Deve-se considerar, ainda, que os dados utilizados apresentam uma defasagem de aproximadamente 10 anos, o que pode comprometer a precisão em relação à realidade atual. No entanto, até a data de realização deste estudo, o Censo Demográfico de 2010 ainda se configurava como a principal referência oficial para a análise das características socioeconômicas dos municípios brasileiros. Para mitigar essa limitação, foram incorporados dados mais recentes, dos anos de 2015 e 2019, com o objetivo de complementar a investigação, embora as análises não tenham como foco a dimensão geográfica dos Estados.

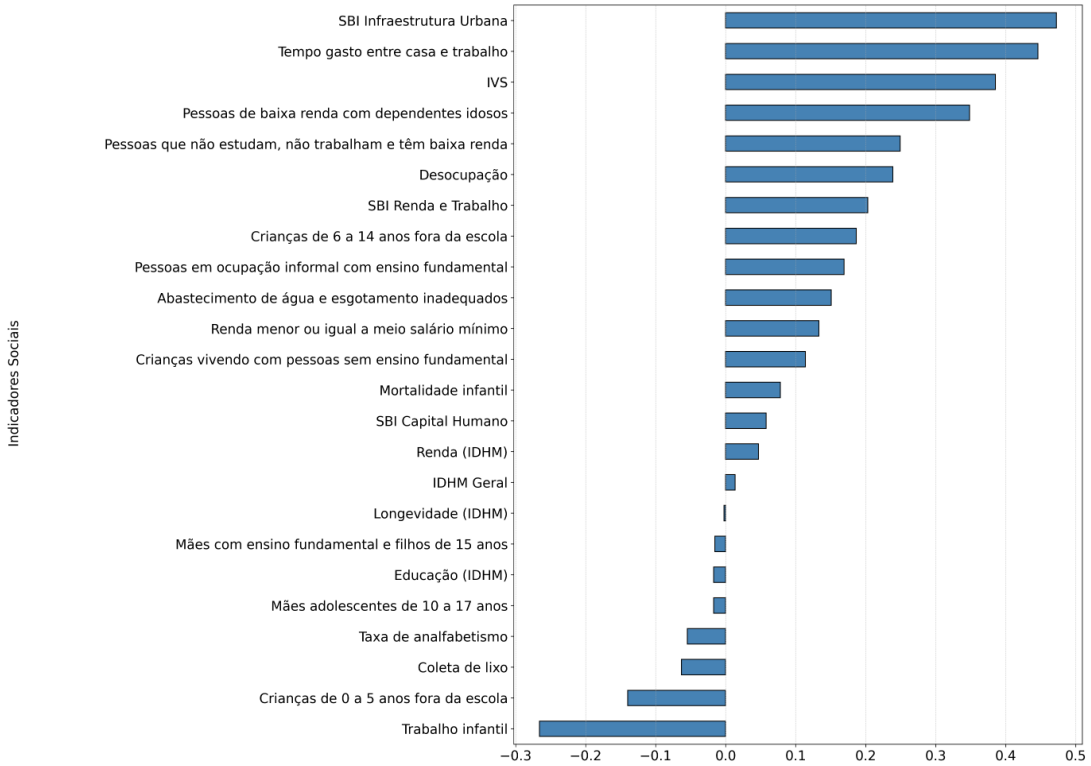
Para compreender o impacto dos indicadores que compõem o IVS no agravamento da pandemia de COVID-19 no Brasil, foram correlacionadas as dimensões do IVS com os dados de mortalidade por COVID-19 (por 100.000 habitantes). Adicionalmente, o IDH foi incorporado ao estudo para ampliar a análise, conforme ilustrado na Figura 24.

Observa-se que os fatores com correlação positiva mais significativa incluem: o percentual de pessoas em domicílios com renda per capita inferior a meio salário-mínimo e que dependem financeiramente de idosos (quando a renda do idoso representa menos de 50% da renda familiar total), o IVS de infraestrutura urbana, o IVS geral e o tempo despendido em afazeres domésticos.

É importante destacar que os valores de correlação não apontam um fator dominante

com alta correlação, o que indica que múltiplos elementos influenciam o avanço da pandemia. No entanto, ainda se observa uma relação estatisticamente significativa entre alguns indicadores, evidenciando como as condições sociais e econômicas impactam diretamente a disseminação da doença.

Figura 24 – Correlação entre óbitos por COVID-19 (por 100 mil hab), IVS, suas dimensões e IDH (Censo 2010)



Fonte: Elaborado pelo autor, (2024).

Outro aspecto relevante da análise de correlação é que os indicadores de vulnerabilidade social, representados pelo IVS, apresentaram uma correlação mais expressiva do que os indicadores de qualidade de vida, incluídos no IDH. Esse resultado sugere que o IVS é um instrumento mais adequado para representar as condições socioeconômicas e compreender os fatores que agravam a pandemia.

Em relação aos quatro indicadores com maior nível de correlação, pode-se observar:

- O percentual de pessoas em domicílios com renda per capita inferior a meio salário-mínimo e dependentes de idosos é um indicador relevante para avaliar o risco social durante a pandemia de COVID-19. Em lares onde o idoso é a principal fonte de renda, ele frequentemente assume tarefas como ir ao mercado e ao banco, aumentando sua exposição ao contágio. Além disso, idosos tendem a utilizar menos ferramentas alternativas, como compras online, o que reforça sua vulnerabilidade.

- O IVS de infraestrutura urbana considera três aspectos essenciais do território analisado: (i) redes de abastecimento de água e serviços de esgotamento sanitário, (ii) serviços de coleta de lixo e (iii) tempo de deslocamento entre casa e trabalho. Ao avaliar a correlação desse IVS com a taxa de mortalidade por 100.000 habitantes, observou-se que condições habitacionais inadequadas, a falta de saneamento básico e a ausência de coleta de lixo favorecem a disseminação de doenças infecciosas, muitas delas transmitidas por fezes e urina de animais em lixões. Além disso, enfermidades sazonais como Dengue, Zika e Chikungunya impactam essas populações, e o longo tempo de exposição em transportes públicos superlotados nas capitais contribui para a propagação em massa de doenças (SILVA, 2020).

- A correlação com o IVS geral reforça a inferência de que as condições de vulnerabilidade social influenciam diretamente na transmissão e no agravamento da COVID-19. Nesse caso, os 16 indicadores que compõem o IVS são considerados de maneira ponderada.

- A relação entre tempo de deslocamento casa-trabalho e o IVS de infraestrutura urbana demonstra que o transporte público desempenhou um papel central na exposição das pessoas a aglomerações. Quanto maior a proporção da população que depende do transporte coletivo e maior o tempo de deslocamento, maior foi o número de casos da doença. Embora diversos especialistas tenham apontado essa relação empiricamente, poucas medidas efetivas foram adotadas para minimizar os riscos no transporte público durante a pandemia.

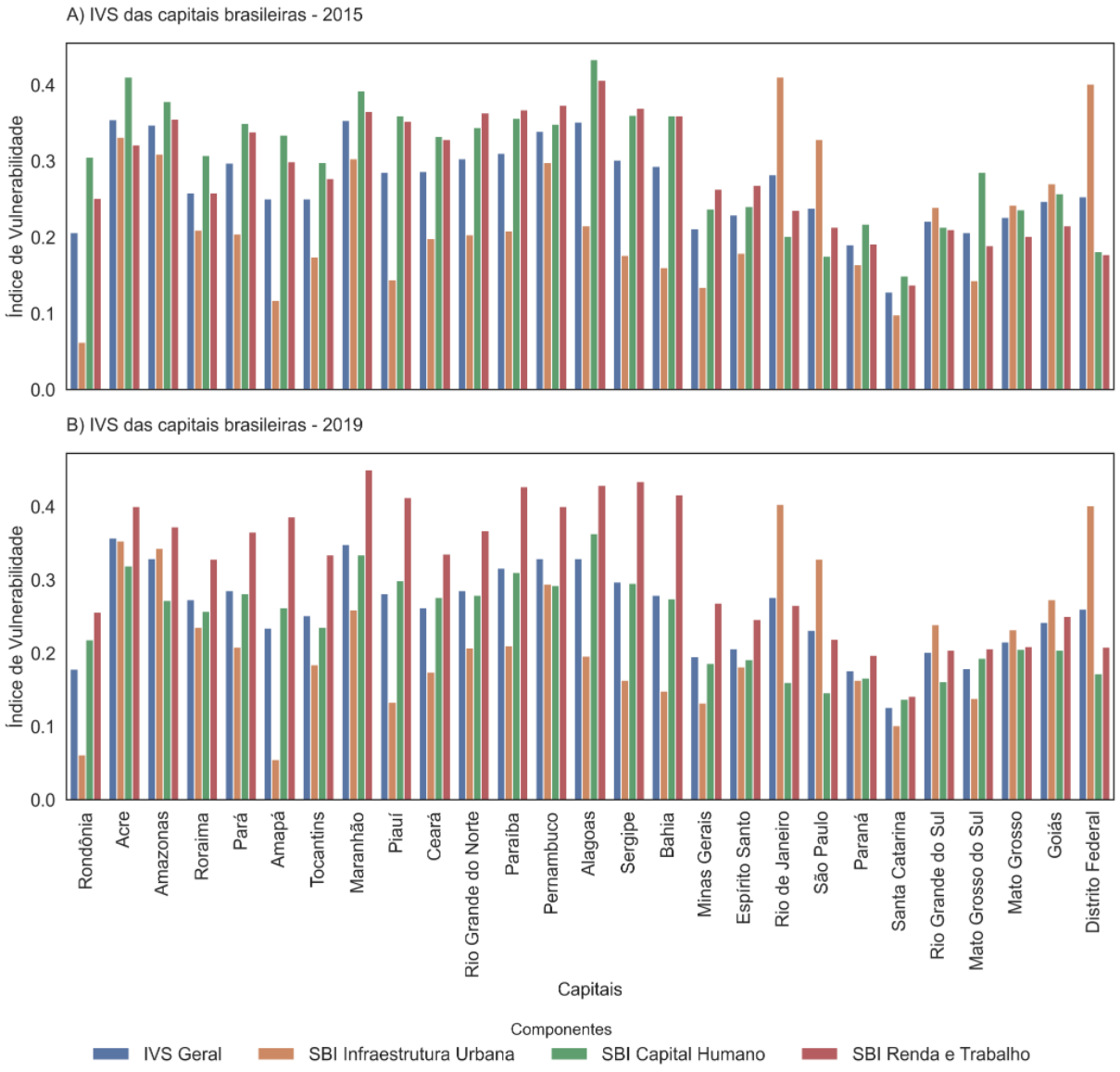
Para avaliar um cenário socioeconômico mais recente, anterior à pandemia de COVID-19, foram utilizados dados da PNAD de 2015 e 2019. Esses anos foram escolhidos porque 2015 foi o último ano com dados completos publicados para o cálculo do IVS, abrangendo todos os 16 indicadores das três dimensões analisadas, enquanto 2019 representa o período imediatamente anterior à pandemia.

É importante destacar que, a partir de 2016, nem todas as dimensões utilizadas no IVS continuaram sendo mensuradas e divulgadas pela PNADC. Exemplos disso são os dados sobre trabalho infantil, crianças de 0 a 5 anos fora da escola e tempo de deslocamento casa-trabalho, que deixaram de ser reportados. Além disso, devido à metodologia empregada pela PNADC, os dados possuem representatividade apenas para Estados e regiões geográficas amplas, impossibilitando uma análise detalhada em nível municipal.

A Figura 25 apresenta o IVS e suas dimensões para os Estados brasileiros e o Distrito Federal nos anos de 2015 e 2019. Nota-se que, em ambos os períodos, a vulnerabilidade social dos Estados mantém um padrão semelhante ao observado nos dados do Censo de 2010 (Figura22 e 23). Em média, os Estados das regiões Sul e Sudeste continuam

apresentando melhores indicadores de vulnerabilidade social em comparação às demais regiões do país.

Figura 25 – IVS e dimensões nos estados brasileiros e Distrito Federal.



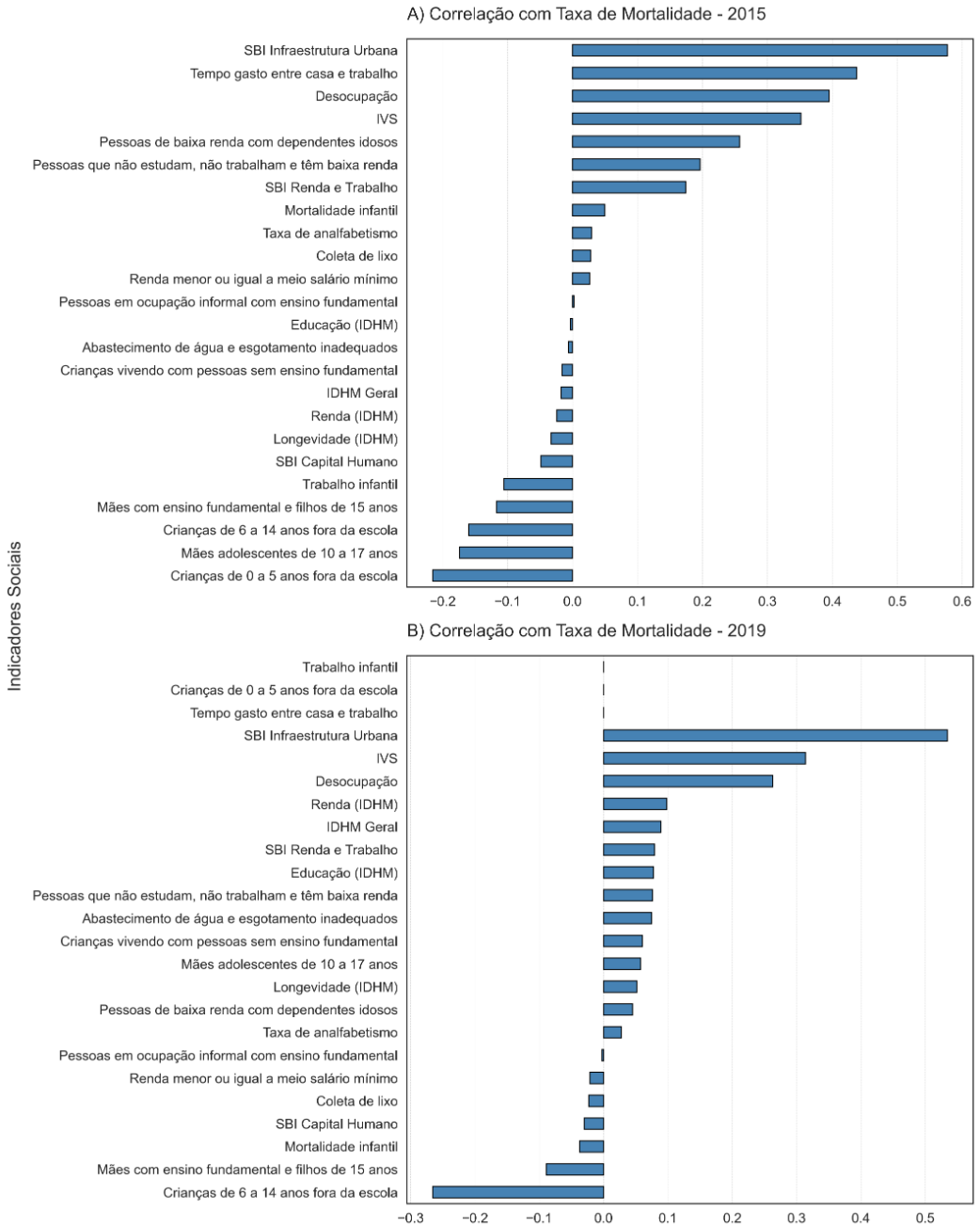
Fonte: Elaborado pelo autor, (2024).

Observa-se que, em 2015, a dimensão de Capital Humano apresentou os maiores índices de vulnerabilidade, enquanto, em 2019, a dimensão de Renda e Trabalho tornou-se a mais vulnerável. Esse comportamento está relacionado à ausência de determinados indicadores de Capital Humano em 2019, uma vez que algumas dimensões sociais não foram consideradas no cálculo da vulnerabilidade devido à defasagem das informações do último Censo.

As correlações entre as taxas de óbitos por 100 mil habitantes em decorrência da COVID-19 e os indicadores de vulnerabilidade social e qualidade de vida foram analisadas separadamente para os anos de 2015 e 2019, com base nos dados disponíveis pela PNADC

e Atlas IVS. O objetivo foi investigar possíveis padrões associativos entre o grau de vulnerabilidade das capitais brasileiras e a gravidade dos impactos da pandemia. Para isso, foram utilizados coeficientes de correlação de *Pearson*, que permitem avaliar a intensidade e a direção da relação linear entre as variáveis. A Figura 26 apresenta os gráficos de dispersão que ilustram essas correlações, permitindo visualizar a distribuição dos dados e identificar tendências regionais ou discrepantes ao longo dos dois períodos analisados.

Figura 26 – Correlação entre óbitos por COVID-19 (por 100 mil hab.)



Fonte: Elaborado pelo autor, (2024).

Ao analisar os dados referentes ao ano de 2015, observa-se uma continuidade nos padrões de correlação identificados a partir do Censo Demográfico de 2010. Em ambas

as análises, procurou-se compreender a relação entre indicadores socioeconômicos e as taxas de óbitos por 100 mil habitantes devido à COVID-19. Essa constância evidencia que fatores estruturais relacionados à vulnerabilidade social já se encontravam presentes no período pré-pandêmico, influenciando diretamente a forma como a população brasileira foi afetada pela crise sanitária.

Entretanto, os coeficientes de correlação observados para 2015 foram mais elevados do que os verificados em 2010, mesmo com a limitação de os dados estarem agregados por unidade federativa. Esse aumento sugere que as variáveis analisadas apresentaram maior significância estatística na explicação dos impactos da pandemia. Dessa forma, os dados de 2015 revelam-se mais informativos para compreender os mecanismos de propagação e os efeitos desiguais da COVID-19 entre os territórios brasileiros.

Entre os diversos indicadores investigados, destaca-se o componente de Infraestrutura Urbana do IVS, que apresentou um coeficiente de correlação de 0,57% em 2015. Esse valor expressivo reforça a hipótese de que a precariedade dos serviços urbanos essenciais — como abastecimento de água, esgotamento sanitário e transporte — influenciou diretamente os níveis de exposição e letalidade da doença. A defasagem estrutural observada em muitos municípios, sobretudo em áreas periféricas, pode ter dificultado a implementação de medidas preventivas e contribuído para a amplificação dos efeitos da pandemia.

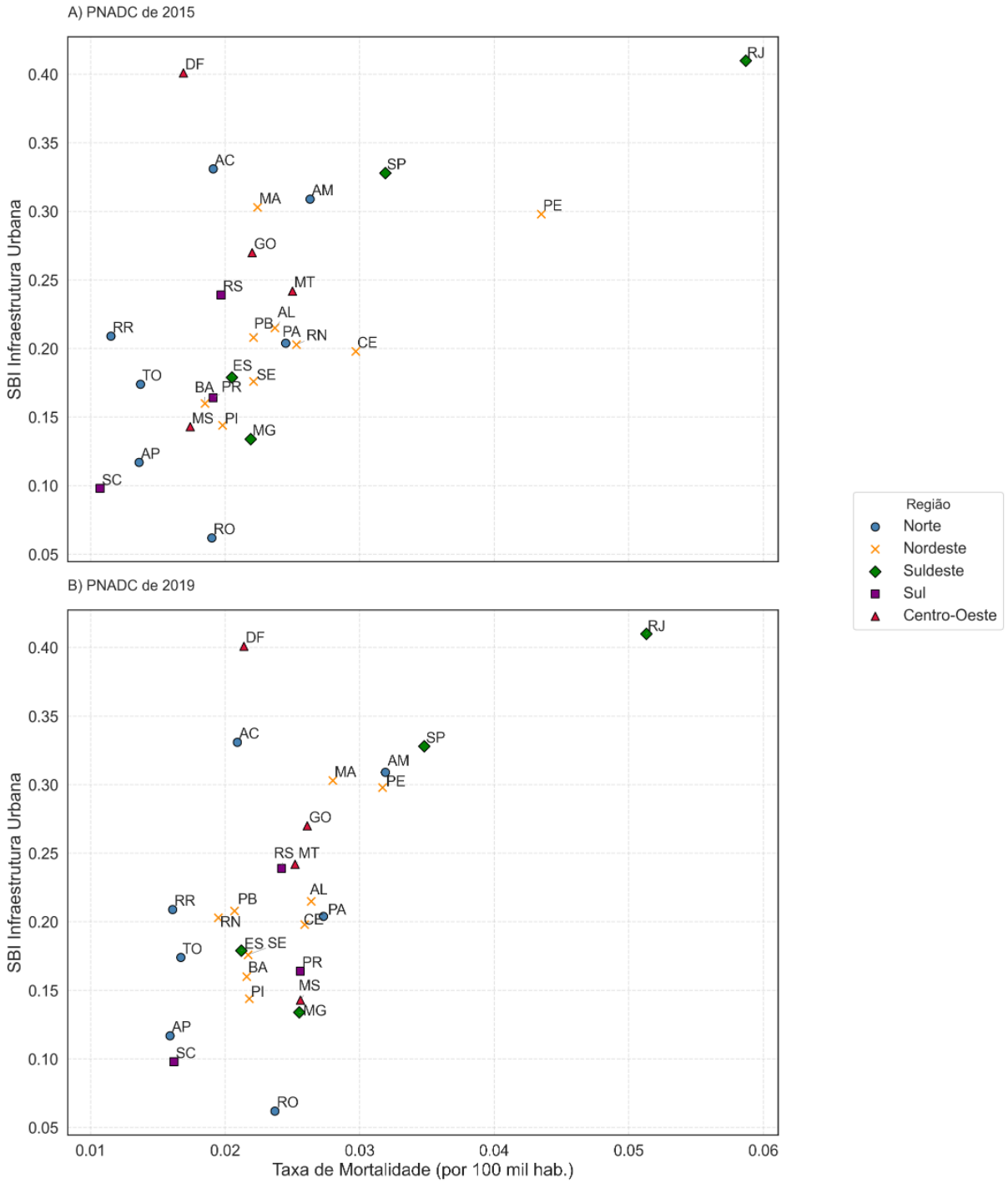
A análise dos dados de 2019 confirma essa tendência, uma vez que o IVS Infraestrutura Urbana manteve-se como o indicador com o maior nível de correlação. Tal persistência evidencia a ausência de melhorias significativas na infraestrutura urbana ao longo dos anos, refletindo uma condição crônica de vulnerabilidade em diversas regiões do país. Isso aponta para a necessidade urgente de políticas públicas voltadas à requalificação urbana, sobretudo em territórios com históricos de exclusão social.

Outro fator de destaque nos dados de 2015 é o tempo médio de deslocamento entre a residência e o local de trabalho, que também apresentou um aumento nos níveis de correlação com as taxas de mortalidade por COVID-19. Essa variável revela o impacto da mobilidade urbana sobre a saúde pública, especialmente em contextos de pandemia. Longos trajetos diários, realizados majoritariamente em condições precárias de transporte coletivo, elevam o risco de contágio e dificultam o distanciamento social, comprometendo a eficácia das estratégias de contenção.

Contudo, essa análise não pôde ser replicada para o ano de 2019, devido à ausência de dados atualizados sobre o tempo de deslocamento nas bases utilizadas. Essa lacuna ressalta a importância da manutenção contínua de sistemas de informação socioeconômica, essenciais para a formulação de diagnósticos precisos e para o desenvolvimento de políticas públicas eficazes. As Figuras 27 e 28 ilustram, respectivamente, os gráficos de dispersão entre a taxa de óbitos por 100 mil habitantes nos anos de 2020 e 2021 com o IVS Infraestrutura Urbana e com o tempo médio de deslocamento casa-trabalho, contribuindo

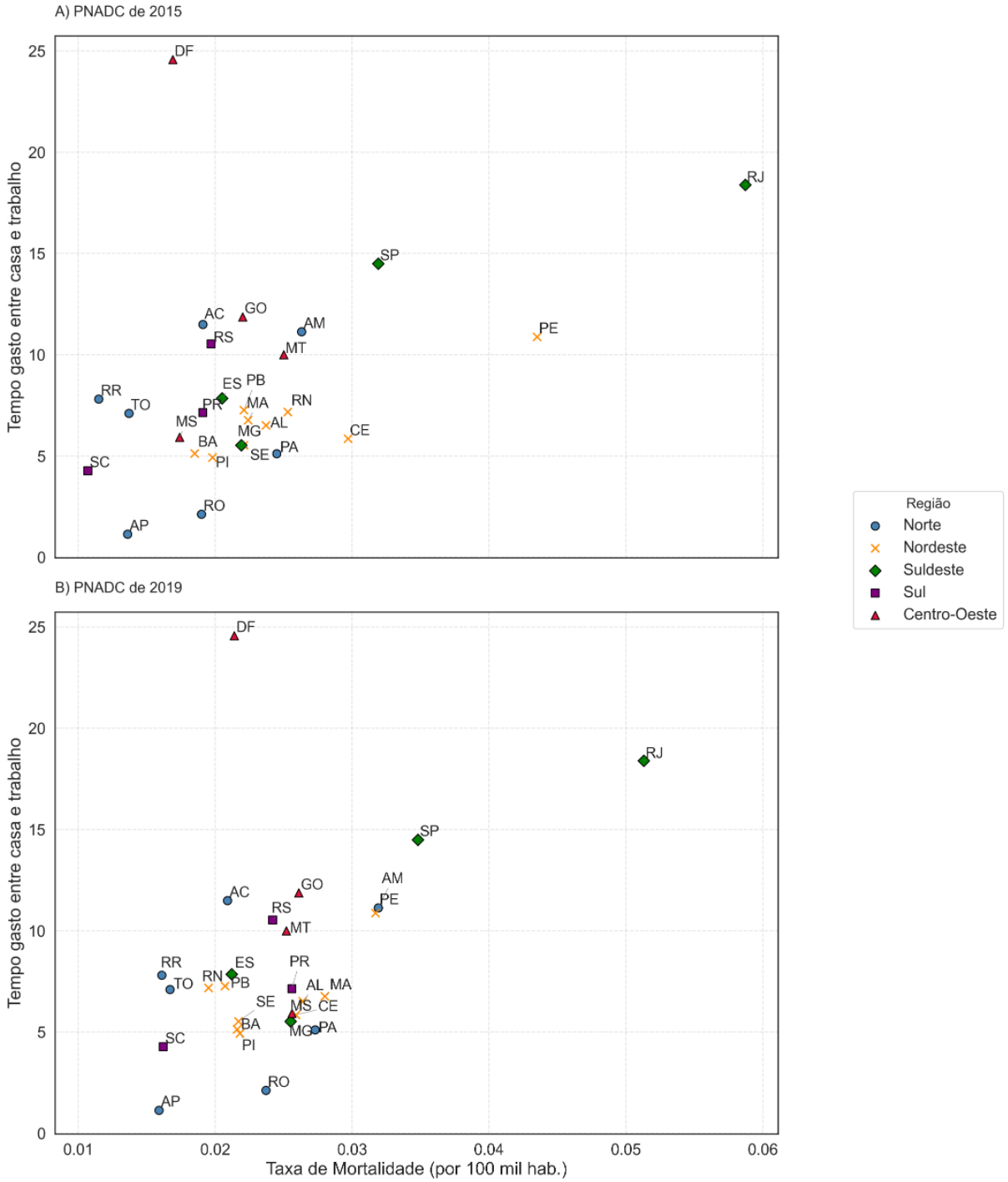
visualmente para a compreensão dessas relações.

Figura 27 – Dispersão da taxa de mortalidade por COVID-19 (por 100 mil hab.) em relação à infraestrutura urbana do IVS nos estados brasileiros.



Fonte: Elaborado pelo autor, (2024).

Figura 28 – Dispersão da taxa de mortalidade por COVID-19 (por 100 mil hab.) nos estados em relação ao tempo casa-trabalho.



Fonte: Elaborado pelo autor, (2023).

Embora o agravamento da pandemia de COVID-19, refletido principalmente na taxa de óbitos, tenha impactado os estados brasileiros de forma desigual entre 2020 e 2021, observa-se um padrão comum: os estados com maiores taxas de mortalidade são aqueles que apresentam maior vulnerabilidade em infraestrutura urbana, incluindo dificuldades no transporte público e longos deslocamentos diários entre casa e trabalho.

Diante desse cenário, torna-se essencial analisar como as desigualdades no acesso a serviços básicos de saneamento impactam a saúde pública, especialmente em momentos de crise sanitária. Assim, este estudo também investiga as disparidades regionais na oferta de água tratada, na cobertura da rede de esgoto e na regularidade da coleta de lixo domiciliar. A precariedade nesses serviços não apenas compromete a qualidade de vida da população, mas também favorece a propagação de doenças infecciosas, incluindo arboviroses e enfermidades de veiculação hídrica, que se tornam ainda mais preocupantes em contextos de epidemias e pandemias.

Os dados obtidos indicam que as regiões Norte e Nordeste apresentam os maiores déficits de cobertura desses serviços, com destaque para municípios de menor porte, localizados em áreas rurais e de difícil acesso geográfico. Tais regiões concentram a maior proporção de domicílios sem acesso à infraestrutura básica de saneamento, refletindo a carência histórica de investimentos públicos e a consequente vulnerabilidade socioeconômica dessas populações.

Por outro lado, as regiões Sul e Sudeste apresentam os melhores índices de cobertura em saneamento básico, com a maioria dos municípios atendidos por serviços de abastecimento de água tratada, rede de esgotamento sanitário e coleta regular de resíduos sólidos. No entanto, mesmo nessas regiões, persistem desigualdades importantes, especialmente em áreas rurais e periferias urbanas, que ainda enfrentam desafios significativos no acesso a esses serviços essenciais. A Tabela 22 apresenta a distribuição percentual dos déficits de saneamento básico por região do Brasil, possibilitando uma análise comparativa das disparidades existentes.

Tabela 22 – Distribuição regional dos déficits em saneamento básico no Brasil, segundo o Censo de 2022.

Região	**Sem Água Tratada (%)**	**Sem Esgoto (%)**	**Sem Coleta de Lixo (%)**
Norte	43,55	75,59	19,52
Nordeste	22,60	56,94	16,70
Centro-Oeste	17,13	34,00	4,61
Sul	13,11	36,27	4,61
Sudeste	8,83	13,32	2,97

Fonte: Elaborado pelo autor, (2024).

Observa-se que a Região Norte apresenta os piores indicadores de acesso, com aproximadamente 43,55% dos domicílios sem abastecimento de água tratada, 75,59% sem conexão à rede de esgoto e 19,52% sem coleta regular de lixo. Esses dados revelam um cenário alarmante de precariedade na infraestrutura básica, que contribui diretamente para a exposição da população a riscos sanitários e à proliferação de doenças infecciosas, como as arboviroses (dengue, zika e chikungunya), além de enfermidades de veiculação hídrica.

A Região Nordeste, embora apresente uma situação menos crítica em comparação ao Norte, também registra índices elevados de deficiência: 22,60% dos domicílios sem acesso à água tratada, 56,94% sem esgotamento sanitário adequado e 16,70% sem coleta de lixo. Nessas áreas, as condições ambientais frequentemente favorecem a disseminação de doenças, agravadas pela limitação no acesso a serviços de saúde.

Nas regiões Centro-Oeste, Sul e Sudeste, os indicadores de saneamento são significativamente mais positivos. A Região Sudeste destaca-se por apresentar os menores percentuais de domicílios sem acesso a esses serviços: 8,83% sem água tratada, 13,32% sem rede de esgoto e apenas 2,97% sem coleta de lixo. Ainda assim, é importante ressaltar as desigualdades intra-regionais, pois localidades rurais e periferias urbanas continuam enfrentando déficits expressivos.

Essa análise reforça a urgência de políticas públicas direcionadas à universalização do saneamento básico e à redução das desigualdades regionais. Investimentos contínuos em infraestrutura são essenciais para melhorar as condições de vida da população, mitigar os efeitos de futuras crises sanitárias e promover maior equidade em saúde pública no Brasil.

6.2.2 Relação entre Infraestrutura Urbana e Doenças Infecciosas no Brasil

Os resultados desta pesquisa evidenciam de forma contundente a relação direta entre a precariedade da infraestrutura urbana e os impactos adversos sobre a saúde pública no Brasil, com ênfase nas taxas de mortalidade por COVID-19. A análise confirma que a ausência ou insuficiência de serviços essenciais de saneamento básico — como o fornecimento de água tratada, o esgotamento sanitário e a coleta regular de resíduos sólidos — intensifica a vulnerabilidade de determinadas populações à exposição a agentes infecciosos, sejam eles de transmissão respiratória ou vetorial (BTASIL, 2025; (FUNASA), 2022).

Estudos anteriores já demonstravam que áreas com elevados déficits em infraestrutura urbana enfrentam desafios significativos para a contenção de doenças infecciosas, devido às condições sanitárias precárias e à maior exposição a vetores de enfermidades, como o Aedes aegypti (BTASIL, 2025; UNDP, 2023).

As taxas de mortalidade por COVID-19, expressas por 100 mil habitantes, apresentaram distribuição desigual entre as regiões brasileiras, refletindo disparidades estruturais históricas. Como apresentado na Tabela 23, as regiões Centro-Oeste e Sudeste, caracterizadas pelos maiores déficits de infraestrutura urbana, registraram as maiores taxas de mortalidade em 2021, considerado o período mais crítico da pandemia. Esses resultados corroboram a literatura que aponta a associação entre desigualdade regional e piores desfechos de saúde (BTASIL, 2025; (OPAS), 2023).

Tabela 23 – Evolução das taxas de mortalidade por COVID-19 (por 100 milhab) nas regiões do Brasil, de 2020 a 2022.

Região	**2020**	**2021**	**2022**
Norte	97,80	160,18	20,70
Nordeste	83,66	126,63	24,20
Centro-Oeste	109,52	254,42	38,11
Sudeste	100,97	232,46	43,76
Sul	73,72	251,60	41,00

Fonte: Elaborado pelo autor, (2024).

A análise revela que a precariedade da infraestrutura urbana no Brasil, especialmente nas regiões Norte e Nordeste, amplia a vulnerabilidade da população às doenças infecciosas. De acordo com o relatório da (FUNASA) (2022), apenas 49% da população brasileira possui acesso à rede de esgotamento sanitário, o que contribui para a perpetuação de cenários de risco sanitário. Além dos déficits estruturais, é necessário considerar o impacto das mudanças climáticas, que têm ampliado o período de transmissão das arboviroses e dificultado as estratégias de controle vetorial (BTASIL, 2025; UNDP, 2023).

Diante desse cenário, a ampliação dos investimentos em saneamento básico — com foco na expansão das redes de esgoto, no acesso universal à água tratada e na coleta regular de resíduos — é essencial para reduzir as desigualdades regionais e conter a disseminação de doenças infecciosas. Tais investimentos devem ser integrados a políticas públicas regionais que considerem as especificidades socioambientais de cada território.

É fundamental que futuras pesquisas explorem intervenções regionais específicas, considerando as características locais de infraestrutura e vulnerabilidade social. Além disso, é recomendada a integração de políticas de saúde pública e planejamento urbano, visando à redução das disparidades sanitárias no Brasil e ao fortalecimento do sistema de saúde pública (CAPRARA, 2023).

6.3 Considerações Finais

Este capítulo apresentou os resultados obtidos com a aplicação do método proposto, destacando o uso de algoritmos para a pré-seleção de dados e a análise conduzida em dois estudos de caso. Os padrões extraídos por meio de inferência com redes Bayesianas demonstraram a capacidade do modelo em identificar relações significativas entre as variáveis analisadas. De modo geral, os resultados atendem aos objetivos do trabalho, evidenciando que a abordagem proposta é objetiva, eficiente e de fácil compreensão para os usuários nas áreas de aplicação. Além disso, o método mostrou-se eficaz como suporte à tomada de decisão, oferecendo uma ferramenta analítica adequada a cenários complexos e marcados por incertezas, como os observados nos contextos estudados.

7 Conclusões e Trabalhos Futuros

Este capítulo apresenta as conclusões e considerações finais deste trabalho, com destaque para os principais resultados obtidos nos estudos de caso sobre o impacto da pandemia de COVID-19 no desempenho dos estudantes no ENEM e sua relação com a infraestrutura urbana deficiente. A pesquisa contribui para a compreensão das desigualdades educacionais e sanitárias, ao mesmo tempo em que evidencia as dificuldades enfrentadas na obtenção e integração dos dados utilizados. Por fim, são indicadas perspectivas para trabalhos futuros, com ênfase na ampliação das análises, no desenvolvimento de modelos preditivos integrados e na formulação de políticas públicas baseadas em evidências, voltadas à redução das desigualdades estruturais e ao fortalecimento da resiliência social.

7.1 Conclusões

Este estudo realizou uma análise aprofundada sobre o impacto da pandemia de COVID-19 no desempenho acadêmico e na participação dos estudantes no ENEM, com foco no estado do Pará, entre 2019 e 2022. Por meio do uso de Redes Bayesianas, foi possível identificar os fatores socioeconômicos e contextuais que mais influenciaram o acesso e o desempenho escolar durante o período pandêmico.

Apesar de eficientes para modelar relações probabilísticas, as Redes Bayesianas apresentaram limitações decorrentes das condições emergenciais impostas pela pandemia, as quais comprometeram a suposição de independência entre variáveis. A coleta de dados socioeconômicos também foi prejudicada por informações incompletas. Além disso, as disparidades na implementação de políticas educacionais e na infraestrutura das diferentes mesorregiões podem ter influenciado os resultados de maneira desigual.

Embora o estudo ofereça uma visão abrangente dos impactos da pandemia sobre o ENEM, ele apresenta limitações que podem ser superadas em pesquisas futuras, por meio da inclusão de novas variáveis e da adoção de métodos de coleta mais robustos, visando a uma compreensão mais aprofundada dos fatores que influenciam o desempenho educacional em cenários de crise.

No campo empírico, os resultados indicam que estudantes oriundos de famílias de baixa renda, majoritariamente matriculados em escolas públicas, foram os mais afetados pela interrupção das atividades presenciais e pela transição abrupta para o ensino remoto emergencial. A falta de acesso a recursos tecnológicos, especialmente à internet, configurou-se como um dos principais obstáculos à continuidade dos estudos, comprometendo significativamente a participação no ENEM e os resultados obtidos nas provas.

Destaca-se, entretanto, que o uso de computadores não se mostrou um fator relevante nos resultados da pesquisa, uma vez que grande parte dos estudantes utilizou telefones celulares como principal meio de acesso às plataformas de ensino. Nas regiões mais vulneráveis, como o Marajó e o Sudeste Paraense, observou-se um aumento expressivo nas taxas de evasão escolar. Em contraste, estudantes de escolas privadas mantiveram um desempenho mais estável, evidenciando a persistência das desigualdades estruturais no sistema educacional brasileiro.

A análise também revelou que a escolaridade dos pais — especialmente a formação da mãe — e a ocupação profissional dos responsáveis foram determinantes para a manutenção da participação e do desempenho escolar durante a pandemia. Estudantes cujas mães possuíam ensino superior apresentaram menores taxas de abandono escolar e melhores notas no ENEM, reforçando o papel da estrutura familiar como um fator de proteção diante dos desafios impostos pela crise sanitária.

Com base nos dados e resultados obtidos, foi possível confirmar as hipóteses propostas nesta tese, tanto por meio da análise estatística quanto pelas evidências empíricas observadas:

A Hipótese 1 foi validada pela identificação de padrões de desempenho escolar mais baixos e de maiores índices de evasão em regiões com infraestrutura urbana precária e indicadores sociais críticos, como o Marajó e o Sudeste Paraense. Os dados analisados evidenciaram que a vulnerabilidade social exerceu um papel central na intensificação dos impactos da pandemia, reforçando a necessidade de políticas públicas territorializadas e sensíveis às desigualdades estruturais presentes em diferentes contextos regionais.

A Hipótese 2 foi confirmada pelos resultados que demonstram que estudantes oriundos de famílias de baixa renda — matriculados em escolas públicas e com acesso limitado a recursos tecnológicos e a serviços básicos — foram os mais afetados durante o período pandêmico. O desempenho escolar inferior e a elevada taxa de evasão observados nesse grupo evidenciam o impacto direto da desigualdade social nas consequências educacionais geradas pela crise sanitária.

Além da investigação educacional, foi realizada uma análise complementar no Estudo de Caso II, com foco na relação entre o IVS e o IDH, no período de 2018 a 2023. Os resultados indicaram que o IVS apresentou uma correlação mais consistente com os indicadores de saúde analisados, ao contrário do IDH, cuja relação foi menos significativa. Municípios com maiores deficiências em saneamento básico — como ausência de água tratada e coleta inadequada de lixo — registraram maior incidência de doenças e taxas mais elevadas de mortalidade, tanto por arboviroses quanto por COVID-19.

A conexão entre infraestrutura urbana, saúde pública e educação ficou evidente ao se observar que as mesmas regiões mais afetadas por doenças infecciosas e deficiência

em serviços básicos também concentraram os piores indicadores educacionais durante a pandemia. Esses dados reforçam a necessidade de compreender a desigualdade social de forma intersetorial, reconhecendo que os determinantes sociais de saúde e educação estão interligados e exigem respostas integradas das políticas públicas.

Em síntese, os principais achados desta pesquisa demonstram que:

- A pandemia de COVID-19 ampliou as desigualdades educacionais preexistentes, especialmente no acesso ao ensino superior por meio do ENEM;
- O acesso a recursos tecnológicos foi um fator decisivo para a continuidade do aprendizado e o desempenho dos estudantes, aprofundando as disparidades entre alunos da rede pública e privada;
- A infraestrutura urbana precária, especialmente no Norte e Nordeste do país, contribuiu para o agravamento dos impactos sanitários e educacionais, criando um ciclo de vulnerabilidade social;
- Políticas públicas que priorizem investimentos em infraestrutura tecnológica, digital e urbana, aliadas a ações de inclusão social e apoio psicossocial, são fundamentais para a redução das desigualdades e o fortalecimento da resiliência frente a futuras crises sanitárias e educacionais.

Portanto, este trabalho reforça a urgência na adoção de políticas públicas baseadas em dados, que considerem as especificidades regionais e as múltiplas dimensões da desigualdade social, buscando a construção de um sistema educacional e de saúde mais justo, inclusivo e preparado para lidar com desafios emergenciais.

7.2 Trabalhos Futuros

A partir dos resultados apresentados, destacam-se as seguintes propostas para trabalhos futuros:

1. **Análise Longitudinal e Expansão Geográfica**: Ampliar a análise para outros estados brasileiros, considerando as diferentes realidades regionais, com foco especial nos estados do Norte e Nordeste, a fim de comparar os impactos educacionais e identificar padrões nacionais;
2. **Avaliação do Impacto das Políticas Públicas Pós-Pandemia**: Investigar a efetividade das políticas públicas implementadas após a pandemia, como programas de recuperação de aprendizagem, reforço escolar e inclusão digital, avaliando seus efeitos sobre a redução das desigualdades educacionais;

3. **Integração de Dados de Saúde e Educação para Modelos Preditivos**: Desenvolver modelos integrados que combinem dados educacionais, de saúde pública e de infraestrutura urbana, com o objetivo de criar sistemas de alerta precoce para a identificação de populações vulneráveis durante crises sanitárias e educacionais;

4. **Estudos sobre Saúde Mental e Bem-Estar dos Estudantes**: Explorar o impacto da pandemia na saúde mental dos estudantes, investigando a relação entre bem-estar emocional e desempenho acadêmico, especialmente em comunidades de baixa renda;

5. **Desenvolvimento de Ferramentas de Apoio à Tomada de Decisão**: Propor o desenvolvimento de plataformas digitais baseadas em ciência de dados que permitam o monitoramento em tempo real de indicadores educacionais e sanitários, com vistas a subsidiar decisões estratégicas de gestores públicos em contextos de emergência.

Em síntese, os desdobramentos propostos buscam promover respostas integradas e baseadas em evidências para reduzir desigualdades estruturais e fortalecer a resiliência social, com foco na educação básica e na saúde pública das populações mais vulneráveis.

7.3 Publicações

Durante o processo de pesquisa acadêmica, foram desenvolvidos e publicados trabalhos voltados à aplicação de *Data Science* sobre dados abertos. Dentre essas contribuições, destaca-se a seguinte publicação derivada do Estudo de Caso I desta tese:

- SANTOS, Sandio Maciel dos; SILVA, Marcelino Silva da; FRANÇA LOBATO, Fábio Manoel; FRANCÊS, Carlos Renato Lisboa. **Use of Bayesian networks in Brazil high school educational database: analysis of the impact of COVID-19 on ENEM in Pará between 2019 and 2022**. *Frontiers in Big Data*, v. 8, 2025. Disponível em: <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2025.1485493>. Acesso em: 11 abr. 2025. DOI: 10.3389/fdata.2025.1485493. ISSN 2624-909X.

Referências

ABUQUERQUE, R. L. Fernandes de. Enem durante a pandemia? um estudo de caso das percepções de docentes da rede estadual de educação do rio de janeiro sobre a realização do enem 2020. *Olhar de Professor*, v. 23, p. 1–6, set. 2020. Disponível em: <https://revistas.uepg.br/index.php/olhardeprofessor/article/view/15649>. Citado na página 68.

ANACONDA INC. *Anaconda Software Distribution*. Anaconda Inc., 2020. Disponível em: <https://readthedocs.com/projects/continuumio-conda/downloads/pdf/latest/>. Citado na página 37.

ANKAN, A.; TEXTOR, J. pgmpy: A python toolkit for bayesian networks. *Journal of Machine Learning Research*, v. 25, n. 265, p. 1–8, 2024. Disponível em: <http://jmlr.org/papers/v25/23-0487.html>. Citado na página 38.

ATLAS IVS. *Atlas da Vulnerabilidade Social*. IPEA DATA, 2025. Disponível em: <http://ivs.ipea.gov.br/index.php/pt/sobre>. Citado 3 vezes nas páginas 34, 35 e 39.

BARATA, R. d. C. B. O desafio das doenças emergentes e a revalorização da epidemiologia descritiva. *Revista de Saúde Pública*, SciELO Brasil, v. 31, p. 531–537, 1997. Citado na página 3.

BARBOSA, J. A. A aplicabilidade da tecnologia na pandemia do novo coronavírus (covid-19). *Revista da FAESF*, v. 4, 2020. Citado na página 3.

BARMAN, A. et al. Respiratory rehabilitation in patients recovering from severe acute respiratory syndrome: A systematic review and meta-analysis. *Heart & Lung*, v. 53, p. 11–24, 2022. ISSN 0147-9563. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0147956322000061>. Citado na página 17.

BARTHOLO, T. L. et al. Learning loss and learning inequality during the covid-19 pandemic. *Ensaio: Avaliação e Políticas Públicas em Educação*, Fundação CESGRANRIO, v. 31, n. 119, p. e0223776, 2023. ISSN 0104-4036. Disponível em: <https://doi.org/10.1590/S0104-40362022003003776>. Citado 2 vezes nas páginas 10 e 11.

BERKLEY, S. *COVAX explained. To end this global health crisis we don't just need COVID-19 vaccines, we also need to ensure that everyone in the world has access to them.* 2020. <https://www.gavi.org/vaccineswork/covax-explained>. Online; accessed 06 July 2022. Citado na página 19.

BIENER, C.; LANDMANN, A.; SANTANA, M. I. Contract nonperformance risk and uncertainty in insurance markets. *Journal of Public Economics*, v. 175, p. 65–83, 2019. ISSN 0047-2727. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0047272719300660>. Citado na página 69.

BLACKWELL, C. K. et al. Sociodemographic differences in parental hesitancy to the covid-19 vaccine. *Vaccine*, v. 55, p. 127041, 2025. Citado 2 vezes nas páginas 9 e 10.

BOSCHETTI, A.; MASSARON, L. *Python Data Science Essentials - Learn the fundamentals of Data Science with Python*. Birmingham: Packt Publishing, 2015. ISBN 978-1-78528-042-9. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=e880b26d877ff1fa8f20a6f0bb5b89bd>. Citado na página 26.

BRAGA, J. V. *Gestão da informação Governamental: Um Modelo Harmonizado de Curadoria Digital Para Dados Abertos Governamentais*. Tese (Doutorado) — Universidade Fernando Pessoa (Portugal), 2023. Citado na página 4.

BRASIL. *Guia para Investigações de Surtos ou Epidemias / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis*. Brasília, 2018. Citado 2 vezes nas páginas 15 e 16.

BRASIL. *Guia para Investigações de Surtos ou Epidemias / Ministério da Saúde, Secretaria de Vigilância em Saúde, Departamento de Vigilância das Doenças Transmissíveis – Brasília : Ministério da Saúde*. 2018. <https://legis.senado.leg.br/sdleg-getter/documento?dm=8075954&ts=1584647908386&disposition=inline>. Online; accessed 17 December 2022. Citado na página 3.

BRASIL. *Ministério da Educação. Mais de 3,9 milhões de candidatos participam do primeiro dia do Enem 2019*. Gov.br, 2019. Disponível em: <http://portal.mec.gov.br/pronatec/oferta-voluntaria/418-noticias/enem-946573306>. Citado na página 57.

BRASIL. *CORONAVÍRUS - BRASIL. Painel de casos de doença pelo coronavírus 2019 (COVID-19) no Brasil pelo Ministério da Saúde*. 2020. <https://covid.saude.gov.br/>. Online; accessed 17 December 2024. Citado 2 vezes nas páginas 2 e 17.

BRASIL. *Câmara dos Deputados. Reconhece, para os fins do art. 65 da Lei Complementar n. 101, de 4 de maio de 2000, a ocorrência do estado de calamidade pública, nos termos da solicitação do Presidente da República encaminhada por meio da Mensagem n. 93*. 2020. <https://legis.senado.leg.br/sdleg-getter/documento?dm=8075954&ts=1584647908386&disposition=inline>. Online; accessed 06 July 2024. Citado na página 17.

BRASIL. *DATASUS. Sistema de Informação da Vigilância Epidemiológica da Gripe (SIVEP Gripe)*. Ministério da Saúde, 2020. Online; accessed 15 July 2024. Disponível em: <https://datasus.saude.gov.br/notifica/>. Citado na página 20.

BRASIL. *Diário Oficial da União. Portaria Nº 454, de 20 de Março de 2020. Declara, em todo o território nacional, o estado de transmissão comunitária do coronavírus (covid-19)*. 2020. <https://www.in.gov.br/en/web/dou/-/portaria-n-454-de-20-de-marco-de-2020-249091587>. Online; accessed 06 July 2024. Citado na página 18.

BRASIL. *Ministério da Saúde. Portaria MS/GM n. 188, de 3 de fevereiro de 2020. Declara Emergência em Saúde Pública de importância Nacional em decorrência da Infecção Humana pelo novo Coronavírus (2019-nCoV)*. 2020. Https://www.in.gov.br/en/web/dou/-/portaria-n-188-de-3-de-fevereiro-de-2020-241408388. Citado na página 2.

BRASIL. *Ministério da Saúde/Gabinete do Ministro. Diário Oficial da União. Portaria Nº 188, de 3 de Fevereiro de 2020. Declara Emergência em Saúde Pública de importância Nacional (ESPIN) em decorrência da Infecção Humana pelo novo Coronavírus (2019-nCoV)*. 2020. <https://legis.senado.leg.br/sdleg-getter/documento?dm=8075954&

ts=1584647908386&disposition=inline>. Online; accessed 06 July 2024. Citado na página 17.

BRASIL. *PORTARIA Nº 1.792, DE 17 DE JULHO DE 2020. Altera a Portaria nº 356/GM/MS, de 11 de março de 2020, para dispor sobre a obrigatoriedade de notificação ao Ministério da Saúde de todos os resultados de testes diagnóstico para SARS-CoV-2 realizados por laboratórios da rede pública, rede privada, universitários e quaisquer outros, em todo território nacional.* [S.l.]: Ministério da Saúde, 2020. Online; accessed 15 July 2024. Citado na página 20.

BRASIL. *Orientações sobre notificação e registros de casos de Covid-19 no Brasil.* [S.l.]: gov.br, 2021. Online; accessed 15 July 2024. Citado na página 20.

BRASIL. *Serviços e Informações do Brasil. Taxa de comparecimento no 1º dia do Enem chega a 74% dos inscritos.* Gov.br, 2021. Disponível em: <http://portal.mec.gov.br/pronatec/oferta-voluntaria/418-noticias/enem-946573306>. Citado 2 vezes nas páginas 57 e 58.

BRASIL. *DATASUS. SIM - Sistema de Informação sobre Mortalidade.* Secretária de Vigilância em Saúde, 2022. Online; accessed 15 July 2024. Disponível em: <http://sim.saude.gov.br/default.asp>. Citado na página 21.

BRASIL. *Vacinas - Covid-19.* [S.l.]: Agência Nacional de Vigilância Sanitária - Anvisa, 2022. Online; accessed 07 July 2024. Citado na página 19.

BRASIL. *Ministerio da Saúde - DATASUS.* 2023. Disponível em: <https://datasus.saude.gov.br/>. Acessado em: 14 de Julho 2023. Citado na página 15.

BRITO, C. *Rio registrou aumento de 31% no número de turistas durante o carnaval.* 2020. <https://g1.globo.com/rj/rio-de-janeiro/carnaval/2020/noticia/2020/03/02/rio-registrou-aumento-de-31percent-no-numero-de-turistas-durante-o-carnaval.ghtml>. Online; accessed 06 July 2022. Citado na página 17.

BRITO, W. H. d.; PEDROSO, F. P. Impactos de variáveis socioeconômicas no desempenho no enem no primeiro biênio da pandemia de covid-19. *Regae: Revista de Gestão e Avaliação Educacional*, v. 12, n. 21, 2023. Citado 2 vezes nas páginas 68 e 69.

BTASIL, I. T. B. T. *Saneamento é saúde: Como a falta de acesso à infraestrutura básica afeta as incidências de doenças relacionadas ao saneamento ambiental inadequado no Brasil?* 2025. Acesso em: 11 abr. 2025. Disponível em: <https://tratabrasil.org.br/estudos-trata-brasil/>. Citado 2 vezes nas páginas 87 e 88.

BUENO, F. T. C.; SOUTO, E. P.; MATTA, G. C. Notas sobre a trajetória da covid-19 no brasil. In: . [S.l.]: Fiocruz, 2021. Citado 5 vezes nas páginas 1, 2, 17, 18 e 19.

BUTANTAN. *Entenda o que é uma pandemia e as diferenças entre surto, epidemia e endemia.* [S.l.]: Instituto Butantan, 2022. Online; accessed 20 Agu 2022. Citado na página 15.

CADY, F. *The Data Science Handbook.* [S.l.]: Wiley, 2017. Citado na página 26.

CAO, L. Data science: Challenges and directions. *Commun. ACM*, Association for Computing Machinery, New York, NY, USA, v. 60, n. 8, p. 59–68, jul 2017. ISSN 0001-0782. Disponível em: <https://doi.org/10.1145/3015456>. Citado na página 26.

CAPRARA, A. *Vigilância Participativa da Comunidade para o Combate das Arboviroses Transmitidas pelo Aedes Aegypti em uma Intervenção Eco-Bio-Social (Tracking)*. 2023. Acesso em: 11 abr. 2025. Citado na página 88.

CARDOSO, E. H. S. et al. Characterizing the impact of social inequality on covid-19 propagation in developing countries. *IEEE Access*, v. 8, p. 172563–172580, 2020. Citado na página 4.

CARMO, R. V. do; HECKLER, W. F.; CARVALHO, J. V. de. Uma análise do desempenho dos estudantes do rio grande do sul no enem 2019. *RENOTE*, v. 18, n. 2, p. 378–387, 2020. Citado na página 48.

CASTRO, M. C. et al. Demand for hospitalization services for covid-19 patients in brazil. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Disponível em: <https://www.medrxiv.org/content/early/2020/04/01/2020.03.30.20047662>. Citado na página 18.

CASTRO, M. H.; SOARES, J. F. *Nova forma de divulgar dados do Enem e do Censo Escolar contraria interesse público*. 2021. Disponível em: <https://portaliede.org.br/contribuicao/nova-forma-de-divulgar-dados-do-enem-e-do-censo-escolar-pelo-inep-contraria-interesse-publico-e-invi Citado na página 68.

CEREDA, D. et al. *The early phase of the COVID-19 outbreak in Lombardy, Italy*. arXiv, 2020. Disponível em: <https://arxiv.org/abs/2003.09320>. Citado na página 18.

CERQUEIRA, D. R. et al. Escassez hidroambiental na produção do espaço urbano do oeste da região metropolitana do rio de janeiro: crises e conflitos na gestão e provisão de água das redes do sistema paraíba-piraí-guandu. Universidade do Estado do Rio de Janeiro, 2025. Citado na página 3.

CHAPMAN, P. et al. Crisp-dm 1.0: Step-by-step data mining guide. In: . [S.l.: s.n.], 2000. Citado 2 vezes nas páginas 27 e 40.

CHIGNARD, S. *A brief history of Open Data*. 2013. Disponível em: <http://parisinnovationreview.com/articles-en/a-brief-history-of-open-data>. Acessado em: 10 de Novembro 2023. Citado na página 14.

CIOTTI, M. et al. The covid-19 pandemic. *Critical reviews in clinical laboratory sciences*, Taylor & Francis, v. 57, n. 6, p. 365–388, 2020. Citado na página 1.

COELHO, F. C. et al. Assessing the spread of covid-19 in brazil: Mobility, morbidity and social vulnerability. *PLoS One*, Public Library of Science San Francisco, CA USA, v. 15, n. 9, p. e0238214, 2020. Citado na página 2.

COSTA, A. F. d. Desigualdades sociais e pandemia. *Um olhar sociológico sobre a crise Covid-19 em livro*, Observatório das Desigualdades, 2020. Citado na página 3.

CSORBA, L. M.; DABIJA, D.-C. The impact of the covid-19 pandemic on students' future online education behaviour. *Heliyon*, v. 10, n. 20, p. e39560, 2024. Citado 2 vezes nas páginas 11 e 12.

CUNHA, F. P. d. A gripe hespanhola em santa maria (1918): um estudo do flagelo na imprensa local. Universidade Federal de Santa Maria, 2024. Citado na página 1.

DANA, S. et al. Brazilian modeling of covid-19(bram-cod): a bayesian monte carlo approach for covid-19 spread in a limited data set context. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Disponível em: <https://www.medrxiv.org/content/early/2020/05/03/2020.04.29.20081174>. Citado 3 vezes nas páginas 1, 3 e 18.

DARWICHE, A. *Modeling and Reasoning with Bayesian Networks*. [S.l.]: Cambridge University Press, 2009. Citado na página 29.

DAVENPORT, T. H. How strategists use "big data" to support internal business decisions, discovery and production. *Strategy & Leadership*, v. 42, p. 45–50, 2014. ISSN 45-50. Disponível em: <https://www.emerald.com/insight/content/doi/10.1108/SL-05-2014-0034/full/html?skipTracking=true>. Citado 2 vezes nas páginas 24 e 25.

DAVIS, J. T. et al. Estimating the establishment of local transmission and the cryptic phase of the covid-19 pandemic in the usa. *medRxiv*, Cold Spring Harbor Laboratory Press, 2020. Disponível em: <https://www.medrxiv.org/content/early/2020/08/28/2020.07.06.20140285>. Citado na página 18.

DEEPA, N. et al. A survey on blockchain for big data: Approaches, opportunities, and future directions. *Future Generation Computer Systems*, Elsevier, v. 131, p. 209–226, 2022. Citado na página 24.

DEGAN, J. O. C. *Intergração de Dados Corporativos: Uma Proposta de Arquitetura Baseada em Serviços de Dados*. 7-24 p. Dissertação (Mestrado) — Universidade Federal de Campinas, São Paulo, 2005. Citado na página 23.

DELATORRE, E. et al. Tracking the onset date of the community spread of sars-cov-2 in western countries. *Memórias do Instituto Oswaldo Cruz [online]*, v. 115, sept 2020. Disponível em: <https://doi.org/10.1590/0074-02760200183>. Citado na página 17.

DEVORE, J. L. *Probabilidade e Estatística para Engenharia e Ciências*. 8. ed. [S.l.]: Cengage Learning, 2015. ISBN 978-85-221-1183-1. Citado na página 23.

DHARMAYANI, P. N. A.; MIHRSHAHI, S. The prevalence of psychological distress and its associated sociodemographic factors in the australian adults aged 18–64 years during covid-19: Data from the australian national health survey. *Journal of Affective Disorders*, v. 368, p. 312–319, 2025. Citado na página 10.

DIAS, É.; RAMOS, M. N. A educação e os impactos da covid-19 nas aprendizagens escolares. *Ensaio: Avaliação e Políticas Públicas em Educação*, Fundação Cesgranrio, v. 30, n. 117, p. 859–870, 2022. Citado na página 63.

DUCAMP, G.; GONZALES, C.; WUILLEMIN, P.-H. agrum/pyagrum: a toolbox to build models and algorithms for probabilistic graphical models in python. In: *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. PMLR, 2020. (Proceedings of Machine Learning Research, v. 138), p. 609–612. Disponível em: <https://proceedings.mlr.press/v138/ducamp20a.html>. Citado na página 38.

ELLWANGER, J. H.; CHIES, J. A. B. Saúde única (one health): uma abordagem para entender, prevenir e controlar as doenças infecciosas e parasitárias. *Bio Diverso*, v. 2, n. 1, 2022. Citado na página 1.

EMARA, N. et al. Analyzing non-linear and interactive impacts of distance learning on college enrollment post-covid-19. *Economic Analysis and Policy*, v. 85, p. 2310–2325, 2025. Citado na página 12.

EMARA, N. et al. Analyzing non-linear and interactive impacts of distance learning on college enrollment post-covid-19. *Economic Analysis and Policy*, v. 85, p. 2112–2125, 2025. Citado 2 vezes nas páginas 12 e 13.

ENAP, E. N. d. A. P. *RESOLUÇÃO ENAP Nº 7, DE 8 DE ABRIL DE 2022. Aprova o Plano de Dados Abertos (PDA) 2022-2024 da Fundação Escola Nacional de Administração Pública - Enap.* Gov.br, 2022. Disponível em: <https://sei.enap.gov.br/sei/publicacoes/controlador_publicacoes.php?acao=publicacao_visualizar&id_documento=565258&id_orgao_publicacao=0>. Citado na página 14.

FERNANDES, A. C. As grandes pandemias da história da europa e os seus impactos na nossa civilização: desafios da moderna saúde pública. *Cadernos Ibero-Americanos de Direito Sanitário*, v. 10, n. 2, p. 19–30, jun. 2021. Disponível em: <https://www.cadernos.prodisa.fiocruz.br/index.php/cadernos/article/view/780>. Citado na página 1.

FIOCRUZ. *Fundação OsWaldo Cruz.Technologial Order agreement term Nº 01/2020.* 2020. <https://portal.fiocruz.br/sites/portal.fiocruz.br/files/documentos/contrato_vacina_astrazaneca_fiocruz.pdf>. Online; accessed 06 July 2022. Citado na página 19.

FIOCRUZ. *Impactos sociais da pandemia.* 2020. <https://portal.fiocruz.br/impactos-sociais-economicos-culturais-e-politicos-da-pandemia>. Online; accessed 20 december 2022. Citado na página 4.

FORTUNA, S.; SETIAWAN, R. P.; SHARIFI, A. Do spatial and sociodemographic factors affect the transmission pattern of covid-19? evidence from surabaya city, indonesia. *International Journal of Disaster Risk Reduction*, v. 96, p. 103900, 2023. Citado 2 vezes nas páginas 8 e 9.

FOUST, A. M. et al. Pediatric sars, h1n1, mers, evali, and now coronavirus disease (covid-19) pneumonia: what radiologists need to know. *American Journal of Roentgenology*, American Roentgen Ray Society, v. 215, n. 3, p. 736–744, 2020. Citado na página 1.

FRANK, E.; WITTEN, I. *Data Mining: Practical Machine Learning Tools and Techniques With Java Implementations.* [s.n.], 2000. Chapter 8. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=787a145b668449b75f4e6f38d13ad06e>. Citado na página 26.

(FUNASA), F. N. de S. *Manejo de Resíduos Sólidos.* 2022. Acesso em: 11 abr. 2025. Disponível em: <http://www.funasa.gov.br/manejo-de-residuos-solidos>. Citado 2 vezes nas páginas 87 e 88.

GAMMA, J. et al. *Extração de Conhecimento de Dados - Data Mining.* 3. ed. [S.l.]: Sílabo, 2017. ISBN 978-972-618-914-5. Citado 2 vezes nas páginas 23 e 24.

GARCÍA LOZANO, M. et al. Veracity assessment of online data. *Decision Support Systems*, v. 129, p. 113132, 2020. ISSN 0167-9236. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0167923619301617>. Citado na página 25.

GHASEMIAN, A.; HOSSEINMARDI, H.; CLAUSET, A. Evaluating overfit and underfit in models of network community structure. *IEEE Transactions on Knowledge and Data Engineering*, v. 32, n. 9, p. 1722–1735, 2020. Citado na página 24.

GOLDSCHMIDT, R.; PASSOS, E.; BEZERRA, E. *Data Mining: conceito, técnicas, algoritmos orientações e aplicações*. 2. ed. Brasil: GEN LTC, 2015. ISBN 978-8535278224. Citado na página 41.

Green-Jr, S.; CHOW-WHITE, P. Data mining difference in the age of big data: Communication and the social shaping of genome technologies from 1998 to 2007. *International Journal of Communication*, v. 7, p. 1–28, 01 2013. Citado na página 24.

GRISOTTI, M. Doenças infecciosas emergentes e a emergência das doenças: uma revisão conceitual e novas questões. *Ciência & Saúde Coletiva*, SciELO Brasil, v. 15, p. 1095–1104, 2010. Citado na página 3.

GUNTHER, W. A. et al. Debating big data: A literature review on realizing value from big data. *The Journal of Strategic Information Systems*, v. 26, n. 3, p. 191–209, 2017. ISSN 0963-8687. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0963868717302615>. Citado na página 25.

HALLAL, P. C. et al. Remarkable variability in sars-cov-2 antibodies across brazilian regions: nationwide serological household survey. *medRxiv*, 2020. Citado na página 2.

HAYS, J. *Epidemics and Pandemics: Their Impacts on Human History*. 1. ed. ABC-CLIO, 2006. ISBN 9781851096589,9781851096633,1851096582. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=e6aad374d60902df6162ee8e654903f2>. Citado na página 1.

HRUSCHKA, E. R. *Iimputação Bayesiana no Contexto da Mineração de Dados*. Dissertação (Mestrado) — Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2003. Citado na página 32.

HUNTER, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, IEEE COMPUTER SOC, v. 9, n. 3, p. 90–95, 2007. Citado na página 37.

HURWITZ, J. et al. *Big Data For Dummies*. 1. ed. Wiley, 2013. ISBN 1118504224,9781118504222. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=5ff04eb38a838ebe2367dd3cdd52b59e>. Citado na página 23.

HURWITZ, J. et al. *Big Data For Dummies*. 1. ed. [S.l.]: Wiley, 2013. 169-174 p. ISBN 1118504224,9781118504222. Citado na página 24.

IBGE. *PNAD Contínua - Pesquisa Nacional por Amostra de Domicílios Contínua*. 2023. Disponível em: <http://https://www.ibge.gov.br/estatisticas/multidominio/condicoes-de-vida-desigualdade-e-pobreza/17270-pnad-continua.html?=&t=o-que-e>. Acessado em: 14 de Julho 2023. Citado na página 14.

IBGE. *Sistema IBGE de Recuperação Automática - SIDRA*. 2023. Disponível em: <https://sidra.ibge.gov.br/home/pmc/brasil/>. Acessado em: 14 de Julho 2023. Citado na página 14.

IEDE, I. *Mudança do INEP compromete análise de desempenho de alunos e desigualdade, diz ex-diretora*. 2021. Disponível em: <https://portaliede.org.br/contribuicao/g1-inep-omite-dados-do-enem-mudanca-compromete-analise-de-desempenho-de-alunos-e-desigualdade Citado na página 68.

INEP. *Ministerio da Educação*. 2023. Disponível em: <https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados>. Acessado em: 14 de Julho 2023. Citado na página 15.

JENSEN, F. V.; NIELSEN, T. D. *Bayesian Networks and Decision Graphs*. 2. ed. New York: Springer, 2007. Citado na página 66.

JORDAHL, K. et al. *geopandas/geopandas: v0.8.1*. Zenodo, 2020. Disponível em: <https://doi.org/10.5281/zenodo.3946761>. Citado na página 38.

JUSTEN, A.; et all. *Brasil.IO*. 2023. Disponível em: <https://brasil.io/home/>. Acessado em: 14 de Julho 2023. Citado na página 15.

KIM, B. et al. Covid-19 testing, case, and death rates and spatial socio-demographics in new york city: An ecological analysis as of june 2020. *Elsevier Ltd*, Health Place, v. 68, Mar 2021. ISSN 1679-4974. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/33639446/>. Citado 2 vezes nas páginas 1 e 16.

KLAUS-DIETER, G. *Integrated Business Information Systems: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data*. 2. ed. [S.l.]: SPRINGER, 2020. 81-85, 87-110 p. ISBN 9783662598108,3662598108. Citado 2 vezes nas páginas 24 e 25.

KLOSTERMAN, S. *Projetos de ciência de dados com Python: Abordagem de estudo de caso para a criação de projetos de ciência de dados bem-sucedidos usando Python, pandas e scikit-learn*. Brasil: Novatec Editora, 2020. ISBN 6586057116. Disponível em: <https://books.google.com.br/books?id=eX_iDwAAQBAJ>. Citado na página 23.

KLUYVER, T. et al. Jupyter notebooks – a publishing format for reproducible computational workflows. In: LOIZIDES, F.; SCHMIDT, B. (Ed.). *Positioning and Power in Academic Publishing: Players, Agents and Agendas*. [S.l.], 2016. p. 87 – 90. Citado na página 37.

KOLLER, N. F. D. *Probabilistic Graphical Models: Principles and Techniques*. 1. ed. [S.l.]: The MIT Press, 2009. (Adaptive Computation and Machine Learning series). ISBN 0262013193,9780262013192. Citado 2 vezes nas páginas 53 e 66.

KONG, M. et al. The impact of non-face to face education due to covid-19 pandemic on energy consumption and academic achievement. *Energy and Buildings*, v. 319, p. 114532, 2024. Citado na página 12.

KORB, K. B.; NICHOLSON, A. E. *Bayesian Artificial Intelligence*. 2nd. ed. [S.l.]: CRC Press, 2010. Citado na página 28.

KORB, K. B.; NICHOLSON, A. E. *Bayesian Artificial Intelligence*. 2. ed. [S.l.]: Taylor & Francis, 2010. (Chapman & Hall/CRC Computer Science & Data Analysis",). ISBN 1439815917,9781439815915. Citado na página 31.

KRISHNAN, K. *Building Big Data Applications*. 1. ed. [S.l.]: Academic Press, 2019. ISBN 0128157461,9780128157466. Citado na página 24.

LIM, C.; KWANG-JAE, K.; MAGLIO, P. P. Smart cities with big data: Reference models, challenges, and considerations. *Cities*, v. 82, p. 86–99, 2018. ISSN 0264-2751. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0264275117308545>. Citado na página 25.

LIMA, N. T. Pandemia e interdisciplinaridade: desafios para a saúde coletiva. *Saúde em Debate*, SciELO Public Health, v. 46, p. 9–24, 2023. Citado na página 4.

LIU, J. et al. A comparative overview of covid-19, mers and sars: Review article. *International Journal of Surgery*, v. 81, p. 1–8, 2020. ISSN 1743-9191. Disponível em: <https://www.sciencedirect.com/science/article/pii/S1743919120305677>. Citado na página 17.

LOPES, G. V. B.; COSTA, K. F. d. L. Impactos e desdobramentos da pandemia da covid-19 na atenção básica: um relato de experiência. 2020. Citado na página 18.

MAGALHãES, I. B. d. *Avaliação de redes Bayesianas para imputação de variáveis qualitativas e quantitativas*. Dissertação (Mestrado) — Escola Politécnica (USP), São Paulo, 2007. Citado na página 32.

MALAKAR, S. Modelagem geoespacial da vulnerabilidade covid-19 usando uma abordagem mcdm fuzzy integrada: um estudo de caso de bengala ocidental, Índia. *Modelo. Sistema Terra. Ambiente*, v. 8, p. 3103–3116, 2022. Disponível em: <https://link.springer.com/article/10.1007/s40808-021-01287-1>. Citado na página 8.

MARQUES, R. L.; DUTRA, I. Redes bayesianas: o que são, para que servem, algoritmos e exemplos de aplicações. UFRJ, 1999. Disponível em: <<http://www.cos.ufrj.br/~ines/courses/cos740/leila/cos740/Bayesianas.pdf>. Citado na página 31.

MARZ, N.; WARREN, A. *Big Data: Principles and Best Practices of Scalable Real-Time Data Systems*. MANNING, 2015. ISBN 9781617290343. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=cb6216671641e5788d3d6a4b974f59de>. Citado na página 24.

MARí-DELL'OLMO, M. et al. Socioeconomic inequalities in covid-19 in a european urban area: Two waves, two patterns. *International Journal of Environmental Research and Public Health*, v. 18, n. 3, 2021. ISSN 1660-4601. Disponível em: <https://www.mdpi.com/1660-4601/18/3/1256>. Citado na página 75.

MATTA, G. et al. *Os impactos sociais da Covid-19 no Brasil: populações vulnerabilizadas e respostas à pandemia [online]*. Rio de Janeiro: Observatório COVID-19: Editora FIOCRUZ, 2021. Citado na página 2.

MCKINNEY Wes. Data Structures for Statistical Computing in Python. In: WALT Stéfan van der; MILLMAN Jarrod (Ed.). *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 56 – 61. Citado na página 38.

MEDEIROS, C. A. d. *Estatística Aplicada à Educação*. Brasília: Universidade de Brasília, 2007. ISBN 978-85-230-0990-8. Citado na página 23.

MEDEIROS, K. R. d. et al. Lei de responsabilidade fiscal e as despesas com pessoal da saúde: uma análise da condição dos municípios brasileiros no período de 2004 a 2009. *Ciência & Saúde Coletiva*, scielo, v. 22, p. 1759 – 1769, 06 2017. ISSN 1413-8123. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-81232017002601759&nrm=iso>. Citado na página 14.

MORAIS, A. R.; FERES, M. V. C. Estudo de casos múltiplos e o sistema jurídico de patentes: a assimetria entre as doenças negligenciadas esquistossomose e dengue. *Direito, Processo e Cidadania*, v. 3, n. 3, p. 45–77, 2024. Citado na página 3.

MORENS, D. M.; FOLKERS, G. K.; FAUCI, A. S. The challenge of emerging and re-emerging infectious diseases. *Nature*, Nature Publishing Group, v. 430, n. 6996, p. 242–249, 2004. Citado 2 vezes nas páginas 1 e 16.

MOTA, E.; TEIXEIRA, M. G. *Vigilância epidemiológica e a pandemia da Covid-19 no Brasil. Elementos para entender a resposta brasileira e a explosão de casos e mortes*. SciELO Preprints, 2020. Disponível em: <https://doi.org/10.1590/scielopreprints.1317>. Citado na página 19.

NEAPOLITAN, R. *Learning Bayesian Networks*. [s.n.], 2004. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=09ece45700e177091da4620169345ca7>. Citado na página 32.

NETO, N. W. et al. A pandemia da covid-19 impactou o enem? uma análise comparativa de dados dos anos de 2019 e 2020. *RENOTE*, v. 20, n. 1, p. 223–232, 2022. Citado na página 48.

OBAMA, B. *Transparency and Open Government*. 2009. Washington, United States of America: The White House. Citado na página 14.

OLIVEIRA, E. *Educação. Abstenção do Enem 2020 é de 55,3%; pedido de reaplicação deve ser feito a partir desta segunda*. G1 Educação, 2021. Disponível em: <https://g1.globo.com/educacao/enem/2020/noticia/2021/01/24/abstencao-do-enem-2020-e-de-553percent-24-milhoes-foram-aos-locais-de-prova-neste-domingo.ghtml>. Citado na página 57.

OLIVEIRA, J. B. A. e.; GOMES, M.; BARCELLOS, T. A covid-19 e a volta às aulas: ouvindo as evidências. *Ensaio: Avaliação e Políticas Públicas em Educação*, Fundação CESGRANRIO, v. 28, n. Ensaio: aval.pol.públ.Educ., 2020 28(108), Jul 2020. ISSN 0104-4036. Disponível em: <https://doi.org/10.1590/S0104-40362020002802885>. Citado 2 vezes nas páginas 60 e 63.

OPAS. *Organização Pan-Americana de Saúde*. 2021. <https://iris.paho.org/bitstream/handle/10665.2/54449/OPASWBRAPHECOVID-19210043_por.pdf?sequence=1&isAllowed=y>. Online; accessed 07 July 2022. Citado na página 18.

(OPAS), O. P.-A. da S. *Atualização Epidemiológica: Dengue, Chikungunya e Zika*. 2023. Acesso em: 11 abr. 2025. Disponível em: <https://www.paho.org/sites/default/files/2023-01/2023janpheactualizacaoarboviruspor_0.pdf>. Citado na página 87.

OPEN DATA COMMONS. *Open Data Commons*. 2019. Disponível em: <https://opendatacommons.org/licenses/>. Acessado em: 10 de Novembro 2024. Citado na página 14.

PATANÉ, J. et al. Sars-cov-2 delta variant of concern in brazil - multiple introductions, communitary transmission, and early signs of local evolution. *medRxiv*, Cold Spring Harbor Laboratory Press, 2021. Disponível em: <https://www.medrxiv.org/content/early/2021/10/12/2021.09.15.21262846>. Citado na página 19.

PEARL, J. *Causality: Models, Reasoning and Inference*. 2nd. ed. [S.l.]: Cambridge University Press, 2009. Citado 2 vezes nas páginas 28 e 67.

Philip Chen, C.; ZHANG, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on big data. *Information Sciences*, v. 275, p. 314–347, 2014. ISSN 0020-0255. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0020025514000346>. Citado na página 24.

PIATETSKY, G. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. 2014. Https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html. Citado na página 40.

PILAR, A. F.; CASTRO, R. *(CCS/Fiocruz). Boletim destaca marco de 500 mil mortes por Covid-19 no Brasil*. 2020. <https://portal.fiocruz.br/noticia/boletim-destaca-marco-de-500-mil-mortes-por-covid-19-no-brasil/>. Online; accessed 07 July 2024. Citado na página 17.

PRAMIYANTI, A. et al. Public perception on transparency and trust in government information released during the covid-19 pandemic. *Asian Journal for Public Opinion Research*, Center for Asian Public Opinion Research & Collaboration Initiative, v. 8, n. 3, p. 351–376, 2020. Citado na página 2.

QEDU. *Análise de Dados do ENEM 2018–2023: Resultados Preliminares*. 2023. Disponível em: <https://conteudos.qedu.org.br/analise-de-dados-enem-2018-2023-resultados-preliminares>. Citado na página 68.

RAAMKUMAR, A. S.; TAN, S. G.; WEE, H. L. Measuring the outreach efforts of public health authorities and the public response on facebook during the covid-19 pandemic in early 2020: Cross-country comparison. *J Med Internet Res*, v. 22, n. 5, p. e19334, May 2020. ISSN 1438-8871. Disponível em: <http://www.jmir.org/2020/5/e19334/>. Citado na página 4.

ROCHA, A. M.; CASTRO, A. F. de; OLIVEIRA, A. G. de. Análise de dados: perfil e desempenho dos participantes das edições do enem 2019 a 2022 sob a perspectiva da covid-19. *Revista Caribeña de Ciencias Sociales*, v. 12, n. 7, p. 3100–3120, 2023. Disponível em: <https://doi.org/10.55905/rcssv12n7-008>. Citado na página 69.

ROCHA, F. B. N. da et al. Análise dos microdados de matemática do enem de 2017–2019 do nordeste. *Research, Society and Development*, v. 11, n. 10, p. e207111032716–e207111032716, 2022. Citado na página 48.

RUSSELL, S.; NORVIG, P. *Artificial Intelligence: A Modern Approach (4th Edition) (Pearson Series in Artifical Intelligence)*. 4. ed. Language: English, 2020. ISBN 0134610997,9780134610993. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=B6BD185942998C39D7E87ABDD927E580>. Citado na página 30.

SCHMIDT, B. et al. Saúde mental e atenção psicossocial a grupos populacionais vulneráveis por processos de exclusão social na pandemia de covid-19. In: . [S.l.]: Fiocruz, 2021. p. 87–97. Citado na página 19.

SCUTARI, M.; DENIS, J.-B. *Bayesian Networks: With Examples in R*. [S.l.]: Chapman and Hall/CRC, 2018. Citado na página 6.

Secretaria de Estado de Saúde Pública do Pará (SESPA). *Boletim Epidemiológico COVID-19: Estado do Pará - Atualização semanal*. 2021. <http://www.saude.pa.gov.br/covid-19/boletins/>. Acesso em: 28 abr. 2025. Citado na página 2.

SILVA, C. M. L. D. *Estatística Descritiva - Manual de Auto-Aprendizagem*. 3. ed. [S.l.]: Edições Sílabo, 2018. ISBN 9726189683. Citado na página 23.

SILVA, C. M. L. d. *Estatística Descritiva - Manual de Auto-Aprendizagem*. 3. ed. [S.l.]: Sílabo, 2018. ISBN 9726189683. Citado na página 45.

SILVA, L. J. d. O controle das endemias no Brasil e sua história. *Ciência e Cultura*, scielocec, v. 55, p. 44–47, 01 2003. ISSN 0009-6725. Disponível em: <http://cienciaecultura.bvs.br/scielo.php?script=sci_arttext&pid=S0009-67252003000100026&nrm=iso>. Citado na página 15.

SILVA, P. H. I. O mundo do trabalho e a pandemia de covid-19: Um olhar sobre o setor informal. *Caderno de Administração*, v. 28, p. 66–70, jun. 2020. Disponível em: <https://periodicos.uem.br/ojs/index.php/CadAdm/article/view/53586>. Citado na página 80.

SILVEIRA, M.; MARCOLIN, C.; FREITAS, H. Uso corporativo do big data: Uma revisão de literatura. *Revista de Gestão e Projetos*, v. 6, n. 3, p. 44–59, 2015. ISSN 2236-0972. Disponível em: <https://periodicos.uninove.br/gep/article/view/9627>. Citado na página 24.

SOLARTE, Y.; PEÑA, M.; MADERA, C. Transmisión de protozoarios patógenos a través del agua para consumo humano. *Colombia médica*, Universidad del Valle, v. 37, n. 1, p. 74–82, 2006. Citado na página 1.

SOTT, M. K.; BENDER, M. S.; BAUM, K. da S. Covid-19 outbreak in brazil: Health, social, political, and economic implications. *International Journal of Health Services*, SAGE Publications Sage CA: Los Angeles, CA, v. 52, n. 4, p. 442–454, 2022. Citado na página 2.

SOUSA, R. C. *Vulnerabilidade, vida precária e luto: os impactos da pandemia da COVID-19 no Brasil*. [S.l.]: UNIFESSPA, 2020. Online; accessed 07 July 2022. Citado na página 17.

SOUZA, L. C. d. et al. Sars-cov, mers-cov e sars-cov-2: a narrative review of the main coronaviruses of the century. *Brazilian Journal of Health Review*, v. 4, n. 1, p. 1419–1439, 2021. ISSN 2595-6825. Citado na página 16.

SOUZA, M. P. R. d. Qualidade de dados dentro do contexto de big data: uma revisão global. 2023. Citado 2 vezes nas páginas 24 e 25.

SU, S. et al. Epidemiology, genetic recombination, and pathogenesis of coronaviruses. *Trends in microbiology*, v. 24, n. 6, p. 490–502, 2016. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/27012512/>. Citado na página 16.

SUS. *MANUAL DE ORIENTAÇÕES DA COVID-19 (vírus SARS-CoV-2). Este manual consolida e revoga as orientações técnicas: Nota Técnica Conjunta nº 002/2020 – COSEMS/SUV/SPS/SES/SC – COE; Nota Técnica Nº. 003/2020 – DIVE/SUV/SES/SC; NOTA INFORMATIVA CONJUNTA Nº. 001/2020 – SUV/DIVE/LACEN/SES/SC – COE;Nota Informativa nº. 002/2020 –DIVE/SUV/SES/SC; Nota Informativa Conjunta nº. 003/2020 – DIVE/LACEN/SUV/SES/SC e Nota Técnica nº 003/2020 SES/SUV/SC – COE*. [S.l.]: GOVERNO DE SANTA CATARINA, 2020. Citado na página 20.

TAN, P.-N.; STEINBACH, M.; KUMAR, V. *Introdução ao Data Mining. Mineração de Dados*. 1. ed. [S.l.]: Ciência Moderna, 2009. ISBN 978-8573937619. Citado 3 vezes nas páginas 30, 41 e 42.

TAUBENBERGER, J. K.; MORENS, D. M. 1918 influenza: The mother of all pandemics. *Emerging Infectious Diseases*, v. 12, n. 1, p. 15–22, 2006. Citado na página 16.

TORRE-LUQUE, A. de la et al. Suicide mortality in spain during the covid-19 pandemic: Longitudinal analysis of sociodemographic factors. *European Neuropsychopharmacology*, v. 82, p. 29–34, 2024. Citado na página 9.

UDDIN, S. et al. How did socio-demographic status and personal attributes influence compliance to covid-19 preventive behaviours during the early outbreak in japan? lessons for pandemic management. *Elsevier Ltd*, Pers Individ Dif, v. 175, Jun 2021. ISSN 1679-4974. Disponível em: <https://pubmed.ncbi.nlm.nih.gov/33639446/>. Citado 2 vezes nas páginas 1 e 16.

(UFOP), L. L. *ENEM 2020: A Prova das Desigualdades*. 2021. Disponível em: <https://sites.ufop.br/lamparina/blog/enem-2020-prova-das-desigualdades>. Citado na página 68.

UJVARI, S. C. *História das epidemias*. [S.l.]: Editora Contexto, 2020. Citado na página 1.

ULAK, M. B.; YAZICI, A.; ZHANG, Y. Analyzing network-wide patterns of rail transit delays using bayesian network learning. *Transportation Research Part C: Emerging Technologies*, v. 119, p. 102749, 2020. ISSN 0968-090X. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0968090X20306628>. Citado na página 28.

UNDP, U. N. D. P. *Relatório de Desenvolvimento Humano 2021-22*. 2023. Acesso em: 11 abr. 2025. Disponível em: <https://www.undp.org/pt/brazil/desenvolvimento-humano/publications/relatorio-de-desenvolvimento-humano-2021-22>. Citado 2 vezes nas páginas 87 e 88.

VAN ROSSUM, G.; DRAKE JR, F. L. *Python reference manual*. [S.l.]: Centrum voor Wiskunde en Informatica Amsterdam, 1995. Citado na página 38.

VANDERPLAS, J. *Python Data Science Handbook: Essential Tools for Working with Data*. [S.l.]: O'Reilly Media, Inc., 2016. ISBN 978-1-491-91205-8. Citado na página 26.

WALLACE, R. *Pandemia e agronegócio: doenças infecciosas, capitalismo e ciência*. [S.l.]: Editora Elefante, 2020. Citado na página 1.

World Health Organization. *WHO Coronavirus (COVID-19) Dashboard*. 2023. urlhttps://covid19.who.int/. Accessed: 2025-04-28. Citado na página 1.

YAZEEDI, B. M. A. et al. Sociodemographic and health determinants of lifestyle changes during the covid-19 pandemic in oman. *Heliyon*, v. 10, n. 6, p. e40358, 2024. Citado na página 9.

YOUNGERMAN, S. F. B. *Pandemics and Global Health*. Facts on File, 2008. (Global Issues). ISBN 9780816070206,0816070202. Disponível em: <http://gen.lib.rus.ec/book/index.php?md5=5730391e2dbcc2b020828a7ae1208584>. Citado 2 vezes nas páginas 15 e 21.

YU, L.; ZHOU, N. *Survey of Imbalanced Data Methodologies*. arXiv, 2021. Disponível em: <https://arxiv.org/abs/2104.02240>. Citado na página 24.

ZHUO, J.; HARRIGAN, N. Low education predicts large increase in covid-19 mortality: The role of collective culture and individual literacy. *Public Health*, v. 221, p. 201–207, 2023. Citado na página 11.

Anexos

ANEXO A – Publicação

frontiers | Frontiers in Big Data

TYPE Original Research
PUBLISHED 12 March 2025
DOI 10.3389/fdata.2025.1485493

OPEN ACCESS

EDITED BY
Immanuel Azaad Moonesar,
Mohammed Bin Rashid School of Government, United Arab Emirates

REVIEWED BY
Karthikeyan Umapathy,
University of North Florida, United States
Gustavo Cunha de Araujo,
Federal University of North Tocantins (UFNT), Brazil

*CORRESPONDENCE
Sandio Maciel Dos Santos
✉ sandio.maciel@gmail.com

RECEIVED 23 August 2024
ACCEPTED 20 February 2025
PUBLISHED 12 March 2025

CITATION
Santos SMD, Silva MSd, França Lobato FM and Francês CRL (2025) Use of Bayesian networks in Brazil high school educational database: analysis of the impact of COVID-19 on ENEM in Pará between 2019 and 2022.
Front. Big Data 8:1485493.
doi: 10.3389/fdata.2025.1485493

Use of Bayesian networks in Brazil high school educational database: analysis of the impact of COVID-19 on ENEM in Pará between 2019 and 2022

Sandio Maciel Dos Santos[1]*, Marcelino Silva da Silva[2], Fábio Manoel França Lobato[3] and Carlos Renato Lisboa Francês[4]

[1]Graduate Program in Electrical Engineering, Federal University of Pará, Belém, Brazil, [2]Graduate Program in Electrical Engineering, Institute of Engineering and Geosciences, Federal University of Western Pará, Santarém, Pará, Brazil, [3]Institute of Engineering and Geosciences, Federal University of Western Pará, Santarém, Pará, Brazil, [4]Graduate Program in Electrical Engineering, Institute of Technology, Federal University of Pará, Belém, Brazil

This study examines the impact of the COVID-19 pandemic on academic performance and student participation in the National High School Exam (ENEM) in the state of Pará, Brazil, focusing on the interaction between socioeconomic factors, access to technology, and regional disparities. The research employed a mixed-methods approach, analyzing quantitative data from ENEM results (2020–2022) and qualitative interviews with educators and students. The findings indicate that the pandemic exacerbated pre-existing educational inequalities, particularly affecting low-income students and those enrolled in public schools. The highest dropout rates were recorded among students with a family income of up to one minimum wage, highlighting the barriers posed by limited access to technology and infrastructure for remote learning. A statistical analysis revealed a 20% increase in scores among students with access to computers and the Internet, particularly in private schools. The study also found significant regional differences across Pará's mesoregions, with Marajó and Southeast Pará facing more persistent challenges in reducing dropout rates compared to the Metropolitan Region of Belém. These results underscore the urgent need for region-specific public policies that address disparities in educational resources, including targeted investments in digital infrastructure and teacher training for remote education. The study concludes that comprehensive support programs, including psychological assistance for students, are essential for building a more resilient and equitable educational system capable of withstanding future crises.

KEYWORDS

COVID-19, ENEM, educational inequality, remote learning, regional disparities

Introduction

The COVID-19 pandemic brought transformative changes across various sectors, including healthcare, education, and urban infrastructure. Its impact exposed and exacerbated preexisting social inequalities, shaping how different groups navigated the challenges posed by the crisis. In education, the shift from in-person to remote learning introduced significant difficulties, including limited access to technology and the internet, increased stress levels, and disruptions to the learning environment—challenges that were particularly pronounced among low-income students in developing countries such as Brazil.

The pandemic most affected the education sector, leading to an abrupt transition from in-person to remote learning. Students worldwide face significant challenges, including limited access to technology and the internet, difficulties concentrating, and increased stress due to changes in the learning environment (Alqahtani and Rajkhan, 2020; Zhu et al., 2022). In countries like Brazil, these challenges were exacerbated by socioeconomic inequalities that directly impacted how students accessed and benefited from remote learning (Ferreira et al., 2022; Silva and Ribeiro-Alves, 2021).

Recent studies have demonstrated that the shift to remote learning, particularly in developing countries, led to significant learning losses, disproportionately affecting low-income students who lacked adequate technological resources to engage in virtual classes (Van Lancker and Parolin, 2020). Other studies have emphasized the crucial role of educational policies implemented during the pandemic in mitigating these impacts, highlighting that interventions providing access to technology and psychological support are essential for reducing educational inequalities (Bartholo et al., 2023).

This study aims to examine the impact of the COVID-19 pandemic on the academic performance of high school students in Pará, focusing on an analysis of microdata from the National High School Exam (ENEM) from 2019 to 2022. While the pandemic introduced new challenges, including prolonged school closures and remote learning, this study seeks to identify the key social factors that influenced student performance during this critical period.

The analysis examines the correlation between factors such as household income, parental educational level, and access to technological resources with academic performance, aiming to understand how these elements contributed to variations in ENEM scores before and during the pandemic.

References to infection rates are maintained in this study to illustrate how infection peaks and social restriction measures, such as lockdowns, affected the learning environment, particularly in regions with high levels of social inequality. Previous studies indicate that these restrictions disproportionately impacted students from low-income families, who had limited access to educational resources (Hawkins et al., 2020; Park and Awan, 2023). Therefore, understanding the relationship between infection rates and the educational policies implemented during these periods is crucial for contextualizing the challenges students faced throughout the pandemic.

Studies conducted in other countries, such as Nigeria and China, have shown that social factors, including family composition and income, directly influence academic performance in remote learning contexts (Ariyo et al., 2022; Zhu et al., 2022). In Brazil, the pandemic highlighted regional and socioeconomic inequalities, which were reflected in students' performance on national exams such as the ENEM (Weber Neto et al., 2022; Gonçalves and Pereira, 2024).

The study by Livingston et al. (2022) reveals that the COVID-19 pandemic exposed inequalities in digital access to education, with the lack of adequate infrastructure hindering remote learning in various regions. The research emphasizes the urgent need for investments in digital inclusion to address these disparities, a challenge that is equally relevant for Brazil and its diverse regions. This study contributes to the literature by examining how these factors specifically manifested in Pará, a region with unique socioeconomic characteristics within the Amazonian context.

Methods

The methodology employed in this study involves the application of data science techniques, specifically Educational Data Mining (Filatro, 2021; Mouromtsev and d'Aquin, 2016), as the primary approach for knowledge extraction from databases, utilizing the gathered information to support decision-making processes. The analysis focuses on educational data from high school students and graduates to investigate the impacts of the COVID-19 pandemic.

For this study, datasets from the ENEM exams for the years 2019 (pre-pandemic period) and 2020–2022 (pandemic period) were selected. These years were chosen due to the significant increase in COVID-19 infections, alongside the corresponding school censuses for the same periods, which serve as sources of microdata for ENEM. This selection allows for an examination of student performance amid the challenges posed by the pandemic, particularly in the context of national exam responses, with the aim of determining the influence of school closures during periods of high epidemic risk (Pereira Junior et al., 2021; Karakose, 2021; Reimers, 2022).

The ENEM microdata for 2019 and 2022 consists of datasets of 2.24, 1.88, and 1.40 gigabytes, respectively, each containing a set of 76 variables. Together, these datasets represent over 14 million instances, corresponding to the number of exam participants nationwide. Among the 76 analyzed variables, 22 were selected based on their stronger correlation with performance scores, as presented in Table 1. This selection was made to optimize the construction of the representative Bayesian Network (BN) for the problem at hand (Murphy and Russell, 2002).

Unlike previous studies that relied solely on average scores as a performance criterion (Boneti and de Oliveira, 2017; Ferrari Bravin et al., 2019; Vinicios do Carmo et al., 2021; da Silveira et al., 2015), this study adopts a more comprehensive approach. Bayesian Networks were selected for their ability to model complex probabilistic relationships and incorporate latent variables that may influence student performance. While traditional metrics, such as Pearson or Spearman correlations, are useful for measuring linear and monotonic associations between variables, Bayesian Networks provide a more flexible approach for identifying non-linear dependencies and causal inferences, facilitating a more detailed analysis of interactions between sociodemographic variables and academic performance.

Data preprocessing

The data were cleaned to remove inconsistencies and fill in missing values. Categorical variables were encoded, and continuous variables were normalized to facilitate analysis.

TABLE 1 Socioeconomic and academic variables in educational data analysis.

Parameter	Variables	Description
Sex	TP_SEXO	Sex
Color and race	TP_COR_RACA	Color and race
Dependence_ADM	TP_DEPENDENCIA_ADM_ESC	Administrative dependency
Abstention	TP_PRESENCA_CN	AE natural sciences
Abstention	TP_PRESENCA_CH	AE human sciences
Abstention	TP_PRESENCA_CL	AE code and language
Abstention	TP_PRESENCA_MT	AE mathematics and technology
Performance	NU_NOTA_CN	Score in natural sciences
Performance	NU_NOTA_CH	Score in human sciences
Performance	NU_NOTA_CL	Score in code and language
Performance	NU_NOTA_MT	Score in math and technology
Father's education	Q001	Father's level of education
Mother's education	Q002	Mother's level of education
Father's occupation	Q003	Father's occupation
Mother's occupation	Q004	Mother's occupation
People household	Q005	Nº. of people in the household
Family income	Q006	Average family income
Bathrooms	Q008	Nº of bathrooms
Bedrooms	Q009	Nº of bedrooms
Computer	Q024	Computer
Internet	Q025	Internet

AE, attendance in the exam, **Nº**, **number**.

Performance stratification

The study categorizes performance using quartiles, calculated based on the minimum and maximum score values in each knowledge area (Bendikson et al., 2011; Waheed et al., 2019), while also considering the number of dropouts per exam edition ξ. As shown in Equation 1:

$$K_Q \frac{P(\vec{u}+1)}{4} \quad (1)$$

To calculate the position of the $K_{Q-\text{th}}$ quartile in an ordered dataset, where:

- P represents the percentile (in the case of quartiles, P ranges from 1 to 3, corresponding to the first, second, and third quartiles).
- $\vec{u}$ is the total number of observations.

Table 2 illustrates the discretization into three groups using the quartile method. The $K_{Q \leq 25\%}$ group represents students with performance below 25%, $_{25\%} < K_{Q < 75\%}$ includes those with scores between 26 and 74%, and $K_{Q \geq 75\%}$ encompasses students with performance above 75%. The variable ξ refers to the number of dropouts per exam edition. This categorization is essential for understanding the real impacts of COVID-19 on sociodemographic dimensions and its influence on student performance during educational disruptions.

TABLE 2 Distribution of ENEM participants by socioeconomic parameters and dropout rate (2019–2022).

Edition	ξ	$K_{Q \leq 25\%}$	$25\% < K_Q < 75\%$	$K_{Q \geq 75\%}$
2019	Inf%	–inf–443	444–546	557–inf+
2020	Inf%	–inf–439	440–544	545–inf+
2021	Inf%	–inf–443	444–572	573–inf+
2022	Inf%	–inf–484	484–602	603–inf+

inf%, absent participants; –inf, scores below 25%; inf+, scores above 75%.

Modeling with Bayesian networks

Bayesian Networks were constructed using the PGMPY library (Ankan and Panda, 2015), chosen for its ease of configuration and usability, as well as its intuitive generation of probabilistic relationships and display of Conditional Probability Tables (CPTs) for each node. Visualization was facilitated by the pyAgrum API (Ducamp et al., 2020).

TABLE 3 Score comparison for Bayesian network structures.

Edition	K2Score	BicScore	BdeuScore
2019	1.32×10^7	4.99×10^5	-4.98×10^5
2020	7.34×10^6	-3.68×10^5	-3.66×10^5
2021	1.06×10^7	-3.14×10^5	-3.11×10^5
2022	1.54×10^7	-3.95×10^5	-3.94×10^5

Selected variables

The variables representing scores in different knowledge areas were grouped into four performance analysis groups, as described in Table 2. For the Monthly Household Income variable (Q006), which consists of income ranges (e.g., "from R$0.00 to R$998.00"), the lowest salary and the number of people per household (Q005) were used to replace the original text and group them according to the ENEM variable dictionary (Brasil, 2022).

The data used in this study were obtained from the public ENEM microdata and are available for consultation through the microdata[1] repository. This allows other researchers to replicate the analysis, promoting transparency and validation of results.

While Bayesian Networks offer a significant advantage in capturing complex relationships, they have inherent limitations, such as the requirement for conditional independence assumptions between variables when employing the Hill-Climb Search algorithm (Koller and Friedman, 2009). To mitigate these limitations, a structure validation analysis was performed using scoring metrics such as K2Score, BicScore, and BdeuScore to ensure the robustness and reliability of the results, as shown in Table 3 (Koller and Friedman, 2009). These metrics provide quantitative measures of the network's structural quality, balancing model fit and complexity.

- **K2Score**: Higher values indicate a better fit under the K2 metric, reflecting how well the structure aligns with the data.
- **BicScore** and **BdeuScore**: Negative values reflect penalization for model complexity, which helps prevent overfitting by discouraging overly complex structures that do not significantly improve the model's performance.

Table 3 compares the scores for different Bayesian Network structures across multiple editions, providing a quantitative basis for evaluating model robustness. Higher K2Score values indicate a better fit, while BicScore and BdeuScore values reflect the trade-offs between accuracy and simplicity. These metrics are instrumental in validating the network structure, ensuring that it captures underlying dependencies without overfitting or introducing unnecessary complexity.

However, the absence of a detailed discussion or interpretation of these scores limits the understanding of their implications for structure validation in Bayesian Networks. Future research should build on these findings by incorporating a comprehensive analysis of the scoring metrics and exploring their theoretical and practical impacts. Additionally, qualitative analyses or empirical validations should complement these results, offering further insights into the model's performance and applicability in real-world scenarios.

The statistical and probabilistic inferences drawn from the ENEM microdata and the School Census aim to compare the sociodemographic effects of successive epidemic outbreaks, confirmed cases, and deaths on student performance. This comprehensive approach seeks to identify those most likely to be affected when a public health alert is declared.

Results

The findings of this study reveal significant trends regarding the impact of the COVID-19 pandemic on academic performance and student participation in the Brazilian National High School Exam (ENEM) in the state of Pará, Brazil, from 2019 to 2022.

As shown in Table 4, participants with a household income below the minimum wage exhibited the highest dropout rates from ENEM in 2020 and 2021 compared to 2019. These data underscore the disproportionate impact of epidemic outbreaks, such as the COVID-19 pandemic, on low-income populations, where prolonged public institution closures directly hindered educational access for these groups (Dutra et al., 2023; Ferreira et al., 2022; Torres et al., 2020). This impact reflects a scenario where socioeconomic conditions restrict access to remote learning alternatives, particularly in more vulnerable regions.

The data presented in Table 4 reveal a concerning trend of increasing dropout rates among low-income participants during the years most impacted by the pandemic. This observation suggests that socioeconomic inequalities were exacerbated during this period, particularly for individuals reliant on public institutions who faced greater challenges in adapting to remote learning.

Further analysis of participants scoring above 75% shows that students attending or who had attended private schools during the pandemic performed better than their public school counterparts. These findings suggest that resource availability, such as access to computers and the internet, played a crucial role in academic success, especially during remote learning periods. Table 5 highlights a clear relationship between access to these resources and higher exam scores. For instance, private school participants with internet access exhibited an average performance increase of 20% compared to their peers in public schools.

The data suggest that access to technological resources significantly impacted academic performance during the pandemic. Students with home access to a computer and internet achieved higher scores, underscoring the importance of ensuring adequate infrastructure for remote learning, particularly during periods of school disruption.

The data also indicate that higher maternal employment and education levels correlated with improved student performance. This finding suggests that the home environment can substantially influence academic outcomes beyond direct access to material resources. Parental involvement and education provide additional support, either by fostering a more structured study environment or by promoting the value of continuous learning (Fernandes et al., 2023; Navarro et al., 2021).

1 https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem

TABLE 4 Dropout percentage of ENEM participants by socioeconomic parameter (2019–2022).

Parameters	ENEM edition							
	2019		2020		2021		2022	
ADM dependence	Public	15.45	Public	41.49	Public	26.78	Public	33.52
Color and race	Brown	69.16	Brown	66.84	Brown	64.25	Brown	64.71
Mother's education level	Elementary	39,85	Elementary	35.51	Elementary	30.60	Elementary	32.72
Father's education level	Elementary	45.44	Elementary	42.37	Elementary	38.53	Elementary	24.43
Mother's occupation	Group 2	46.07	Group 2	46.98	Group 2	47.11	Group 2	46.07
Father's occupation	Group 1	37.24	Group 1	32.31	Group 1	28.77	Group 1	31.21
Number of people	+4	45.10	+4	41.31	+4	41.07	+4	49.45
*Family income	[0-1] Salary	63.73	[0-1] Salary	70.00	[0-1] Salary	62.41	[0-1] Salary	67.36
Number of bathrooms	1	87.35	1	83.74	1	82.15	1	81.51
Number of rooms	2	51.39	2	51.89	1	82.15	2	51.02
Computer	No	87.80	No	84.10	No	80.88	No	83.90
*Internet	No	64.79	Yes	52.46	Yes	69.42	Yes	70.31

ADM, Administrative; Elementary, Elementary School; High, High School; and Group, Occupation details in the table.

TABLE 5 Academic performance of participants scoring above 75% in the ENEM by socioeconomic parameter (2019–2022).

Parameters	ENEM edition							
	2019		2020		2021		2022	
ADM dependence	Private	66.36	Private	52.05	Private	53.70	Private	43.05
Color and race	Brown	59.55	Brown	54.76	White	50.69	Whine	47.68
*Mother's education level	High	37.26	High	37.93	High	35.98	High	37.50
Father's education level	High	43.39	High	40.58	High	38.03	High	35.04
*Mother's occupation	Group 4	28.11	Group 4	37.91	Group 4	38.58	Group 4	48.47
Father's occupation	Group 4	38.79	Group 4	41.22	Group 4	43.61	Group 4	42.16
Number of people	4	35.58	4	39.40	4	40.20	4	39.85
Family income	[0-1] Salary	32.23	[0-1] Salary	20.20	[0-1] Salary	15.20	[0-1] Salary	19.64
Number of bathrooms	1	61.56	1	48.60	1	44.24	1	37.15
Number of rooms	2	48.85	2	44.84	2	43.67	2	43.53
*Computer	Yes	47.91	Yes	65.83	Yes	64.42	Yes	80.64
*Internet	No	70.78	Yes	87.04	Yes	71.64	Yes	97.91

ADM, Administrative; Elementary, Elementary School; High, High School; and Group, Occupation details in the table.

This initial analysis aimed to clarify the influence of social parameters on the ENEM performance of participants in Pará. A Bayesian probabilistic analysis was conducted to investigate how the rise in respiratory syndrome cases during the COVID-19 pandemic affected student performance. This analysis employed techniques such as Hill-Climb Search, K2 Score, and Variable Elimination, supported by the pyAgrum library, to visualize Bayesian Networks from 2019 to 2022, as illustrated in Figure 1.

The Bayesian networks derived from 2019 data underscore key variables significantly impacting ENEM participants' performance, including parental education level, family income, computer access, and the administrative status of the household, as illustrated in Figure 1A. This organizational structure defines a probabilistic dependency flow among selected parameters, establishing a solid foundation for performance analysis.

Applying the same methodology to structure Bayesian networks with educational data from 2021 and 2022 (Figures 1B, C) reveals a marked shift directly influenced by the COVID-19 pandemic: household computer presence no longer emerges as a primary variable of importance. This phenomenon is particularly relevant, considering that the 2020 ENEM occurred amidst substantial educational disruptions, with many students facing challenges in accessing the technology required for remote learning (Guia do Estudante, 2021; de Albuquerque, 2020).

The analysis of 2020 data, therefore, faces unique challenges, as the pandemic unpredictably altered relationships among variables

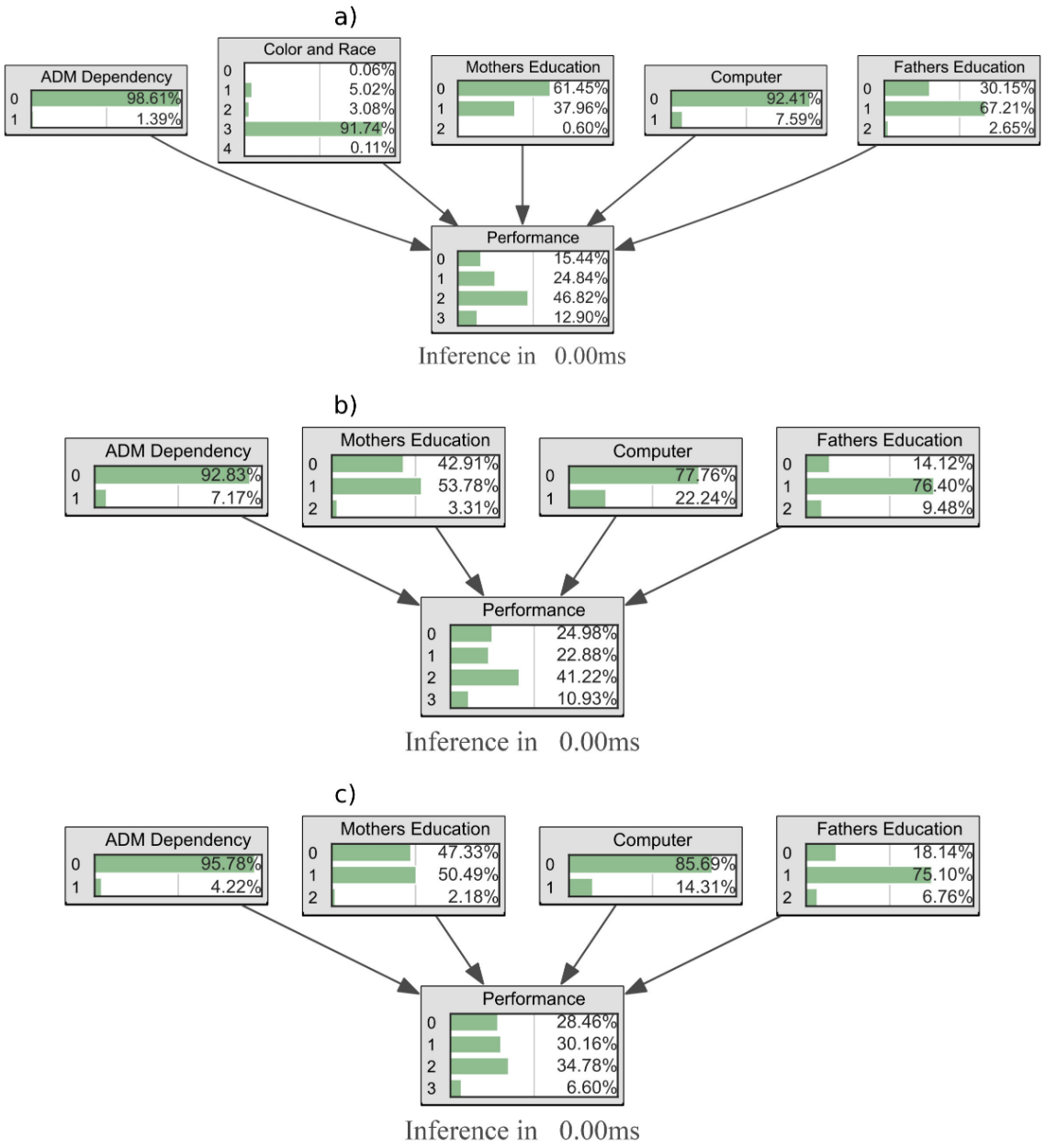


FIGURE 1
Bayesian network with ENEM Data from 2019 to 2022: Pre- and Post-COVID-19 analysis. **(A)** Bayesian inference for in 2019. **(B)** Bayesian inference for in 2021. **(C)** Bayesian inference for in 2022.

traditionally associated with academic performance. In a context of emergency remote learning and unequal access to resources, the data reflect atypical patterns, with socioeconomic variables such as family income and parental education, becoming even more unstable and less predictable.

Bayesian Networks (BNs) effectively model these complex interdependencies among educational and sociodemographic variables, allowing for causal inferences and identification of latent variables affecting student performance (Murphy and Russell, 2002). However, when dealing with 2020 ENEM data, BNs encounter limitations, as the pandemic's profound impact on low-income students led to record absenteeism and disparities in performance across different socioeconomic contexts (de Andrade and Bocardi, 2024; de Albuquerque, 2020).

This pandemic context highlights the need for critical evaluation of probabilistic models such as BNs. While robust, these networks depend on assumptions of conditional independence that may be compromised under extreme conditions, as imposed by the pandemic. Result interpretation thus requires caution, taking into account the limitations and potential biases within the data (Murphy and Russell, 2002).

Variable selection by the Bayesian network, which identifies the most relevant conditional dependencies, shows that higher levels of parental education correlate with better participant performance, as shown in Table 6 (Biener et al., 2019). However, ENEM dropout

TABLE 6 Relationship between father's education level and students' academic performance.

ENEM edition	Father's level of education			
	Group	Elementary school	High school	University education
2019	ξ	*20.68	13.58	*10.88
	$K_{Q \leq 25\%}$	30.55	23.05	19.25
	$_{25\%} < K_{Q} < _{75\%}$	41.95	49.55	47.39
	$K_{Q \leq 75\%}$	6.81	13.82	22.48*
2020	ξ	*51.36	38.01	29.60*
	$K_{Q \leq 25\%}$	21.83	19.01	13.24
	$_{25\%} < K_{Q} < _{75\%}$	24.05	35.35	36.36
	$K_{Q \geq 75\%}$	2.76	7.64	20.80*
2021	ξ	*34.73	25.30	*17.75
	$K_{Q \leq 25\%}$	29.34	24.08	19.20
	$_{25\%} < K_{Q} < _{75\%}$	32.11	41.82	40.86
	$K_{Q \geq 75\%}$	3.82	8.80	20.19*
2022	ξ	*39.76	27.90	*18.92
	$K_{Q \leq 25\%}$	35.51	32.06	25.72
	$_{25\%} < K_{Q} < _{75\%}$	23.20	34.92	39.88
	$K_{Q \geq 75\%}$	1.54	5.12	15.98*

rates (ξ) increased by 19% from pre- to post-pandemic periods for parents with only primary education and by 8% for those with higher education. During the 2020 pandemic, dropout rates were ∼31% for parents with primary education and 10% for those with higher education. By 2021, these rates decreased to around 14 and 9%, respectively, reflecting a slight recovery in educational conditions.

Beyond the general analyses, the study also explored regional variations within Pará, as illustrated in Figure 3. The Metropolitan Region of Belém and Northeast Pará managed to reduce dropout rates during the pandemic between 2020 and 2021, in contrast to other regions that maintained high dropout rates. This finding suggests possible differences in implementing remote educational support strategies and local infrastructure.

An important aspect to highlight is the conditional probability between administrative dependency and the availability of a computer in the household for educational purposes. The inferences reveal a significant correlation, especially among public school students with computer access, showing a strong association with their ENEM scores. Analyzing the scores of students classified in the $K_Q < 75\%$ group, there is a marked disparity between those with and without computer access, indicating a significant increase in performance for the former. Specifically, there was a 13% increase among private school students, as shown in Table 7.

Another crucial aspect to consider is the conditional probability between administrative dependence and the availability of a home computer for educational activities. Inferences indicate a

TABLE 7 Relationship between administrative dependency and families with computer access at home in the 2019 ENEM.

Group	Administrative dependency		
	Public	Private	Computer
ξ	16.09	10.79	None
$K_{Q \leq 25\%}$	25.92	17.26	None
$_{25\%} < K_{Q} < _{75\%}$	47.30	48.25	None
$K_{Q \geq 75\%}$	*10.68	23.70	None
ξ	7.21	3.51	At least one
$K_{Q \leq 25\%}$	11.34	5.29	At least one
$_{25\%} < K_{Q} < _{75\%}$	34.57	24.25	At least one
$K_{Q \geq 75\%}$	*46.87	66.95	At least one

significant correlation, particularly among public school students with computer access, showing notable improvements in ENEM scores compared to those without access. Among students in the $_{25\%} < K_Q < _{75\%}$ group, a considerable increase in scores is observed for those with computer access. Specifically, private school students showed a 20% increase, as detailed in Table 8.

Moreover, the analysis of family income reported by participants reveals a strong relationship between higher income levels (C6; *) and student scores, as illustrated in Table 8. Consistent with this inference, examining the pre-established family income brackets shows a decline in performance among students reporting incomes up to one minimum wage (C1). Among those scoring in the $K_{Q \geq 75\%}$ group, there was a notable reduction of ∼6.5% in the participants within this income bracket.

A more detailed analysis assessed the impact on performance by considering participants' administrative dependence and family income. It was observed that the proportion of public school students in the $K_{Q \geq 75\%}$ group decreased when associated with incomes up to one minimum wage. Conversely, the dropout rate increased by 30%. Figure 2 provides a visual representation of participant performance based on family income.

Figure 2 shows a notable increase in the number of participants from private schools in the $_{25\%} < K_Q < _{75\%}$ group between 2020 and 2021. This shift may be attributed to the challenges posed by remote learning during peak COVID-19 case numbers in Brazil. In contrast, most dropouts in the national exam occurred among public school students (Navarro et al., 2021).

A more specific analysis of educational data from the state of Pará, focusing on the relationship between its six mesoregions and the school census, clarifies whether the impact of the COVID-19 pandemic had uniform effects on dropout rates and the overall performance of participants, as shown in Figure 3.

Figure 3 suggests that regional differences played a crucial role in the impact of the pandemic on education. While some regions implemented strategies that helped mitigate dropout rates, others faced significant challenges, such as high dropout rates. Among the mesoregions of Pará presented in Figure 4, it stands out that only the Metropolitan Region of Belém and Northeast Pará significantly reduced ENEM dropout rates during the COVID-19 pandemic between 2020 and 2021. In

TABLE 8 Relationship between family income and student performance.

ENEM edition	Family income							
	2019		2020		2021		2022	
Group	C1	C6	C1	C6	C1	C6	C1	C6
ξ	16.22	6.39	42.32	22.53	27.84	12.85	30.96	15.79
$K_{Q \leq 25\%}$	22.51	10.41	20.35	8.07	26.49	9.50	33.96	14.71
$_{25\%} < K_{Q} < _{75\%}$	47.01	41.06	31.94	41.59	39.59	39.85	31.61	39.59
$K_{Q \geq 75\%}$	(+) 10.25	41.14 (*)	(+) 5.39	37.81 (*)	(+) 6.08	37.80 (*)	(+) 3.47	29.90 (*)

C1, Income from zero up to 1 Minimum Wage; C6, more than 3 Minimum Wages.

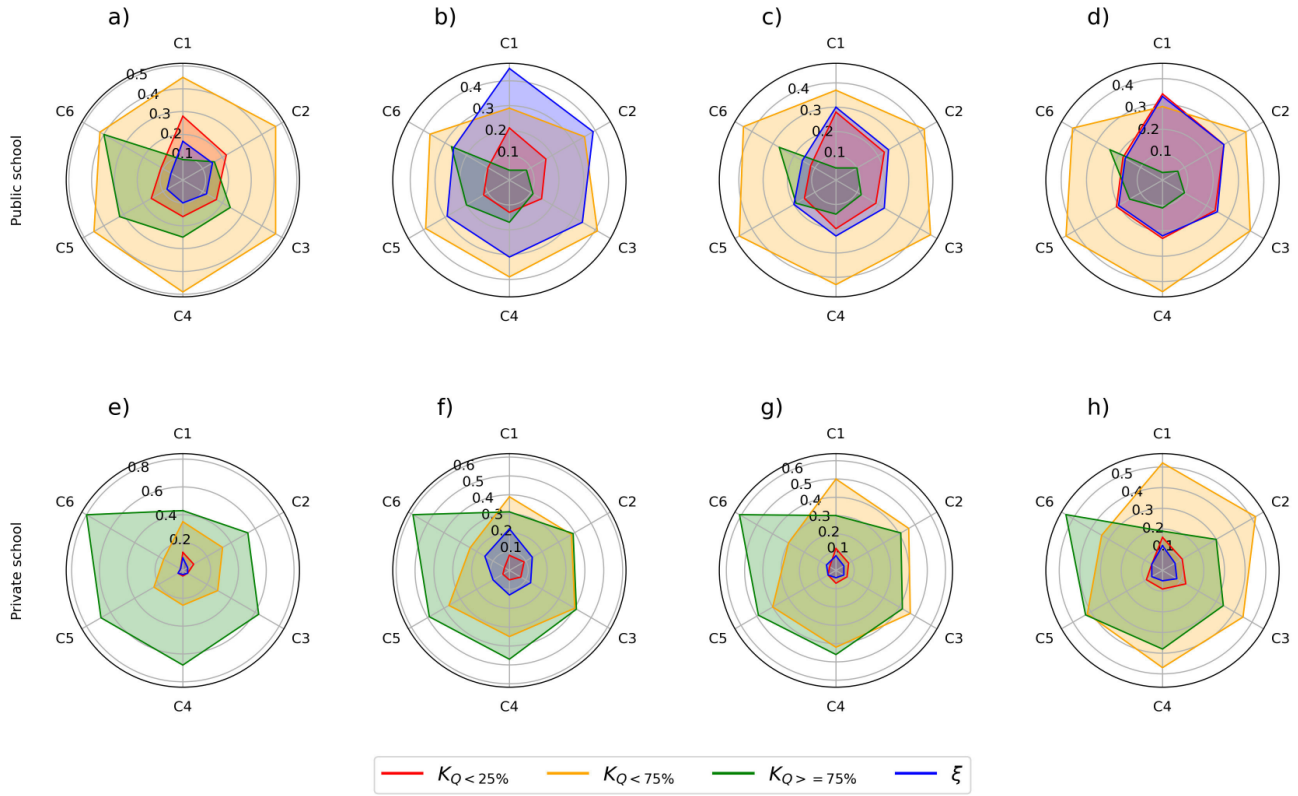


FIGURE 2
Performance radar of students through family income and administrative dependency. C1: Up to 1 minimum wage; C2: 1.5 minimum wages; C3: 2 minimum wages; C4: 2.5 minimum wages; C5: 3 minimum wages; C6: More than 3 minimum wages. Colors represent performance percentages: Blue: Dropouts; Orange: Scores between [0–25]; Green: Scores between [26–74]; Red: Scores between [75–100]. **(A)** Edition 2019. **(B)** Edition 2020. **(C)** Edition 2021. **(D)** Edition 2022. **(E)** Edition 2019. **(F)** Edition 2020. **(G)** Edition 2021. **(H)** Edition 2022.

contrast, the remaining regions maintained persistently high dropout rates, with percentages exceeding 20% during the same period.

The Marajó region was one of the most severely impacted after the onset of the COVID-19 pandemic. Notably, between 2020 and 2022, public school students exhibited a substantial decline in performance, with fewer than 10% achieving scores above 75% in assessments. Additionally, it is essential to highlight the significant increase in absenteeism among private school students during the ENEM. This trend may be related to mobility restrictions imposed by lockdowns and the closure of educational institutions on the island, as illustrated in Figure 4.

The findings suggest that the COVID-19 pandemic exacerbated existing socioeconomic inequalities, particularly concerning exam access and student performance. The forced transition to remote learning exposed structural weaknesses and highlighted the need for policies that ensure more equitable access to education, regardless of students' economic and regional conditions. Factors such as access to technological resources and the home environment proved to be decisive for academic success during this period.

The results of this study indicate that the COVID-19 pandemic significantly impacted students' participation and performance in the National High School Exam (ENEM), particularly in the more

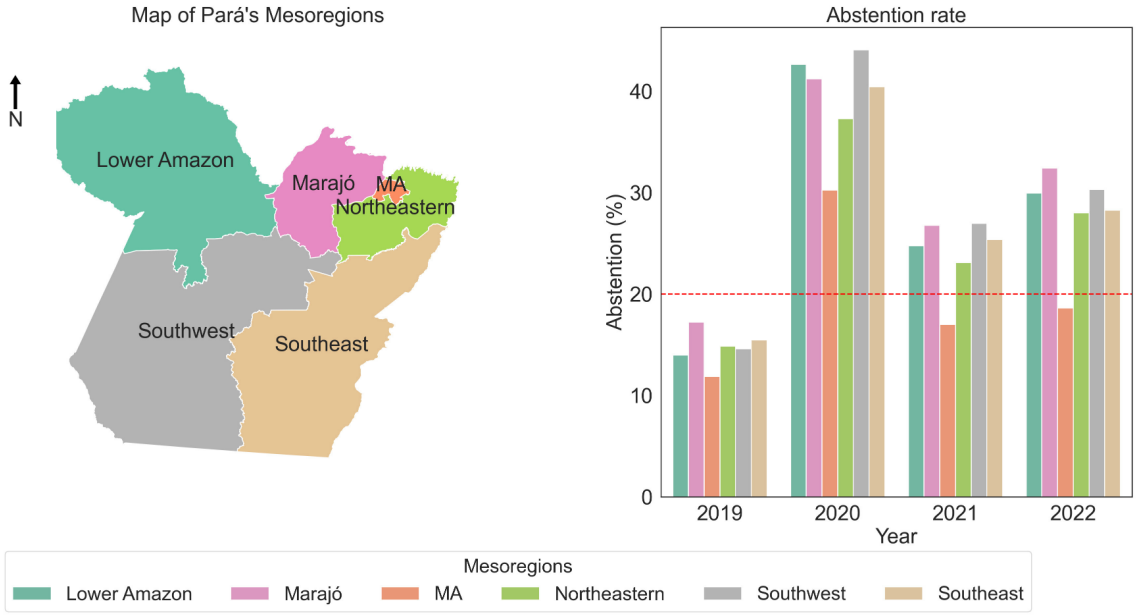


FIGURE 3
Abstention of students from the state of Pará by mesoregion.

vulnerable regions of the state of Pará, Brazil. Students from low-income families with limited access to technological resources were the most affected, exhibiting the highest dropout rates between 2020 and 2021. These findings highlight the exacerbation of socioeconomic inequalities during the pandemic, with the interruption of in-person classes and the difficulty of adapting to remote learning primarily hindering public school students from lower-income backgrounds.

Access to technological resources, such as computers and the internet, played a crucial role in academic performance. Students from private schools, who often had better access to these resources, showed superior performance compared to their peers in public schools. The analysis also underscored the importance of parental education and occupation, which, when higher, contributed to better academic outcomes for students, suggesting the significance of a more structured family environment.

Additionally, the Bayesian network analysis and regional variations in Pará indicated that the pandemic affected the state's different mesoregions unevenly. While the Metropolitan Region of Belém and the Northeast of Pará were able to reduce dropout rates, other areas, such as the Island of Marajó, faced greater challenges, showing significantly reduced performance and higher abandonment rates.

Limitations of the study

While the analysis provided valuable insights into the effects of the pandemic on academic performance, some limitations must be acknowledged. First, the use of Bayesian Networks, although effective in modeling probabilistic dependencies, relies on assumptions of conditional independence that may have been compromised in the emergency context of the pandemic. This could have led to distortions in the results, particularly when handling outlier data and variables influenced unpredictably by the pandemic. Additionally, the collection of data on socioeconomic and family factors may have been affected by incomplete information or access challenges during the period of restrictions. The analysis of regional variables also faces limitations, as the implementation of educational policies and local infrastructure in each mesoregion could have influenced the results unevenly.

In summary, while the findings provide a comprehensive view of the pandemic's impacts on the ENEM, future studies may need to address these limitations by expanding the analysis to include additional variables or more robust data collection methods, aiming to refine the models and provide a more detailed understanding of the factors influencing educational performance in times of crisis.

Discussion

The results reveal the profound and unequal impact of the COVID-19 pandemic on academic performance and student participation in the ENEM in the state of Pará. A detailed analysis of the different mesoregions and the relationship between socioeconomic factors and performance highlights several trends and challenges that should be considered for the future of education in the region.

The data showed that the pandemic exacerbated existing inequalities, especially among low-income students and those attending public schools. The highest dropout rates were observed among participants with a family income of up to one minimum wage, highlighting the difficulties faced by families unable to adapt to remote learning due to a lack of technological resources and adequate infrastructure. This trend was particularly evident in

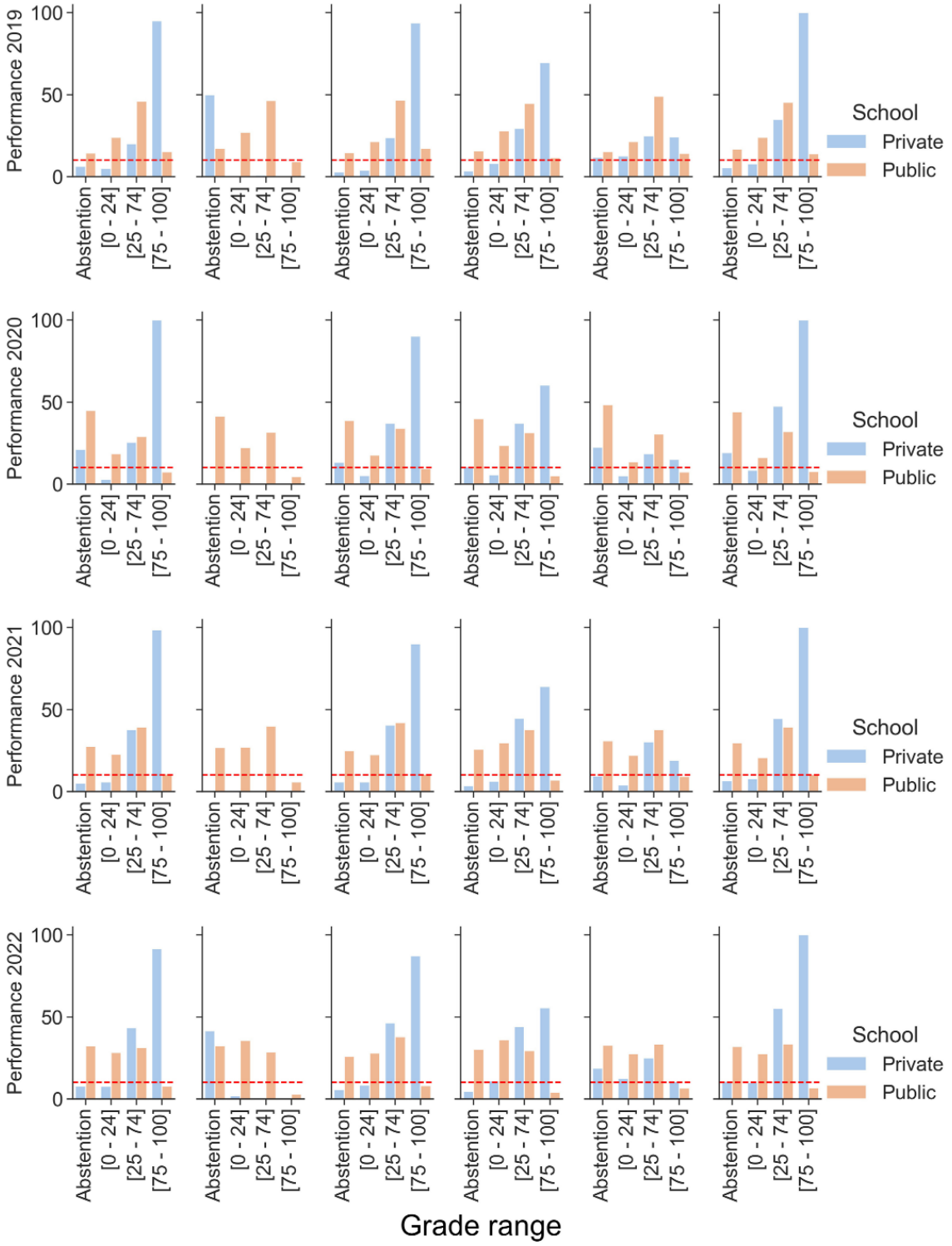


FIGURE 4
Comparison of student performance in Pará by Mesoregion in the ENEM (2019–2022).

Table 2, where low-income groups recorded the highest dropout rates during the peak pandemic (2020 and 2021). This scenario underscores the need for greater attention to inequality and health literacy issues, which are essential to support students' holistic development and education (de Oliveira et al., 2024).

This disparity reflects an urgent need for investments in digital infrastructure and educational support for low-income students. Public policies must ensure universal access to resources such as computers and the Internet to prevent economic inequalities from translating into disparities in educational opportunities.

As illustrated in Table 3, the analysis of academic performance revealed a strong correlation between access to technological resources and academic success during remote learning. Students with access to computers and the internet achieved significantly higher performance, with private school students registering a 20% increase in scores compared to their peers without these resources.

This finding highlights the importance of ensuring that all students, regardless of location or economic status, access tools that enable effective learning. Educational policies should prioritize the distribution of technological resources to minimize the impact of potential future school disruptions.

Regional analysis revealed significant disparities in the impact of the pandemic across the mesoregions of Pará. Figure 3 highlighted that while the Metropolitan Region of Belém and Northeast Pará managed to reduce dropout rates during the pandemic, other regions, such as Marajó and Southeast Pará, continued to face considerable challenges. These regions maintained high dropout rates, suggesting that factors such as local infrastructure, access to technology, and educational support were insufficient to ensure learning continuity.

According to Figure 4, fewer than 10% of public school students achieved scores above 75% between 2020 and 2022, while absenteeism in the ENEM significantly increased among private school students. This scenario may be explained by a combination of factors, including severe mobility restrictions imposed during lockdowns and the closure of educational institutions, which hindered students' access to exams and continuous learning.

This regional analysis demonstrates the need for a more specific, region-based approach to addressing educational inequalities. Support programs that consider each mesoregion's unique characteristics and challenges may be more effective than generic solutions, ensuring that more isolated and economically disadvantaged regions receive the necessary attention.

The results and discussions indicate the need for more inclusive and adaptive educational policies. The pandemic revealed that the educational system must be resilient and prepared to handle emergencies that may disrupt in-person learning. Investments in technology, teacher training for remote education, and programs for psychological and social support for students are essential to build a more robust and equitable educational system.

In summary, the analysis of ENEM data in Pará revealed not only the immediate impact of the COVID-19 pandemic on education but also systemic issues that need to be addressed moving forward. Economic inequalities, regional disparities, and limited resource access hinder educational equity. Public policies and private initiatives must work together to reduce these inequalities, ensuring that all students have equal opportunities for success, regardless of socioeconomic background or geographical location.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found at: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent from the participants legal guardian/next of kin was not required to participate in this study in accordance with the national legislation and the institutional requirements.

Author contributions

SS: Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing – original draft, Writing – review & editing. MS: Conceptualization, Funding acquisition, Resources, Supervision, Validation, Writing – original draft, Writing – review & editing. FF: Investigation, Supervision, Visualization, Writing – original draft, Writing – review & editing. CF: Funding acquisition, Project administration, Resources, Supervision, Visualization, Writing – original draft, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. To CNPq—National Council for Scientific and Technological Development and CAPES (Coordination for the Improvement of Higher Education Personnel), for funding my research through a scholarship.

Acknowledgments

Thanks to Hydro for the support and funding of this survey. Since 2019, the company has collaborated with UFPA in several initiatives through a technical and scientific cooperation agreement.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

References

Alqahtani, A. Y., and Rajkhan, A. A. (2020). E-learning critical success factors during the COVID-19 pandemic: a comprehensive analysis of e-learning managerial perspectives. *Educ. Sci.* 10:216. doi: 10.3390/educsci10090216

Ankan, A., and Panda, A. (2015). *pgmpy: Probabilistic Graphical Models using Python*. Em *Python in Science Conference*. Austin, Texas, 1–7. Available online at: https://conference.scipy.org/proceedings/scipy2015/ankur_ankan.html (accessed June 10, 2024).

Ariyo, E., Amurtiya, M., Lydia, O. Y., Oludare, A., Ololade, O., Taiwo, A. P., et al. (2022). Socio-demographic determinants of children home learning experiences during COVID 19 school closure. *Int. J. Educ. Res. Open* 3:100111. doi: 10.1016/j.ijedro.2021.100111

Bartholo, T. L., Koslinski, M. C., Tymms, P., and Castro, D. L. (2023). Learning loss and learning inequality during the Covid-19 pandemic. *Ensaio* 31:e0223776. doi: 10.1590/s0104-40362022003003776

Bendikson, L., Hattie, J., and Robinson, V. (2011). Identifying the comparative academic performance of secondary schools. *J. Educ. Adm.* 49, 433–449. doi: 10.1108/09578231111146498

Biener, C., Landmann, A., and Santana, M. I. (2019). Contract nonperformance risk and uncertainty in insurance markets. *J. Public Econ.* 175, 65–83. doi: 10.1016/j.jpubeco.2019.05.001

Boneti, L. W., and de Oliveira, G. M. (2017). Enem: analysis of school performance in the 2009-2013 editions. *Rev. Esp. Pedag.* 24, 371–386. doi: 10.5335/rep.v24i2.7420

Brasil (2022). *Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira | INEP*. Relatório de desempenho escolar 2023. Available online at: https://www.gov.br/inep/ (accessed August 8, 2024).

da Silveira, F. L., Barbosa, M. C. B., and da Silva, R. (2015). Exame nacional do ensino médio (ENEM): uma análise crítica. *Rev. Bras. Ens. Fís.* 37:1101. doi: 10.1590/S1806-11173710001

de Albuquerque, R. L. F. (2020). *ENEM durante a pandemia? Um estudo de caso das percepções de docentes da rede estadual de educação do Rio de Janeiro sobre a realização do ENEM 2020. Rev. Olhar Prof.* 23, 15649–209209225856. doi: 10.5212/OlharProfr.v.23.2020.15649.209209225856.0601

de Andrade, R. J., and Bocardi, J. M. B. (2024). Impacto da pandemia de Covid-19 nos resultados do enem do estado do paraná. *Rev. Gest. Aval. Educ.* 13:e86282. doi: 10.5902/2318133886282

de Oliveira, L. M. C., Zanin, L., and Flório, F. M. (2024). Professores do ensino fundamental público: literacia em saúde e fatores associados. *Rev. Contexto Educ.* 39:e13673. doi: 10.21527/2179-1309.2024.121.13673

Ducamp, G., Gonzales, C., and Wuillemin, P.-H. (2020). *aGrUM/pyAgrum: A toolbox to build models and algorithms for probabilistic graphical models in Python*. Em *Proceedings of the 10th International Conference on Probabilistic Graphical Models*. PMLR, 1–8. Available online at: https://proceedings.mlr.press/v138/ducamp20a.html (accessd June 10, 2024).

Dutra, J. F., Firmino Júnior, J. B., and de Souza Fernandes, D. Y. (2023). Fatores que podem interferir no desempenho de estudantes no ENEM: uma revisão sistemática da literatura. *Rev. Bras. Informát. Educ.* 31, 323–351. doi: 10.5753/rbie.2023.3087

Fernandes, L., Mendes, F., Alves da Silva, J., Silva, R., Damaceno, G., and Moura, E. (2023). Análise do desempenho em matemática e suas tecnologias dos participantes do ENEM 2021 em Barra do Corda, Maranhão: Uma comparação entre alunos de escolas públicas e privadas por meio de regressão logística. *Contrib. Cienc. Soc.* 16, 33822–33835. doi: 10.55905/revconv.16n.12-282

Ferrari Bravin, G., Lee, L., and das Dores Rissino, S. (2019). Mineração de dados educacionais na base de dados do enem 2015. *Braz. J. Prod. Eng.* 5, 186–201.

Ferreira, C. A. A., da Costa Lobato, T., and Carvalho, B. d. N. (2022). *ENEM no Norte do Brasil: Uma análise do desempenho e desafios educacionais*. Available online at: https://brsa.org.br/wp-content/uploads/wpcf7-submissions/7559/Artigo_ENEM-NO-NORTE-DO-BRASIL_-identificado.pdf (accessed August 8, 2024).

Filatro, A. (2021). *Data science na educação: Presencial, a distância e corporativa*. Saraiva Educação.

Gonçalves, D., and Pereira, L. (2024). Abandono escolar no ensino médio: uma análise comparativa antes e durante a pandemia em minas gerais. *J. Polít. Educ.* 18. doi: 10.5380/jpe.v18i1.92912

Guia do Estudante (2021). *Enem 2020 fracassa e evidencia desigualdades educacionais*. Available online at: https://guiadoestudante.abril.com.br/atualidades/enem-2020-fracassa-e-evidencia-desigualdades (accessed January 25, 2021).

Hawkins, R. B., Charles, E. J., and Mehaffey, J. H. (2020). Socio-economic status and COVID-19-related cases and fatalities. *Public health* 189, 129–134. doi: 10.1016/j.puhe.2020.09.016

Karakose, T. (2021). The impact of the COVID-19 epidemic on higher education: opportunities and implications for policy and practice. *Educ. Process Int. J.* 10, 7–12. doi: 10.22521/edupij.2021.101.1

Koller, D., and Friedman, N. (2009). *Probabilistic graphical models: Principles and techniques (1st ed.)*. Cambridge, MA: The MIT Press.

Livingston, E., Houston, E., Carradine, J., Fallon, B., Akmeemana, C., Nizam, M., and McNab, A. (2022). Global student perspectives on digital inclusion in education during COVID-19. *Glob. Stud. Childhood.* 13, 341–357. doi: 10.1177/20436106221102617

Mouromtsev, D., and d'Aquin, M. (eds.). (2016). *Open Data for Education: Linked, Shared, and Reusable Data for Teaching and Learning* (1ª ed.). Cham: Springer International Publishing.

Murphy, K. P., and Russell, S. J. (2002). "Dynamic Bayesian networks: Representation, inference, and learning," in *Proceedings of the 2002 Conference*. Available online at: https://api.semanticscholar.org/CorpusID:919497 (accessed August 8, 2024).

Navarro, D., Ianello, M., Muneratto, F., and Watanabe, G. (2021). Impacts of natural science knowledge on ENEM performance: considerations on scientific-technological inequality for social justice. *Rev. Bras. Pesq. Educ. Ciênc.* 21:e26002. doi: 10.28976/1984-2686rbpec2021u12171246

Park, A., and Awan, O. A. (2023). COVID-19 and virtual medical student education. *Acad. Radiol.* 30, 773–775. doi: 10.1016/j.acra.2022.04.011

Pereira Junior, L., Nasser Matos, S., and Bronoski Borges, H. (2021). Análise dos perfis de alunos do ensino superior sobre a realização de aulas na modalidade a distância durante pandemia da covid-19 usando algoritmos de aprendizagem de máquina. *Rev. Nov. Tecnol. Educ.* 18, 336–345. doi: 10.22456/1679-1916.110252

Reimers, F. M. (2022). "Learning from a pandemic. the impact of COVID-19 on education around the world" in *Primary and Secondary Education During Covid-19*, ed. F. M. Reimers (Springer, Cham).

Silva, J., and Ribeiro-Alves, M. (2021). Social inequalities and the pandemic of COVID-19: the case of Rio de Janeiro. *J. Epidemiol. Community Health.* 75, 975–979. doi: 10.1136/jech-2020-214724

Torres, R., de Pereira, M. M., Bender Filho, R., and Lisbinski, F. C. (2020). Determinantes do desempenho dos participantes da prova do enem: evidências para o rio grande do sul. *Desenv. Questão.* 18, 352–368. doi: 10.21527/2237-6453.2020.53.352-368

Van Lancker, W., and Parolin, Z. (2020). The impact of COVID-19 school closures on children's learning: a critical review of the literature. *Front. Educ.* 5, e243–e244. doi: 10.1016/S2468-2667(20)30084-0

Vinicios do Carmo, R., Felipe Heckler, W., and Varella de Carvalho, J. (2021). Uma análise do desempenho dos estudantes do rio grande do sul no ENEM 2019. *Rev. Nov. Tecnol. Educ.* 18, 378–387. doi: 10.22456/1679-1916.110257

Waheed, H., Hassan, S.-U., Aljohani, N. R., Hardman, J., and Nawaz, R. (2019). Predicting academic performance of students from VLE big data using deep learning models. *Comput. Hum. Behav.* 104:106189. doi: 10.1016/j.chb.2019.106189

Weber Neto, N. C., Soares, R., Reis Coutinho, L., and Soares Teles, A. (2022). A pandemia da COVID-19 impactou o ENEM? Uma análise comparativa de dados dos anos de 2019 e 2020. *Rev. Nov. Tecnol. Educ.* 20, 223–232. doi: 10.22456/1679-1916.126655

Zhu, W., Liu, Q., and Hong, X. (2022). Implementation and challenges of online education during the COVID-19 outbreak: a national survey of children and parents in China. *Early Child. Res. Q.* 61, 209–219. doi: 10.1016/j.ecresq.2022.07.004