

Análise Exploratória de Dados para Identificar o Impacto da Pandemia da COVID-19 no ENEM dos Estados do Ceará, Maranhão e Piauí

Nelson Weber Neto¹, Raimundo C. Soares¹, Luciano R. Coutinho¹, Ariel S. Teles^{1,2}

¹Universidade Federal do Maranhão (UFMA)

²Instituto Federal do Maranhão (IFMA)

nelson.weber@discente.ufma.br

Abstract. *In Brazil, the main exam for evaluating the educational performance of basic education is the Exame Nacional do Ensino Médio (ENEM), which is also used for entry of students into higher education. In 2020, with the arrival of the COVID-19 virus, basic education institutions needed to change their educational model with face-to-face teaching to the use of remote teaching methodologies. This work aims to identify the main impacts caused by the pandemic on ENEM in the states of Ceará, Maranhão and Piauí, in its first year. For this, a data exploratory analysis of the 2019 ENEM, the year before the beginning of the pandemic, and the 2020 ENEM, the first year of the pandemic, is carried out. Results show an improvement in the general performance of the states, an increase in the number of registrants who did not show up for the test, and an increase in inequality between participants from public and private institutions.*

Resumo. *No Brasil, o principal exame de avaliação do desempenho educacional do ensino básico é o Exame Nacional do Ensino Médio (ENEM), o qual é também usado para entrada de estudantes no ensino superior. No ano de 2020, com a chegada do vírus da COVID-19, as instituições de nível básico precisaram alterar seu modelo educacional com ensino presencial para o uso de metodologias de ensino remoto. Este trabalho objetiva identificar os principais impactos causados pela pandemia no ENEM nos estados do Ceará, Maranhão e Piauí, no seu primeiro ano. Para isso, é realizado uma análise exploratória de dados do ENEM de 2019, ano anterior ao início da pandemia, e do ENEM de 2020, primeiro ano da pandemia. Os resultados mostram uma melhora no desempenho geral dos estados, um aumento no número de inscritos que não compareceram à prova, e um crescimento da desigualdade entre os participantes provenientes de instituições públicas e privadas.*

1. Introdução

Em 11 de março de 2020, a Organização Mundial da Saúde (OMS) declarou que a epidemia da COVID-19 se tornava uma pandemia internacional. Os estudos sobre o vírus da COVID-19 mostram que o vírus é altamente contagioso e pode causar sérios danos aos seres humanos [Daniel 2020]. De acordo com [Robbins 2020], os altos custos sociais, educacionais e econômicos do fechamento de escolas afetam as pessoas em todas as comunidades, mas atingem desproporcionalmente os grupos vulneráveis e suas famílias.

Em 25 de fevereiro de 2020, o Brasil registrou seu primeiro caso. De acordo com Boletim Epidemiológico COVID-19 do Governo Federal, desde a confirmação do primeiro caso em fevereiro de 2020 até o final de junho de 2022, o Brasil registrou cerca de mais de 32 milhões de casos e mais de 670 mil óbitos.

Com a chegada da pandemia da COVID-19, as instituições de ensino começaram a criar métodos para que a educação não fosse comprometida e o processo de ensino pudesse prosseguir [Educação 2020]. A pandemia da COVID-19 afetou diretamente a vida dos estudantes e professores de diferentes maneiras, independentemente da escolaridade ou nível do curso. O fechamento de escolas mudou a estrutura de aprendizagem e escolarização, bem como os métodos de ensino e avaliação [Alqahtani and Rajkhan 2020]. Ao aumentar a capacidade de ensinar remotamente, escolas tiveram que se adaptar a utilizar métodos de ensino remoto e à distância, requerendo que alunos e professores se adequassem e aprendessem sobre esses métodos rapidamente [Gularte et al. 2021].

No Brasil, existem diversos programas de avaliação da educação, desde a avaliação do ensino infantil até o nível superior. Programas como o Índice de Desenvolvimento da Educação Básica (IDEB), Sistema de avaliação da Educação Básica (SAEB), Censo Escolar, Exame Nacional do Ensino Médio (ENEM), Exame Nacional de Desempenho dos Estudantes (ENADE), dentre outros, servem de apoio para estados e municípios no processo de tomada de decisão. O ENEM é uma avaliação criada e aplicada pelo Ministério da Educação para medir o desempenho de alunos que concluíram o ensino médio. Ele foi originalmente concebido como um exame final para avaliar os alunos ao final da educação básica. No entanto, em 2009, o exame teve o objetivo de substituir o vestibular das universidades federais.

A mineração de dados educacionais tem como foco o desenvolvimento de métodos que buscam extrair *insights* usando dados coletados em ambientes educacionais [de ANDRADE et al. 2021]. Seu principal objetivo é desenvolver ou adaptar técnicas e algoritmos de mineração com o intuito de facilitar a compreensão dados em ambientes educacionais. Uma das etapas do processo de mineração de dados é a análise exploratória, a qual visa o uso de métodos estatísticos descritivos e ferramentas gráficas para melhor compreender os dados.

Esse estudo apresenta um processo de análise exploratória dos microdados do ENEM dos anos de 2019 e 2020 no Maranhão, Ceará e Piauí, objetivando identificar se a pandemia impactou o ENEM nesses estados. Os estados do Ceará, Maranhão e Piauí, estão na região nordeste do Brasil, a qual possui os dados mais preocupantes sobre a alta desigualdade social do país, afetando diretamente na oferta de educação com qualidade [Oliveira and Clementino 2020]. Os estados estudados foram também selecionados devido à sinergia entre eles, o que permitiu a criação da Escola Regional de Computação dos Estados do Ceará, Maranhão e Piauí (ERCEMAPI).

O restante deste artigo está organizado como segue. A Seção 2 discute os trabalhos relacionados. A Seção 3 explica como a análise exploratória foi conduzida, enquanto a Seção 4 apresenta e discute os resultados encontrados. Por fim, a Seção 5 traz as considerações finais.

2. Trabalhos Relacionados

A análise de dados tem sido usada para estudar a educação básica brasileira [Soares et al. 2021]. Segundo [Peña-Ayala 2014], os resultados da aplicação de análise de dados podem fornecer apoio para o processo de tomada de decisão da gestão escolar. No Brasil existem vários estudos que utilizam técnicas de análise de dados no contexto educacional, analisando todo o território nacional. [Barcellos et al. 2018] utilizam técnicas de mineração de dados nos microdados do ENEM de 2018, com o intuito de correlacionar dados financeiros dos participantes com seu desempenho no Exame. O artigo [Simon and Cazella 2017] usa técnicas de árvore de decisão para gerar um modelo preditivo com os resultados do ENEM 2015 e melhorar o indicador de desempenho médio na área de conhecimento de Ciências da Natureza e suas tecnologias. O artigo de [Nakazone and Bortolotti 2021] analisa os microdados do ENEM dos anos de 2015 a 2019 para entender as diferenças entre as provas ao longo dos anos. O estudo de [de Moraes et al. 2021] compreende os fatores e as relações das variáveis apresentadas nos microdados do ENEM durante a pandemia e avalia o desempenho dos alunos na prova de Matemática e suas tecnologias.

Vários estudos ao redor do mundo têm focado no impacto da pandemia na educação. Por exemplo, para ressaltar a importância do tema, o editorial [Reuge et al. 2021] apresenta uma proposta pela *United Nations International Children's Emergency Fund* (UNICEF) em resposta à pandemia da COVID-19. O estudo proposto por [Junior et al. 2021] apresenta uma análise dos perfis dos alunos durante a pandemia. Conforme os resultados do artigo, na modalidade da educação a distância, os alunos com melhor infraestrutura financeira, ou seja, aqueles que possuíam acesso à Internet e computadores de maior qualidade, acabaram lidando melhor com essa modalidade de ensino.

Existem vários trabalhos usando análise de dados educacionais para entender questões educacionais [Peña-Ayala 2014, Soares et al. 2021]. Entretanto, para o melhor do nosso conhecimento, não existe artigo publicado na literatura científica que tenha conduzido um estudo de modo a comparar os dados, dos anos de 2019 e 2020 e dos estados do Ceará, Maranhão e Piauí, visando compreender o impacto da pandemia da COVID-19 no ENEM. Esse trabalho contribui, portanto, na melhora no entendimento da situação educacional dos estados durante o período pandêmico.

3. Materiais e Métodos

Este estudo visa analisar os dados do ENEM dos estados do Ceará, Maranhão e Piauí de 2019 e 2020. Os microdados do ENEM fornecem dados educacionais para todos os estados e territórios do Brasil, por dependência administrativa (Municipal, Estadual, Federal e Particular), notas médias dos candidatos por ano de conclusão acadêmica e outras informações acerca dos participantes e de suas instituições de ensino.

Para esse estudo, foi utilizada a metodologia *Cross-Industry Standard Process for Data Mining* (CRISP-DM) [Azevedo and Santos 2008, Shearer 2000] para medir a qualidade dos resultados em um projeto de mineração de dados. O ciclo de vida do projeto na metodologia CRISP-DM é dividido em seis fases: entendimento do negócio, entendimento dos dados, preparação dos dados, modelagem, avaliação e distribuição [Azevedo and Santos 2008, Shearer 2000]. O CRISP-DM permite a

realização de uma análise exploratória seguindo apenas as três primeiras fases. Neste trabalho, usamos a linguagem de programação *Python*, as bibliotecas *Pandas*, *NumPy*, e *Matplotlib*, com o *Google Colaboratory* como ambiente de desenvolvimento.

3.1. Entendimento do Negócio e dos Dados

Seguindo a metodologia CRISP-DM, a primeira fase é entender o negócio, em que foi feita a identificação dos recursos disponíveis, problemas e objetivos. Para esse estudo, foi feita uma análise das particularidades das mudanças que ocorrem no ENEM no primeiro ano de pandemia.

A próxima fase da metodologia consiste no entendimento dos dados. Para isso, foi feita a leitura dos documentos adicionais (i.e., Matriz de Referência do ENEM, Notas Técnicas, Dicionário dos Dados) presentes na base de dados do ENEM com o intuito de compreender cada coluna do conjunto de dados. Nesse estudo, usamos dois conjuntos de dados: os microdados do ENEM de 2019 e 2020, ambos fornecidos pelo INEP em seu portfólio de dados abertos. O conjunto de dados do ENEM de 2019 possui 5.097.407 linhas e 136 colunas, referentes aos participantes de todo o Brasil. Cada linha corresponde a um participante, e as colunas apresentam diversas informações sobre o participante, tais como o estado onde mora, ano de conclusão do ensino médio, sexo, raça, faixa etária, nota de cada área do conhecimento, dados socioeconômicos, dentre outras. Os microdados do ENEM de 2020 sofreram alterações em relação aos anos anteriores devido à Lei Geral de Proteção de Dados (LGPD). Informações como código de identificação da escola, pedidos de atendimento especializado, dentre outros dados foram retirados em 2020. O conjunto de dados de 2020 possui 5.827.389 linhas e 76 colunas, contendo informações dos participantes de todo o Brasil.

3.2. Preparação dos Dados

A fase de preparação envolve selecionar, limpar e transformar os dados que serão usados para o projeto. Nessa fase, é decidido se será usado todo o conjunto de dados ou um subconjunto para conter determinadas colunas e linhas [Saltz 2021]. O primeiro processo foi a criação de uma nova coluna com a nota média dos participantes, pois o conjunto de dados só possuía as médias de cada uma das áreas de conhecimento. Então foi calculada a média a partir das cinco notas das áreas de conhecimento. Após a criação da coluna, foram removidas as linhas com dados nulos na média de notas, que representam participantes ausentes, seja em apenas um dia do exame, ou ausente nos dois dias. Em seguida, colunas como raça, gênero e dependências administrativas também foram convertidas para outras nomenclaturas. Os dados que antes eram numéricos, passaram a ser categóricos, melhorando o entendimento. Nesta etapa, também foi utilizado o dicionário de microdados do ENEM, ferramenta disponibilizada pelo INEP para auxiliar no entendimento das colunas a serem exploradas. Vale ressaltar que o processo de preparação dos dados é crucial para a fase de análise dos dados. Portanto, os conjuntos de dados de 2019 e 2020 passaram pelo mesmo processo de preparação.

3.3. Análise Exploratória

Nesta etapa foi realizada a investigação preliminar de dados para descobrir padrões, detectar anomalias, testar e validar hipóteses com a ajuda de estatísticas descritivas e representações gráficas [Komorowski et al. 2016]. O propósito desta etapa do estudo foi

explorar dados que pudessem ser traduzidos em informações importantes para entender o impacto da pandemia no ENEM. Além da análise exploratória de dados, o foco principal desta etapa foi gerar informações comparativas e analisar se houve impacto no ENEM. Durante essa fase, a análise foi feita filtrando um estado de cada vez, mas mantendo as demais variáveis, sendo assim, usamos os mesmos dados predefinidos na fase de preparação em todos os estados. Para filtrar apenas as colunas de cada estado, foi usada a indexação booleana disponível na biblioteca *Pandas*. Esse tipo de indexação é usada para filtrar valores de uma coluna no conjunto de dados.

As colunas selecionadas no conjunto de dados foram definidas para possibilitar uma análise mais abrangente. Para isso, foram selecionadas as colunas sobre questões socioeconômicas, informações sobre a escola do ensino médio do participante e desempenho dele. Para todas as análises, consideramos apenas os participantes presentes, em que são caracterizados pela presença nos dois dias de prova, a qual é a maneira como o INEP determina para considerar como presente. Após selecionar as variáveis, geramos uma análise exploratória utilizando dois *notebooks* no *Google Collaboratory*) para cada ano, separadamente, usando gráficos e tabelas dos três estados.

4. Resultados e Discussão

A Figura 1 mostra a quantidade de participantes ausentes e presentes nos três estados. Pode-se perceber que houve um aumento no número de candidatos inscritos no ano de 2020. Também é possível verificar que a quantidade de candidatos ausentes aumentou bastante nos estados analisados.

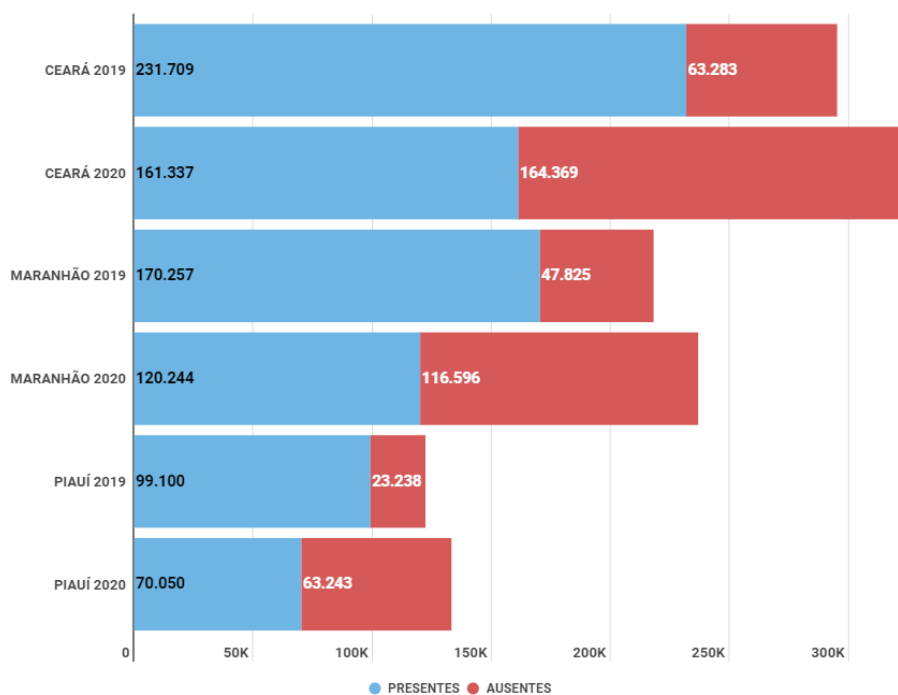


Figura 1. Quantidade de participantes inscritos: presentes e ausentes.

O estado do Ceará teve mais de 50% dos participantes ausentes. Vale ressaltar que a média de ausentes no Brasil chegou a cerca de 52%. Ou seja, dos três estados, apenas o

Ceará ficou acima da média nacional. Comparando os anos, podemos notar que, proporcionalmente, o maior aumento de ausências ocorreu no estado do Piauí. No ano de 2019, o Piauí teve cerca de 19% de ausentes, enquanto no ano de 2020 subiu para 47,4%, um valor ainda abaixo da média nacional, mas que, comparado ao ano anterior, mostra que a pandemia em seu primeiro ano impactou diretamente a falta dos participantes. O Maranhão também ficou abaixo da média nacional, mas, apesar disso, o estado teve um aumento de 68.771 participantes ausentes entre os anos, demonstrando que os participantes do ENEM no estado sofreram um grande impacto no primeiro ano da pandemia. Os números aqui apresentados nos levam a acreditar (i.e., hipóteses) que os participantes podem ter desistido de fazer o exame em 2020 por motivos como: insegurança com o desconhecimento da doença, sua alta taxa de transmissão, questões econômicas, e as dificuldades para adquirir os conhecimentos exigidos no exame, uma vez que os participantes passaram a ter aulas remotamente, o que pode ter levado os alunos a não se sentirem preparados. Outro motivo pode ter sido devido à mudança na data da prova devido à pandemia, que costuma acontecer nos meses de outubro e novembro, mas em 2020 foi realizada em janeiro de 2021.

A Figura 2 mostra gráficos de *boxplot* com as notas médias dos participantes dos estados do Ceará, Maranhão e Piauí, respectivamente, em relação ao tipo de dependência administrativa das escolas. O gráfico mostra que os estados não sofreram alteração significativa no desempenho dos alunos pelo tipo de dependência administrativa da escola. Pode-se notar que nos três estados e em ambos os anos, as escolas privadas estão com as maiores notas, seguido das escolas federais, estaduais e municipais. Em 2019, participantes de escolas privadas no Ceará somavam 6.708, já em 2020, esse número subiu para 7.533, no Maranhão e Piauí também houve o aumento dos participantes de escolas privadas. No que tange as escolas públicas, o Ceará teve menos participantes nos demais tipos de dependência administrativa. No Maranhão, houve aumento apenas no número de participantes de escolas municipais, porém, o aumento foi de apenas 9 pessoas. O estado do Piauí seguiu a mesma trajetória do Maranhão, com um aumento somente em participantes de escolas municipais, com 17 pessoas a mais.

Em 2019, às sete melhores escolas do Brasil com as maiores médias de notas eram privadas, sendo as duas primeiras colocadas do estado do Ceará. Nessa classificação de melhores escolas, o Piauí surge já na 11^a colocação, também com uma escola privada. Já o Maranhão, aparece apenas na 102^a posição. Para mostrar a diferença entre as escolas públicas e particulares, no caso do Ceará, só aparece uma escola pública no ranking, na 802^a posição, sendo uma escola federal. O Maranhão só possui escola pública nesse ranking na 2.396^a posição, também com uma escola de dependência administrativa do tipo federal. O estado do Piauí é o único dentre os três estados que apresenta uma escola estadual melhor qualificada dentre escolas públicas. No conjunto de dados referente ao ano de 2020, essa análise não pôde ser realizada porque o INEP removeu os códigos escolares do conjunto de dados devido a conter dados que podem ir em desacordo com a Lei Geral de Proteção de Dados (LGPD) [INEP 2022].

A Figura 3 mostra gráficos de *violinplot* a nota dos participantes dos estados do Ceará, Maranhão e Piauí, respectivamente, em relação à renda familiar declarada. A ordem de Renda Familiar apresentada no gráfico vai de 'A' até 'Q', em que 'A' representa os participantes que declaram não ter nenhuma renda familiar, e 'Q' são os participantes

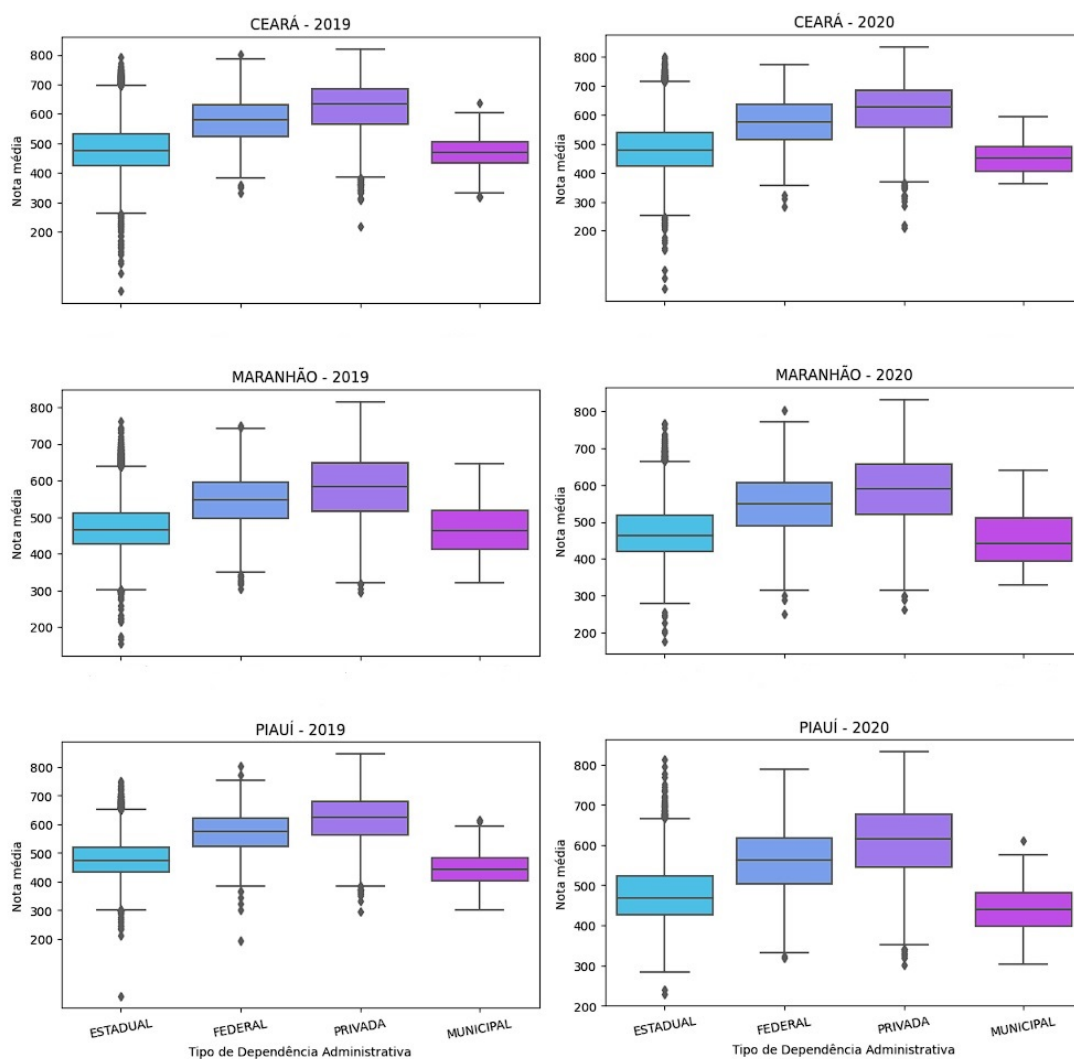


Figura 2. Notas médias dos candidatos nos três estados em relação ao tipo de dependência administrativa das escolas.

que declaram renda familiar mensal acima de R\$ 19.000,00. Pode-se notar que, em todos os estados, o impacto causado no primeiro ano de pandemia em relação ao desempenho por classe social foi baixa. Vale ressaltar que o desempenho médio do Brasil (i.e., todos os estados) aumentou no ano de pandemia.

O Ceará teve uma nota média de 518,77 no ano de 2019, e no ano de 2020, o estado teve um aumento, ficando com 520,47 no primeiro ano de pandemia. O estado do Maranhão também teve um aumento de 0,5 pontos, passando de 493,81 em 2019 para 494,31. Dentre os estados, apenas o estado do Piauí teve diminuição da média, saindo de 506,79 para 506,27 em 2020. Apesar dos aumentos e da diminuição, as diferenças são pequenas. O que nos leva a acreditar que, devido ao exame usar da Teoria da Resposta ao Item (TRI) [INEP 2011], a prova pode ser nivelada em comparação ao desempenho de todos os outros participantes que também realizaram o exame. Vale ressaltar que, nos estados aqui apresentados, tanto nos anos de 2019 quanto 2020, a renda familiar é a característica que possui maior correlação com o resultado da nota média.

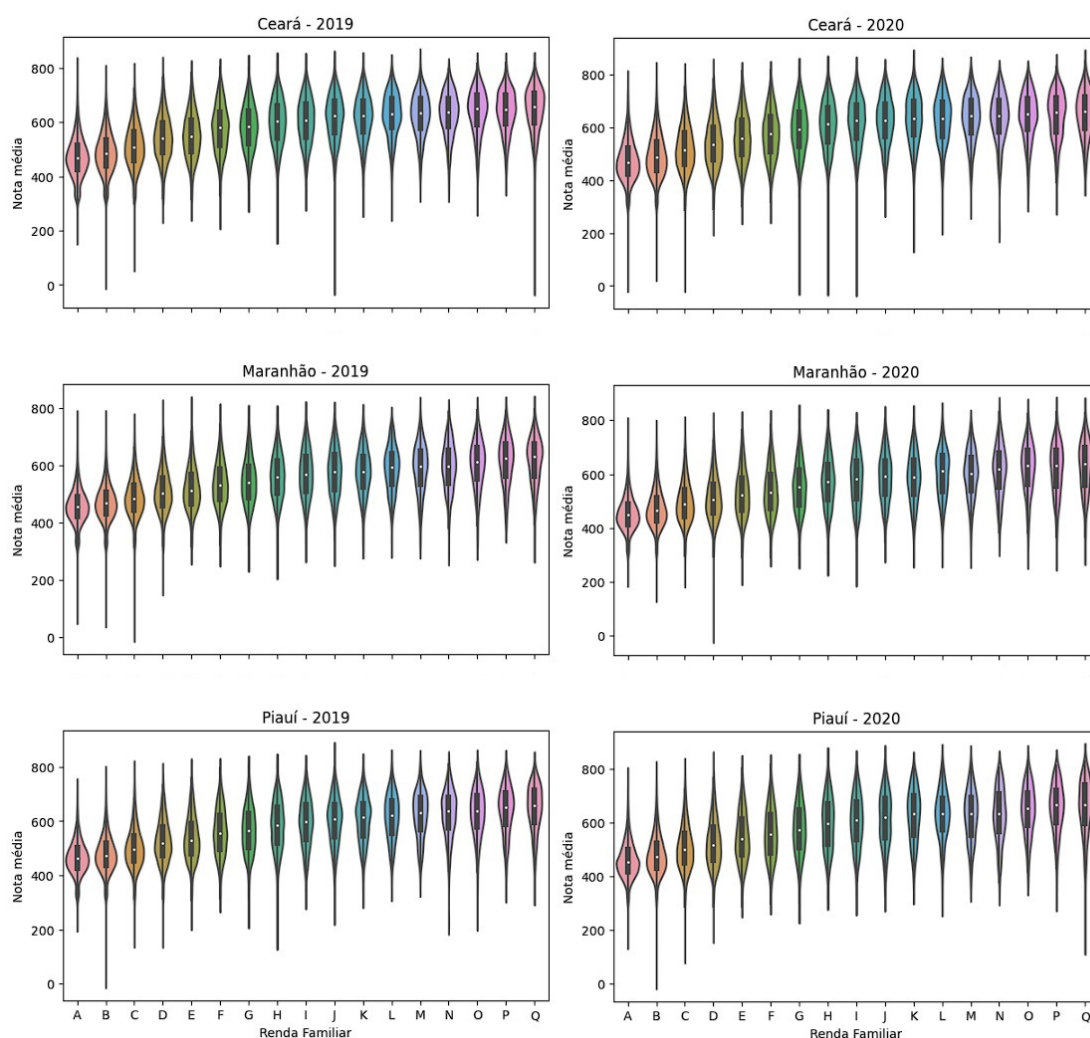


Figura 3. Nota média dos participantes conforme a renda familiar.

Como apresentado na Figura 1, todos os estados mostraram um número significativo de ausentes. Isso leva a supor que os participantes ausentes podem ser de grupos sociais mais vulneráveis, sendo mais propensos a se ausentarem. Nesse sentido, esses participantes social e economicamente mais vulneráveis teriam maior probabilidade de desempenho inferior ao dos candidatos com melhor posição social, econômica e educacional, fazendo, assim, com que a média fosse superior ao ano anterior.

5. Considerações Finais

Utilizando uma análise de dados educacionais, este estudo identificou informações relevantes para responder se a pandemia de COVID-19 impactou os estados do Ceará, Maranhão e Piauí no ENEM. Para isso, o estudo apresentou resultados de uma análise exploratória e comparativa dos microdados do ENEM de 2019 e 2020 dos três estados. Constatamos que a presença de candidatos no ENEM de 2020 foi bastante afetada pela pandemia em ambos os estados. Em relação ao desempenho, o Ceará teve a média geral maior em 2020 do que em 2019, o Maranhão também teve um leve aumento, e o estado do Piauí teve uma diminuição leve na média geral. Sobre as notas em relação às escolas

públicas e privadas, os participantes de escolas municipais, apesar de estarem em menor número que os demais, atingiram menores médias do que os demais tipos de dependência administrativa nos três estados. Pode-se notar que as escolas privadas estão com desempenho melhor no exame, em ambos os estados. Outra informação relevante identificada foi que os participantes com maior renda familiar declarada possuem maior média.

A alteração causada no conjunto de dados do ano de 2020 e a indisponibilidade dos dados do ano de 2021 limitou os achados desse trabalho. Além disso, o estudo considerou apenas o primeiro ano da pandemia. Dessa forma, acreditamos que esta pesquisa está em uma fase inicial. A continuação deste trabalho irá servir para a melhora do entendimento do impacto da pandemia da COVID-19 no ensino básico brasileiro. Portanto, planos para trabalhos futuros incluem a evolução nas análises de forma a executar o processo completo de mineração de dados educacionais, como também a inclusão na análise dos microdados do ano de 2021.

Agradecimentos

Os autores gostariam de agradecer a Fundação de Amparo à Pesquisa e ao Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA) pelo apoio dado a seus projetos de pesquisa.

Referências

- 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. *Education Sciences*, 10(9):216.
- Azevedo, A. I. R. L. and Santos, M. F. (2008). Kdd, semma and crisp-dm: a parallel overview. *IADS-DM*.
- Barcellos, A. A., Isotani, S., Diego, C., and Damasceno, N. (2018). Mineração de dados abertos - enem 2018. *Anais dos Trabalhos de Conclusão de Curso da Pós-Graduação em Computação Aplicada à Educação*.
- Daniel, S. J. (2020). Education and the COVID-19 pandemic. *PROSPECTS*, 49(1-2):91–96.
- de ANDRADE, T. L., RIGO, S. J., and BARBOSA, J. L. V. (2021). Active methodology, educational data mining and learning analytics: A systematic mapping study. *Informatics in Education*, 20(2):171–204.
- de Moraes, C. P., Peres, R. T., and Pedreira, C. E. (2021). Eficácia escolar e variáveis familiares em tempos de pandemia: um estudo a partir de dados do enem. *Interfaces da educação*, 12(35):635–658.
- Educação, T. P. (2020). Ensino a distância na educação básica frente à pandemia da covid-19. *Nota Técnica*.
- Gularte, F. N., do Nascimento, F. K. V., and de Carvalho, A. V. (2021). ENEM em tempos de pandemia: desafios enfrentados pelos docentes frente o ensino remoto da rede estadual de presidente kennedy/TO. *Research, Society and Development*, 10(14):e571101422339.

- INEP (2011). Teoria de resposta ao item. https://download.inep.gov.br/educacao_basica/enem/nota_tecnica/2011/nota_tecnica_tri_enem_18012012.pdf. Online; Acessado em 04 de junho de 2022.
- INEP (2022). Nota de esclarecimento — divulgação dos microdados. <https://www.gov.br/inep/pt-br/assuntos/noticias/institucional/nota-de-esclarecimento-divulgacao-dos-microdados>. Online; Acessado em 07 de maio de 2022.
- Junior, L. P., Matos, S. N., and Borges, H. B. (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. *RENOTE*, 18(2):336–345.
- Komorowski, M., Marshall, D. C., Salciccioli, J. D., and Crutain, Y. (2016). Exploratory data analysis. *Secondary Analysis of Electronic Health Records*, pages 185–203.
- Nakazone, E. and Bortolotti, L. M. (2021). Análise de dados históricos do enem entre 2015 à 2019. In *Congresso de Tecnologia-Fatec Mococa*, volume 4.
- Oliveira, D. A. and Clementino, A. M. (2020). As políticas de avaliação e responsabilização no brasil: uma análise da educação básica nos estados da região nordeste. *Revista Iberoamericana de Educación*, 83(1):143–162.
- Peña-Ayala, A. (2014). Educational data mining: A survey and a data mining-based analysis of recent works. *Expert Systems with Applications*, 41(4, Part 1):1432–1462.
- Reuge, N., Jenkins, R., Brossard, M., Soobrayan, B., Mizunoya, S., Ackers, J., Jones, L., and Taulo, W. G. (2021). Education response to COVID 19 pandemic, a special issue proposed by UNICEF: Editorial review. *International Journal of Educational Development*, 87:102485.
- Robbins, J. (2020). Unesco: 290 million students stay home due to coronavirus. *VOA Learning English*.
- Saltz, J. S. (2021). Crisp-dm for data science: Strengths, weaknesses and potential next steps. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2337–2344.
- Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, 5(4):13–22.
- Simon, A. and Cazella, S. (2017). Mineração de dados educacionais nos resultados do ENEM de 2015. In *Anais dos Workshops do VI Congresso Brasileiro de Informática na Educação (CBIE 2017)*. Brazilian Computer Society (Sociedade Brasileira de Computação - SBC).
- Soares, R. D. C., Neto, N. W., Coutinho, L. R., e Silva, F. J. D. S., dos Santos, D. V., and Teles, A. S. (2021). Mineração de dados da educação básica brasileira usando as bases do INEP: Uma revisão sistemática da literatura. *RENOTE*, 19(1):361–370.