ISSN: 2358-8829

ANÁLISE DO IMPACTO DE VARIÁVEIS SOCIOECONÔMICAS NO DESEMPENHO EM MATEMÁTICA NO ENEM UTILIZANDO MACHINE LEARNING

João Gabriel de Carvalho Santos [1]
João Paulo Barros Sousa [2]
Mairon Lorran Santos Caldas [3]
Francisco Willian Nunes Gomes [4]
Erick MacGregor Santos Lima [5]

INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) consolidou-se como o principal instrumento de avaliação em larga escala do Brasil, servindo simultaneamente como via de acesso ao ensino superior e ferramenta diagnóstica para formulação de políticas públicas (INEP, 2023). Entre as áreas avaliadas, Matemática e suas Tecnologias é estratégica, por estar diretamente relacionada ao raciocínio lógico, ao letramento quantitativo e ao ingresso em cursos das áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM).

Diversos estudos demonstram que o desempenho no ENEM não decorre apenas do esforço individual, mas é fortemente condicionado por fatores socioeconômicos como renda familiar, escolaridade dos responsáveis, tipo de escola frequentada, acesso à internet, posse de computador e condições domiciliares de estudo (SOARES; ANDRADE, 2006; MELO et al., 2021). Tais elementos dialogam com Bourdieu (1998), ao indicar que diferentes volumes de capital econômico e cultural se convertem em oportunidades desiguais de sucesso escolar, contribuindo para a reprodução das desigualdades históricas.

Nesse cenário, técnicas de Machine Learning (ML) aplicadas aos microdados do ENEM surgem como ferramenta promissora para identificar padrões complexos, estimar probabilidades e mensurar a importância relativa de variáveis explicativas (ROMERO; VENTURA, 2010; GERON, 2019). Este trabalho, desenvolvido por estudantes do ensino médio técnico, orientados por docentes do IP Coelho Neto, tem como objetivo analisar o impacto de variáveis socioeconômicas no desempenho em Matemática no ENEM 2023,

[1] Professor do IEMA Pleno Coelho Neto, jgabriel.professional@gmail.com;
[2] Aluno do IEMA Pleno Coelho Neto.
[3] Aluno do IEMA Pleno Coelho Neto.
[4] Aluno do IEMA Pleno Coelho Neto.
[5] Professor do IEMA Pleno Coelho Neto, erickmacgregor2_@hotmail.com

ISSN: 2358-8829

utilizando um modelo de Árvore de Decisão para investigar quais fatores mais se associam à probabilidade de o participante alcançar nota igual ou superior a 500 pontos.

METODOLOGIA (OU MATERIAIS E MÉTODOS)

A pesquisa é quantitativa, de caráter exploratório-descritivo, baseada nos microdados do ENEM 2023 disponibilizados pelo INEP (INEP, 2023). As etapas envolveram seleção da base, tratamento dos dados, definição das variáveis, construção do modelo de ML e análise dos resultados.

Foram extraídas da base oficial as informações referentes à nota de Matemática e às respostas do Questionário Socioeconômico. Para garantir viabilidade computacional e adequação ao contexto de laboratório escolar, foi selecionada uma amostra aleatória simples de 7.729 participantes com registros completos nas variáveis de interesse, preservando distribuição aproximada das categorias socioeconômicas.

A variável dependente foi transformada em binária:

0 – nota de Matemática < 500;

1 – nota de Matemática ≥ 500.

Foram utilizadas como variáveis independentes, principalmente: renda familiar; escolaridade dos responsáveis; tipo de escola do ensino médio (pública/privada); posse de computador; acesso à internet; número de moradores no domicílio; e acesso a alguns bens e recursos educacionais. Variáveis não relacionadas diretamente ao objetivo do estudo foram descartadas para reduzir ruídos.

As respostas categóricas foram convertidas em códigos numéricos, respeitando a ordenação lógica (por exemplo, faixas de renda crescentes). Casos com dados ausentes nas variáveis-chave foram excluídos. Em seguida, os dados foram divididos em 80% para treinamento e 20% para teste.

Foi adotado o algoritmo de **Árvore de Decisão**, por sua interpretabilidade e potencial pedagógico, permitindo visualizar caminhos de decisão e compreender, de forma acessível, o efeito combinado das variáveis (HAN; KAMBER; PEI, 2012). O modelo foi implementado em Python com as bibliotecas *pandas*, *numpy* e *scikit-learn* (GERON, 2019). Avaliaram-se a acurácia no conjunto de teste, a matriz de confusão e a importância relativa das variáveis. Por se tratar de dados públicos anonimizados, o estudo atende às diretrizes éticas de uso de bases secundárias.

REFERENCIAL TEÓRICO

ISSN: 2358-8829

As desigualdades educacionais no Brasil vêm sendo associadas, de forma consistente, ao perfil socioeconômico dos estudantes. Soares e Andrade (2006) evidenciam que renda, escolaridade dos pais e tipo de escola estão diretamente relacionados ao desempenho em avaliações externas. Estudos recentes com dados do ENEM reforçam esse padrão, mostrando que estudantes de maior nível socioeconômico concentram as maiores notas, sobretudo em Matemática (MELO et al., 2021).

A teoria do capital cultural de Bourdieu (1998) contribui para interpretar tais resultados: famílias com maior capital econômico e cultural tendem a oferecer ambientes mais letrados, apoio escolar, acesso a aulas extras e materiais de qualidade, o que favorece trajetórias de sucesso. A escola, em muitos casos, legitima essas diferenças prévias, fazendo com que exames em larga escala reflitam desigualdades estruturais.

Paralelamente, a área de Educational Data Mining e Learning Analytics aponta o uso de algoritmos de ML como estratégia relevante para análise de grandes bases educacionais, permitindo identificar fatores de risco, prever desempenho e orientar políticas de intervenção (ROMERO; VENTURA, 2010). Modelos interpretáveis, como árvores de decisão, têm a vantagem de combinar capacidade preditiva com transparência, aspecto fundamental quando se discutem desigualdades e justiça educacional (HAN; KAMBER; PEI, 2012).

Inserido nessa perspectiva, o presente estudo utiliza ML não apenas como ferramenta técnica, mas como recurso formativo e crítico para leitura dos dados do ENEM, articulando ciência de dados, análise socioeconômica e formação científica na educação básica.

RESULTADOS E DISCUSSÃO

O modelo de Árvore de Decisão obteve acurácia aproximada de 67% no conjunto de teste, indicando desempenho moderado na classificação dos participantes entre os grupos com nota abaixo e igual ou acima de 500 pontos em Matemática. Considerando a complexidade dos condicionantes de aprendizagem, esse resultado é adequado a um estudo exploratório e didático, ainda que não represente um modelo definitivo.

A análise da importância das variáveis revelou três grupos de fatores centrais:

- **Renda familiar** – principal variável do modelo, associada a maior probabilidade de alcançar ≥ 500 pontos, sugerindo que estudantes com melhores condições econômicas têm acesso ampliado a escolas de qualidade, materiais didáticos, aulas de reforço e ambientes mais favoráveis ao estudo.

ISSN: 2358-8829

- **Tipo de escola cursada (pública/privada)** – indicador relevante, que reforça a clivagem entre oportunidades oferecidas pelas redes pública e privada, alinhando-se às evidências de desigualdade de condições de oferta (SOARES; ANDRADE, 2006).
- **Acesso tecnológico (computador e internet em casa)** – variável com peso significativo, indicando que a inclusão digital potencializa o acesso a videoaulas, simulados e plataformas de estudo, fatores importantes na preparação para o ENEM.

Esses resultados convergem com a literatura que aponta a centralidade das condições socioeconômicas na explicação do desempenho em exames padronizados (MELO et al., 2021) e dialogam com Bourdieu (1998), ao evidenciar que os estudantes não partem do mesmo ponto de largada. As trajetórias acadêmicas são atravessadas pela disponibilidade desigual de capitais, o que faz com que o ENEM, frequentemente tomado como parâmetro meritocrático, também reflita assimetrias estruturais.

Do ponto de vista pedagógico, a Árvore de Decisão mostrou-se adequada por permitir que os próprios estudantes autores visualizassem como combinações de renda, tipo de escola e acesso tecnológico alteram a probabilidade de bom desempenho. Essa experiência favoreceu a compreensão crítica de que algoritmos não são neutros; eles apenas explicitam padrões presentes na realidade social e podem servir de base para discussões sobre justiça educacional.

Ao mesmo tempo, os resultados indicam possibilidades de aprimoramento técnico, como a testagem de modelos mais robustos (Random Forest, Gradient Boosting), ampliação da amostra e inclusão de variáveis regionais. Tais avanços podem gerar modelos com maior acurácia e análises mais refinadas, sem perder de vista o compromisso ético com a transparência e a interpretação social dos dados.

CONSIDERAÇÕES FINAIS

A aplicação de Machine Learning aos microdados do ENEM 2023 permitiu identificar que o desempenho em Matemática está fortemente associado a fatores socioeconômicos, especialmente renda familiar, tipo de escola e acesso a recursos tecnológicos. Longe de representar apenas diferenças individuais de esforço, os resultados reforçam que o sucesso em Matemática e o acesso a cursos concorridos permanecem condicionados pela estrutura de oportunidades disponível a cada grupo social.

ISSN: 2358-8829

O estudo aponta a necessidade de políticas públicas que promovam: investimentos na qualidade da escola pública; programas de reforço em Matemática para estudantes em vulnerabilidade; ampliação do acesso à internet banda larga e a equipamentos; e ações de acompanhamento sistemático do desempenho com base em evidências. Defende-se que a leitura crítica dos dados do ENEM deve considerar o contexto socioeconômico, evitando interpretações simplistas que culpabilizem o indivíduo.

Do ponto de vista formativo, a experiência demonstra o potencial de envolver estudantes da educação básica técnica em projetos de ciência de dados aplicada à realidade brasileira, fortalecendo competências em programação, estatística e análise crítica de informações oficiais. Como continuidade, propõe-se a comparação entre diferentes edições do ENEM e a incorporação de outros algoritmos, mantendo como eixo central a reflexão sobre equidade, inclusão e justiça social.

Palavras-chave: ENEM; Desempenho em Matemática; Variáveis socioeconômicas; Machine Learning; Desigualdades educacionais.

AGRADECIMENTOS

Agradecemos aos nossos orientadores Erick MacGregor Santos Lima e João Gabriel de Carvalho Santos pelo incentivo, orientação e oportunidade de participação em atividades de iniciação científica, bem como ao IP Coelho Neto (IEMA) pelo suporte institucional e pela disponibilização da estrutura física necessária à realização deste trabalho.

REFERÊNCIAS

BOURDIEU, P. Escritos de Educação. Petrópolis: Vozes, 1998.

GERON, A. Mãos à Obra: Aprendizado de Máquina com Scikit-Learn, Keras & TensorFlow. Rio de Janeiro: Alta Books, 2019.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. San Francisco: Morgan Kaufmann, 2012.

INEP. Microdados do ENEM 2023. Brasília: Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2023.

MELO, A. et al. Desigualdades educacionais e desempenho no ENEM: uma análise das condições socioeconômicas dos participantes. Revista Brasileira de Estudos em Avaliação Educacional, 2021.

ISSN: 2358-8829

ROMERO, C.; VENTURA, S. Educational data mining: a review of the state-of-the-art. IEEE Transactions on Systems, Man, and Cybernetics, v. 40, n. 6, 2010.

SOARES, J. F.; ANDRADE, R. J. Nível socioeconômico, qualidade e equidade das escolas de Belo Horizonte. Ensaio: Avaliação e Políticas Públicas em Educação, v. 14, n. 52, 2006.