CONEXÃO UNIFAMETRO 2025
XXI SEMANA ACADÊMICA
ISSN: 2357-8645

ANÁLISE DA RELAÇÃO ENTRE RENDA E DESEMPENHO NO ENEM: ESTUDO DE CASO BASEADO EM INTELIGÊNCIA ARTIFICIAL PARA DIFERENTES REGIÕES BRASILEIRAS

Tarcísio Alex Almeida de Paula
Aluno - Centro Universitário Fametro
tarcisioalexalmeida@gmail.com

Ana Clara Oliveira Fernandes
Aluno - Centro Universitário Fametro
clara.of93@gmail.com

Ana Laura Bastos
Aluno - Centro Universitário Fametro
analaurabastos890@gmail.com

Hiury Rodrigo da Silva Alves
Aluno - Centro Universitário Fametro
hiuryalves.contato@gmail.com

Julianny Albuquerque Lima
Aluno - Centro Universitário Fametro
albuquerque.julianny17@gmail.com

Kaio Gefferson de Almeida Mesquita
Orientador - Centro Universitário Fametro
kaio.mesquita@professor.unifametro.edu.br

Kauã Silva do Nascimento
Aluno - Centro Universitário Fametro
kaua.nascimento64@aluno.ce.gov.br

Área Temática: Inovação e Inteligência Artificial.
Área de Conhecimento: Ciências Tecnológicas.
Encontro Científico: XIII Encontro de Iniciação à Pesquisa.

RESUMO

O presente estudo integra um projeto de Iniciação Científica que busca compreender as desigualdades educacionais no Brasil, analisando a relação entre renda familiar, localidade e desempenho dos estudantes no Exame Nacional do Ensino Médio (ENEM) 2023. Considerando a relevância do ENEM como porta de entrada para o ensino superior e as diferenças socioeconômicas entre regiões, o objetivo é investigar como a renda per capita e fatores contextuais influenciam o rendimento dos candidatos. Para isso, foram utilizados os dados disponibilizados pelo INEP, tratados por meio de processos de limpeza, exclusão de registros inconsistentes e padronização das informações. Complementarmente, foram adicionadas variáveis externas, como o Índice de Desenvolvimento Humano (IDH) dos estados e a classificação das localidades em zonas urbanas ou rurais. A análise exploratória, realizada com apoio de bibliotecas em Python, incluiu a construção de matrizes e gráficos de correlação, revelando que a renda familiar (Q006) e o acesso à internet (Q024) se destacaram como

ISSN: 2357-8645

variáveis mais associadas ao desempenho. O modelo preditivo foi desenvolvido por meio de algoritmos de regressão, com divisão em conjuntos de treino e teste, visando avaliar a influência desses fatores nas notas. Como resultados parciais, destaca-se a consolidação de uma base limpa e integrada, além da identificação de variáveis críticas que embasam as próximas etapas. Conclui-se que a preparação rigorosa dos dados é fundamental para análises confiáveis e que o estudo tem potencial de contribuir para a compreensão das desigualdades educacionais e para a formulação de políticas públicas mais eficazes.

Palavras-chave: Inteligência Artificial; Análise socioeconômica; Análise regional; Modelagem.

INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) consolidou-se como principal mecanismo de acesso ao ensino superior no Brasil, influenciando diretamente o ingresso de milhões de estudantes em universidades públicas e privadas. Compreender os fatores que afetam o desempenho dos candidatos tornou-se, portanto, essencial para analisar desigualdades educacionais e propor políticas de enfrentamento. Estudos indicam que aspectos socioeconômicos, como renda familiar e escolaridade dos pais, exercem forte influência sobre os resultados obtidos no exame (SANTOS, 2018; JUSTINIANO; QUEIROZ, 2018). A análise espacial do município de São Paulo evidencia que estudantes oriundos de famílias com maior poder aquisitivo tendem a apresentar desempenho superior, refletindo o acesso mais amplo a recursos pedagógicos e condições de estudo mais favoráveis (JUSTINIANO; QUEIROZ, 2018). Além disso, desigualdades regionais demonstram que a localidade de residência impacta o desempenho, mostrando que fatores estruturais e contextuais se cristalizam via diferenças socioeconômicas (SANTOS, 2018).

A evidência de padrões espaciais sugere que políticas públicas devem considerar não apenas a melhoria do acesso e da qualidade das escolas, mas também intervenções direcionadas a regiões e grupos mais vulneráveis, de modo a reduzir disparidades históricas no desempenho educacional. A consolidação de dados microeconômicos e espaciais permite, assim, compreender a complexidade das desigualdades no ENEM e fornece subsídios para estratégias mais eficazes de equidade educacional (OASISBR, 2018).

Por fim, o objetivo deste trabalho consiste em modelar a relação entre a renda per capita familiar e o desempenho dos estudantes no ENEM 2023, observando as variações entre as diferentes regiões do Brasil, por meio do uso de modelos de inteligência artificial. De forma específica, busca-se: (i) levantar e tratar os dados do ENEM 2023 e sociodemográficos; (ii) realizar análise

ISSN: 2357-8645

exploratória para identificar variáveis críticas; e (iii) aplicar modelos preditivos de regressão para avaliar a influência dos fatores socioeconômicos e regionais;

METODOLOGIA

A metodologia deste trabalho foi divida em 5 etapas, sendo elas: (i) Coleta de dados; (ii) Análise Exploratória; (iii) Complementação e tratamento de dados; (iv) Treinamento do Modelo de IA; e (v) Análise de resultados. A Figura 1 exemplifica o método.

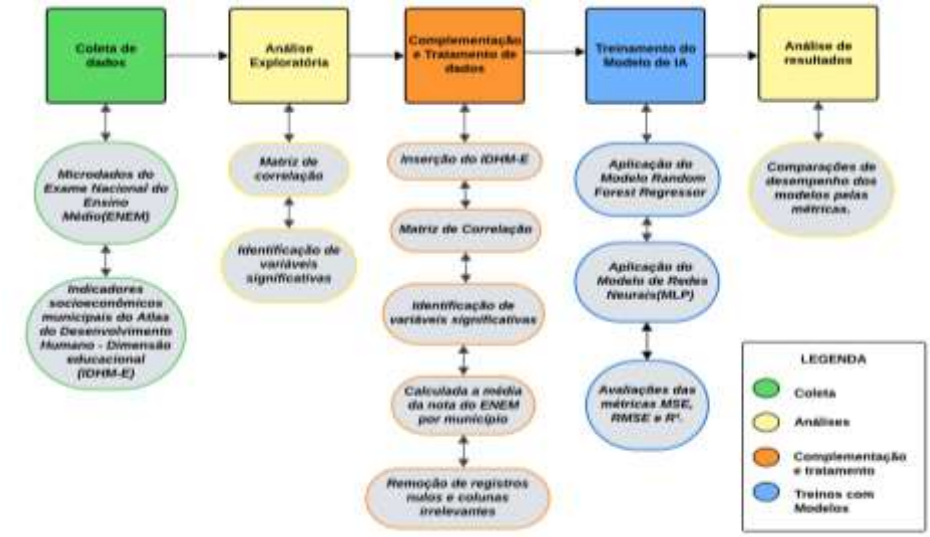

Figura 1 – Proposta metodológica

1. Captura e tratamento de dados

Para esta pesquisa, foram utilizados dados de duas fontes principais: (i) Os microdados do Enem que incluem informações detalhadas sobre provas, gabaritos, itens, notas individuais e questionários respondidos pelos candidatos, permitindo análises de desempenho educacional. (ii) Os dados do Atlas fornecem cerca de 200 indicadores para cada município, abrangendo dimensões de demografia, educação, renda, trabalho, habitação e vulnerabilidade social.

A primeira etapa consistiu na análise preliminar dos microdados do Enem. Investigou-se a correlação entre as variáveis (Figura 2) dos questionários socioeconômicos dos candidatos e suas notas, destacando-se a renda familiar (variável Q006) e o acesso à internet (variável Q024) como os fatores mais fortemente associados ao desempenho, enquanto as demais variáveis disponíveis não foram consideradas nesta análise.

ISSN: 2357-8645

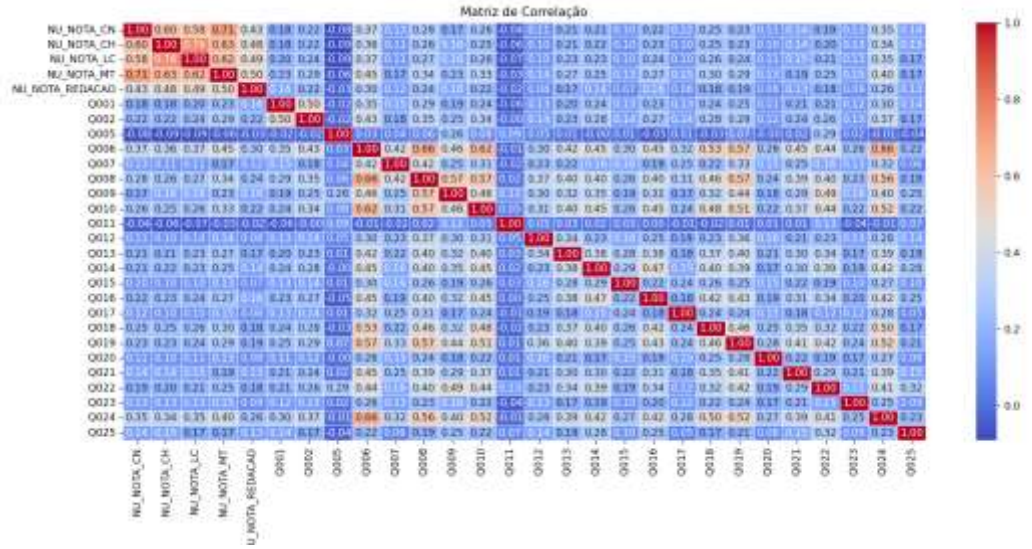

Figura 2 - Matriz de correlação entre os pontos Q006 e Q024 com os indicativos da nota do Enem

A análise foi complementada com dados do Atlas do Desenvolvimento Humano, relacionando indicadores municipais ao Índice de Desenvolvimento Humano Municipal - Dimensão Educação (IDHM-E). A partir de correlações, identificaram-se os dez fatores com maior impacto positivo e os dez com maior impacto negativo, totalizando vinte variáveis socioeconômicas relevantes para o desempenho educacional.

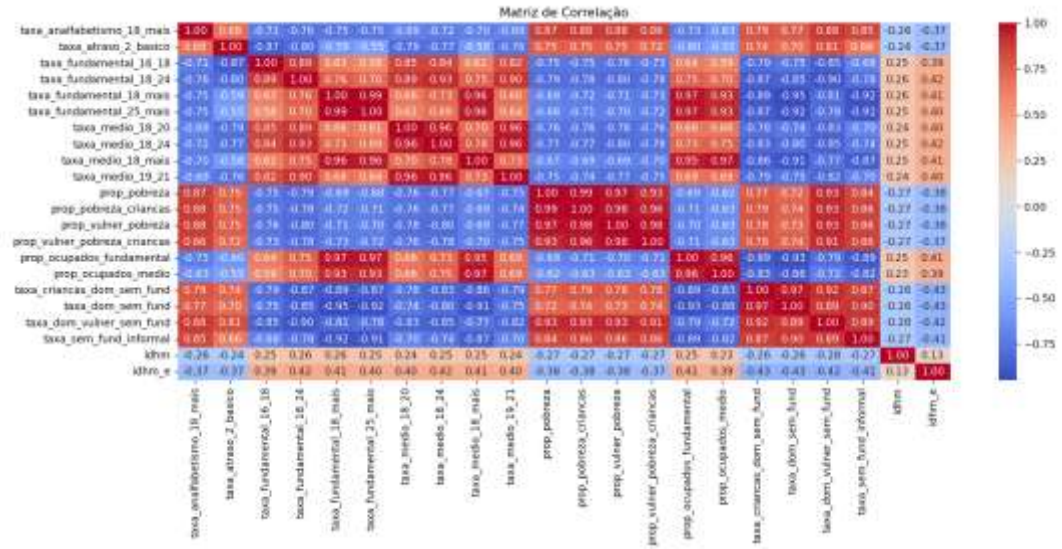

Figura 3 - Matriz de correlação demonstrando os indicadores mais importantes no IDHM-E

Posteriormente, calculou-se a média da nota do Enem por município e ela foi combinada com os indicadores socioeconômicos selecionados, formando uma tabela única com cada município representado por seus indicadores e a nota média do exame. Registros com valores nulos ou sem correspondência entre as bases foram eliminados, assegurando a consistência do conjunto de dados final.

ISSN: 2357-8645

As ferramentas utilizadas para o processamento e análise de dados incluíram *Python* com bibliotecas para manipulação de dados (pandas), visualização (*matplotlib, seaborn*) e cálculo de correlações.

2. Treinamento utilizando os modelos Random Forest e Redes Neurais

Com uma base de indicadores socioeconômicos municipais e médias do Enem, aplicaram-se técnicas de aprendizado de máquina para prever o desempenho médio dos alunos em cada município. Foram utilizados dois modelos, *Random Forest Regressor* e Redes Neurais (MLP Regressor), implementados em Python com a biblioteca *scikit-learn*.

Antes do treinamento, a tabela foi tratada para remoção de valores nulos e inconsistentes. As variáveis explicativas incluíram indicadores socioeconômicos previamente selecionados (20 variáveis), enquanto a variável alvo foi a média da nota do Enem por município. Para o modelo de redes neurais, os dados foram escalonados utilizando ***StandardScaler***, procedimento essencial para o desempenho das redes neurais, enquanto para o *Random Forest* esse passo não foi necessário.

O modelo de *Random Forest* consistiu em uma floresta de 500 árvores de decisão (**n_estimators=500**) com profundidade máxima limitada a 15 (**max_depth=15**) e critérios de divisão mínimos (**min_samples_split=5**, **min_samples_leaf=3**). A divisão entre treino e teste foi de 80%-20%. Este modelo é particularmente robusto a variáveis correlacionadas e permite capturar relações não lineares sem necessidade de escalonamento dos dados.

O modelo de rede neural utilizado foi um Multi-Layer Perceptron (MLP) com duas camadas ocultas, sendo a primeira de 100 neurônios e a segunda de 50 neurônios (**hidden_layer_sizes=(100, 50)**), função de ativação ReLU e otimizador Adam. O treinamento foi configurado para até 1000 iterações (**max_iter=1000**) com divisão de dados 80%-20%. O escalonamento dos dados foi realizado previamente, pois redes neurais são sensíveis à magnitude das variáveis.

RESULTADOS E DISCUSSÃO

Os modelos de aprendizado de máquina aplicados apresentaram desempenhos semelhantes na previsão da média das notas do ENEM por município. O modelo de *Random Forest* obteve um R^2 de 0,270, enquanto o modelo de MLP alcançou um R^2 de 0,270, indicando que

ISSN: 2357-8645

aproximadamente 27% da variabilidade da média das notas foi explicada por ambos os modelos. Ambos os modelos apresentaram erro quadrático médio (MSE) em torno de 1662 (*Random Forest*) e 1663 (Redes Neurais) e erro quadrático médio da raiz (RMSE) próximo a 40,77, sugerindo um desempenho similar.

Estudos anteriores também indicam que o Random Forest é eficaz na análise de dados educacionais. Por exemplo, um estudo comparativo revelou que o *Random Forest* apresentou desempenho superior ao da regressão linear múltipla na previsão de concentrações de neuroquímicos, destacando sua robustez e capacidade de lidar com dados complexos (GRIMM *et al*., 2013).

A análise de correlação entre os indicadores socioeconômicos municipais e o IDHM-E revelou que variáveis como taxa de analfabetismo, taxa de escolarização e renda per capita possuem forte correlação com o desempenho educacional. Esses achados corroboram estudos que destacam a importância de fatores socioeconômicos no desempenho escolar. Por exemplo, um estudo identificou que a renda familiar e o grau de escolaridade materna são determinantes significativos para o desempenho dos alunos (PPUFU, 2024).

Além disso, a análise revelou que municípios com menores taxas de analfabetismo e maior escolarização apresentam melhores desempenhos educacionais. Esses resultados estão alinhados com dados nacionais que indicam avanços na escolaridade média da população brasileira. Em 2024, a média de anos de estudo das pessoas de 25 anos ou mais atingiu 10,1 anos, o maior valor da série histórica (IBGE, 2024).

Os resultados obtidos sugerem que políticas públicas focadas na melhoria dos indicadores socioeconômicos, como aumento da escolaridade e redução do analfabetismo, podem ter um impacto positivo no desempenho educacional. Além disso, a utilização de modelos de aprendizado de máquina pode auxiliar na identificação de municípios com maior necessidade de intervenção, permitindo a alocação mais eficiente de recursos.

CONSIDERAÇÕES FINAIS

O estudo evidencia que fatores socioeconômicos municipais influenciam significativamente o desempenho médio no Enem, com municípios que apresentam maior escolarização e menor taxa de analfabetismo alcançando notas mais elevadas. Modelos de aprendizado de máquina, como *Random Forest* e redes neurais MLP, mostraram desempenho similar na previsão das

notas, explicando aproximadamente 27% da variabilidade observada. Entre os fatores mais determinantes destacam-se renda familiar, acesso à internet e taxa de escolarização, permitindo identificar municípios com maior necessidade de intervenção educacional e subsidiar políticas públicas direcionadas.

Entre as limitações da pesquisa estão o uso de dados agregados, que podem ocultar diferenças internas dentro dos municípios, e a ausência de variáveis relacionadas à infraestrutura escolar e políticas locais. Estudos futuros podem se beneficiar da análise de dados em nível escolar ou individual, além de incorporar indicadores adicionais. Também há a possibilidade de explorar outras técnicas de aprendizado de máquina ou combinações de modelos, visando aprimorar a capacidade de previsão do desempenho educacional e oferecer uma visão mais detalhada das desigualdades.

REFERÊNCIAS

IBGE. PNAD Contínua: Educação 2024. Rio de Janeiro: IBGE, 2024. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv102180_informativo.pdf. Acesso em: 18 set. 2025.

PPUFU. Programa de Pós-Graduação da Universidade Federal de Uberlândia. Determinantes socioeconômicos do desempenho escolar: renda familiar e escolaridade materna. Uberlândia: PPUFU, 2024.

GRIMM, R.; STEIN, N.; HENNIG, C.; LOPES, A. Random Forests versus Multiple Linear Regression in the Prediction of Neurochemical Concentrations. *BMC Bioinformatics*, v. 14, n. 1, p. 1–11, 2013. DOI: https://doi.org/10.1186/1471-2105-14-1.

JUSTINIANO, A.; QUEIROZ, M. Renda, participação e desempenho no ENEM em São Paulo: uma abordagem espacial (2012-2018). *OpenEdition Journals*, 2018. Disponível em: https://journals.openedition.org/confins/38804. Acesso em: 19 set. 2025.

SANTOS, G. Q. Os efeitos das desigualdades regionais nos resultados do ENEM: uma análise a partir dos microdados de 2018. Trabalho de Graduação, Universidade Federal do Rio Grande do Sul, 2018. Disponível em: https://lume.ufrgs.br/handle/10183/205590. Acesso em: 19 set. 2025.

INEP. *Microdados do ENEM*. Brasília: INEP, 2023. Disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem. Acesso em: 18 set. 2025.