Autor: Jeferson Magalhães dos Santos
GitHub | Email

Desenvolvimento e aplicação de técnicas de inteligência artificial para estimar o nível de influência das questões socioeconômicas no desempenho dos candidatos do ENEM 2021

RESUMO

Este relatório apresenta uma análise preditiva utilizando dados do Exame Nacional do Ensino Médio (ENEM) 2021, com o objetivo de investigar a influência das questões socioeconômicas no desempenho dos candidatos. Para isso, foram empregadas técnicas de inteligência artificial e algoritmos de aprendizado de máquina, incluindo Regressão Linear, *Random Forest* e Redes Neurais, para desenvolver modelos preditivos.

Os resultados obtidos revelaram que o modelo de redes neurais demonstrou o melhor desempenho, utilizando métricas como R^2 e Erro Médio Quadrado para avaliar a precisão das previsões. Ao explorar as características socioeconômicas dos candidatos, observou-se que a renda total da família despontou como o fator de maior influência no desempenho, seguida pela presença de computador na residência e o número de pessoas que residem com o candidato.

1. INTRODUÇÃO

De acordo com o Ministério da Educação, o ENEM visa avaliar o desempenho dos estudantes ao fim da escola básica. No entanto, o ENEM também pode ser usado como uma fonte valiosa de informações para entender como fatores socioeconômicos afetam a qualidade do aprendizado dos estudantes do Ensino Médio. Isso é possível porque o exame atrai muitos estudantes de diferentes grupos sociais e econômicos. Na última edição, em 2019, mais de 3 milhões de pessoas se inscreveram no ENEM (INEP, 2021), fornecendo uma amostra representativa dos estudantes que desejam ingressar no ensino superior. Além disso, todos os candidatos são obrigados a responder a um questionário com uma série de perguntas no momento da inscrição. Isso resulta em 136 informações diferentes para cada candidato, incluindo informações básicas como idade, sexo e município de residência, bem como detalhes como escolaridade dos pais, número de banheiros em casa e acesso à internet.

O ENEM foi criado originalmente para avaliar o desempenho dos estudantes no final da educação básica. Depois da reformulação em 2009, o exame passou a ser realizado não apenas como uma forma de avaliação, mas também por estudantes que desejam uma vaga em programas de Educação Superior, como o Sistema de Seleção Unificada (SISU), o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento ao Estudante do Ensino Superior (Fies). Vários trabalhos foram produzidos sobre o ENEM ao longo de sua história, e o INEP tem produzido diversos documentos e bases de dados que mostram os resultados e análises do exame. Apesar disso, a maioria dos trabalhos sobre o ENEM se concentra em classificar alunos e instituições, em vez de criar mecanismos ou descobrir insights que possam ajudar os professores em sua prática diária. Este relatório busca descobrir insights que possam ser usados para melhorar a qualidade da educação nacional, usando os microdados do ENEM 2021 e aplicando técnicas de Aprendizado de Máquina para prever resultados e capturar características relevantes dos participantes.

2. OBJETIVOS

- O objetivo deste relatório é desenvolver, modelar e executar técnicas de inteligência artificial, incluindo aprendizado de máquina e redes neurais, com base em um conjunto de dados de amostra dos participantes do ENEM 2021 em todo o Brasil. Buscaremos estimar o nível de influência das questões socioeconômicas no desempenho dos candidatos.

3. MATERIAIS E MÉTODOS

Os dados do ENEM estão disponíveis no site do INEP e podem ser baixados gratuitamente. Eles incluem provas, gabaritos, informações sobre itens, notas e questionários preenchidos pelos inscritos. Os dados vêm em uma pasta compactada que precisa ser descompactada antes de serem utilizados. A pasta contém cinco subpastas, cada uma com informações diferentes. A subpasta de dados contém dois arquivos do Excel com informações sobre as provas e os questionários preenchidos pelos participantes. A subpasta de dicionário de dados contém informações sobre as variáveis presentes nos dados, o que é importante para compreender a estrutura dos dados e como utilizá-los.

3.1 Pré-processamento e limpeza dos dados

Realizou-se um pré-processamento dos dados, removendo colunas desnecessárias para o objetivo deste relatório, a fim de reduzir o tamanho do arquivo original e o tempo de processamento. Durante a limpeza, observou-se uma grande quantidade de dados ausentes nas variáveis-alvos (notas das provas). Isso ocorre devido à abstenção dos candidatos nas respectivas provas. A abordagem adotada para esse problema foi a exclusão das linhas que continham dados ausentes.

3.2 Candidatos "treineiros"

Foram selecionados apenas os candidatos não treineiros para o treinamento dos modelos.

3.3 Agregações da variável-alvo

Para simplificar o treinamento dos modelos, foi criada uma nova variável-alvo que representa a média de todas as notas das provas dos candidatos (ciências da natureza, ciências humanas, matemática e linguagens).

3.4 Tratamentos de variáveis categóricas

Todas as variáveis categóricas foram transformadas em variáveis numéricas para o adequado treinamento do modelo.

3.5 Modelos de aprendizado de máquina utilizados

Foram construídos três modelos para abordar esse problema: regressão linear (Scikit-Learn), regressão por *Random Forest* (Scikit-Learn) e redes neurais (TensorFlow). O modelo de *Random Forest* foi configurado com uma profundidade máxima de 7 para cada árvore, enquanto outros hiperparâmetros foram mantidos com os valores padrão. A rede neural foi projetada da seguinte forma:

- Camada de entrada com 25 neurônios (1 para cada questão)
- Camada intermediária com 100 neurônios
- Camada intermediária com 25 neurônios
- Camada de saída com 1 neurônio (ativação linear devido às características do problema)
- Utilizou-se o otimizador 'ADAM' com um tamanho de lote e número de épocas igual a 5.

3.6 Métricas utilizadas para avaliação

As métricas utilizadas para avaliar o desempenho dos modelos foram o coeficiente de determinação (R^2), erro médio absoluto (MAE), erro médio quadrado (MSE) e raiz quadrada do erro médio (RMSE).

3.7 Técnicas de permutação

A fim de avaliar a influência das questões socioeconômicas, utilizamos o método de permutação (por meio da função 'permutation_importance' do pacote Sci-kit Learn). Essa técnica foi aplicada para analisar e mensurar a importância das variáveis presentes no conjunto de dados. O procedimento envolve a permutação aleatória dos valores de cada variável e a observação do impacto resultante nas métricas de avaliação do modelo. A diferença entre as métricas originais e as métricas obtidas a partir das permutações revela a importância relativa de cada variável para o desempenho do modelo. Com base nesse processo, obtivemos uma estimativa confiável da importância das variáveis, classificando-as de acordo com suas pontuações médias de importância. Essa análise nos permitiu identificar quais variáveis tiveram uma maior influência nas previsões do modelo, facilitando a tomada de decisões informadas em relação às características mais relevantes para a tarefa em questão.

4. RESULTADOS

Após aplicação das técnicas, obtemos as seguintes métricas avaliativas:

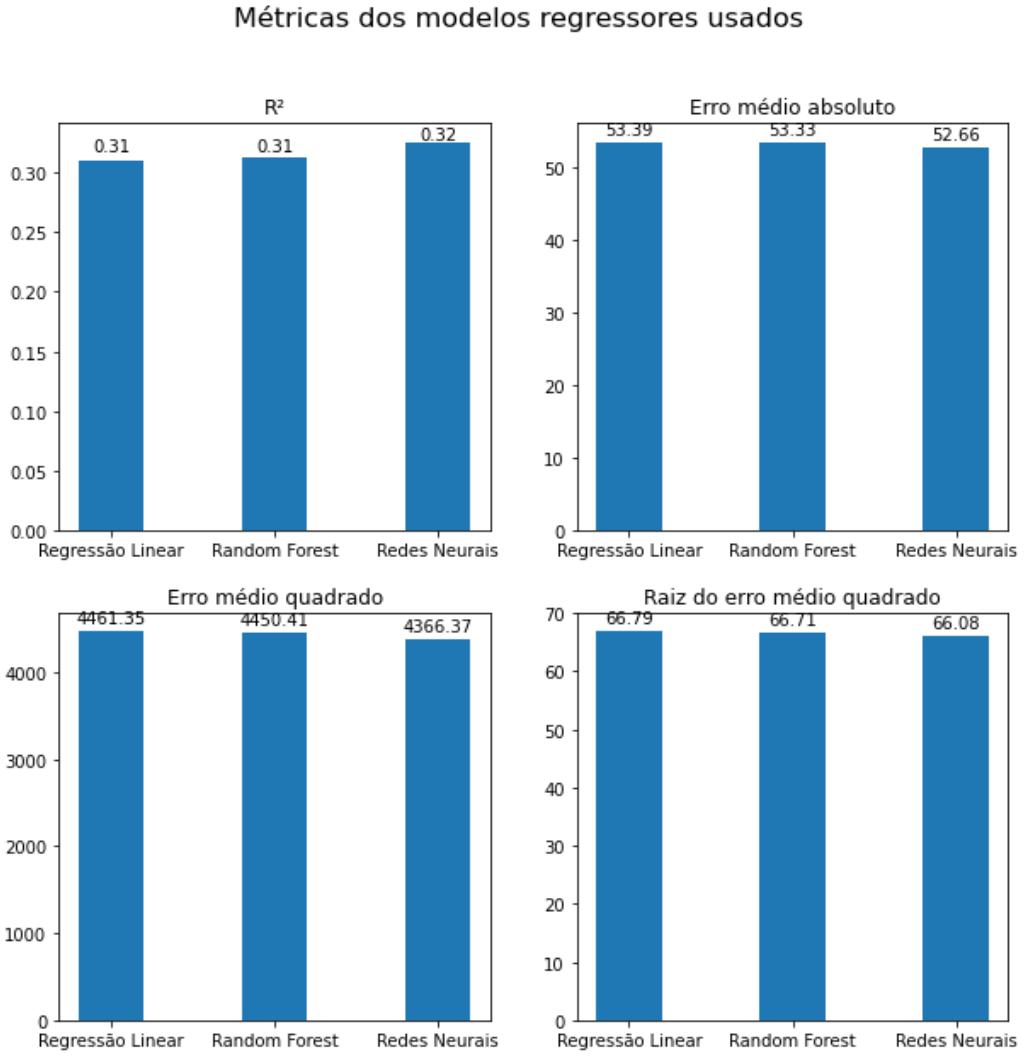

Figura 1 - Métricas avaliativas para os modelos aplicados – Autoria Própria

Foi observado que o modelo de redes neurais apresentou um melhor coeficiente de determinação (R^2) em comparação com a *Random Forest* e o modelo de regressão linear. Essa tendência também foi verificada nas métricas de erro, como o erro médio absoluto (MAE), erro médio quadrado (MSE) e raiz quadrada do erro médio (RMSE), em que os resultados foram melhores (ou seja, os erros foram menores) para o modelo de redes neurais. Com base nesses resultados, optou-se por utilizar o método de redes neurais para a análise das questões (*features*). Através desse modelo preditivo, é possível estimar a importância de cada variável

em relação à variável-alvo. Para isso, foram utilizados o R^2 e o RMSE na construção da permutação randômica.

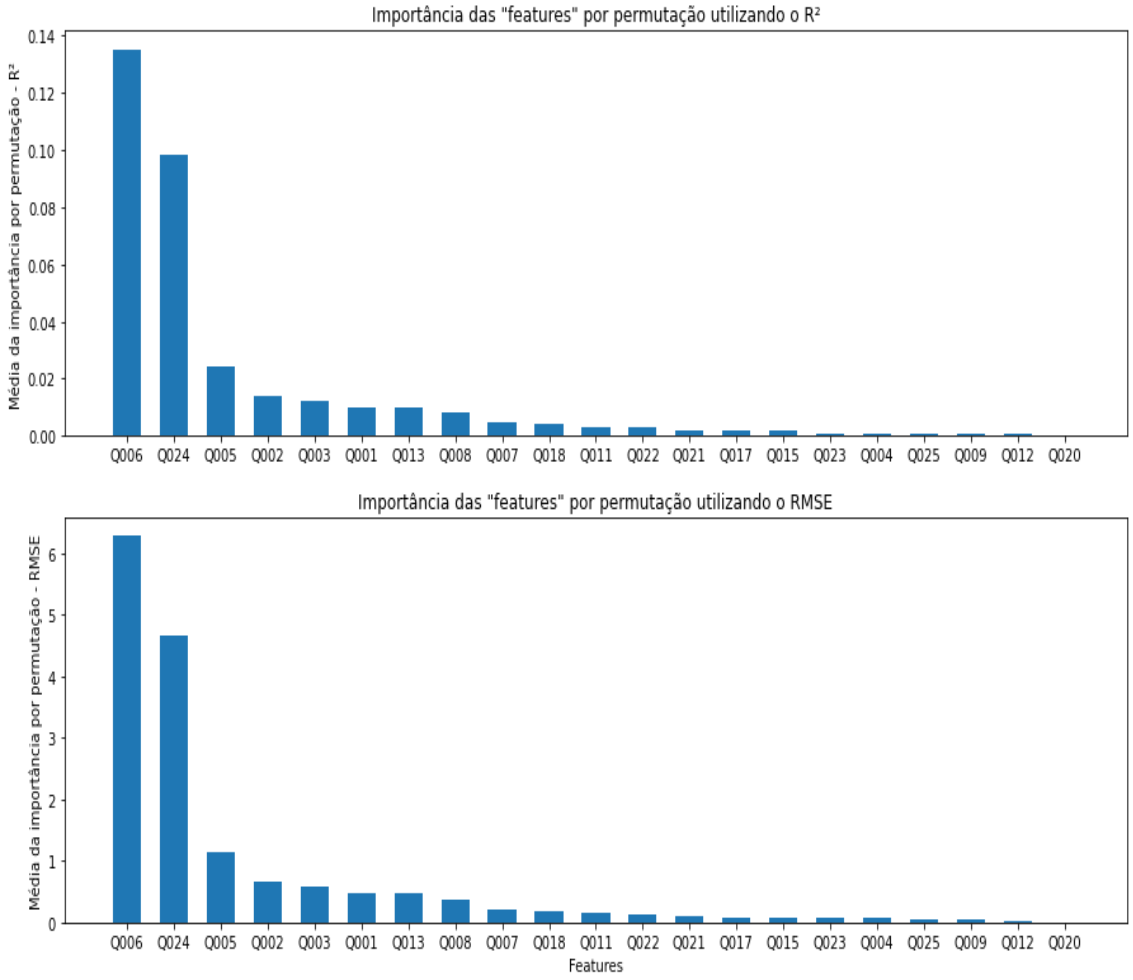

Figura 2 - Gráfico com a média da importância de cada Questão Socioeconômica - Autoria Própria

Ao analisar o gráfico com as métricas de permutação (Figura 2), é possível observar que as questões Q006 e Q024 (Figura 3) se destacam significativamente entre as variáveis consideradas. Essas duas questões em particular apresentaram uma importância considerável em relação à variável-alvo do modelo. Surpreendentemente, quando combinadas, essas duas questões representam aproximadamente 70% da importância total do modelo como um todo. Esse destaque ressalta a influência dessas questões específicas no desempenho dos candidatos, sugerindo que fatores socioeconômicos relacionados à renda familiar e à disponibilidade de recursos podem desempenhar um papel significativo na avaliação do ENEM.

Q006	Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
Q024	Na sua residência tem computador?

Figura 3 - Descrição das questões socioeconômicas 006 e 024 - Autoria Própria

Considerando as informações disponíveis sobre a distribuição das questões e a variável alvo no ENEM de 2021, realizamos uma análise para entender como essas questões se apresentam individualmente e em comparação com a Nota Média.

Ao analisarmos a variável "Renda Familiar" (Questão Q006), Figura 4, constatamos que cerca de 60% dos candidatos possui uma renda familiar abaixo de R$ 2200, sendo que, destes, a maioria possui uma renda inferior a R$ 1100. Esses dados revelam uma concentração significativa de candidatos com renda familiar mais baixa. Essa análise fornece uma visão importante sobre o perfil socioeconômico dos participantes do ENEM em 2021. Outra questão relevante é a disponibilidade de computador nas residências dos candidatos (Figura 5). Os resultados indicam que aproximadamente 40% da população não possui computador em sua residência. Essa informação revela a existência de uma parcela considerável de candidatos que podem enfrentar desafios adicionais em relação ao acesso à tecnologia e recursos digitais, o que influencia seu desempenho no exame.

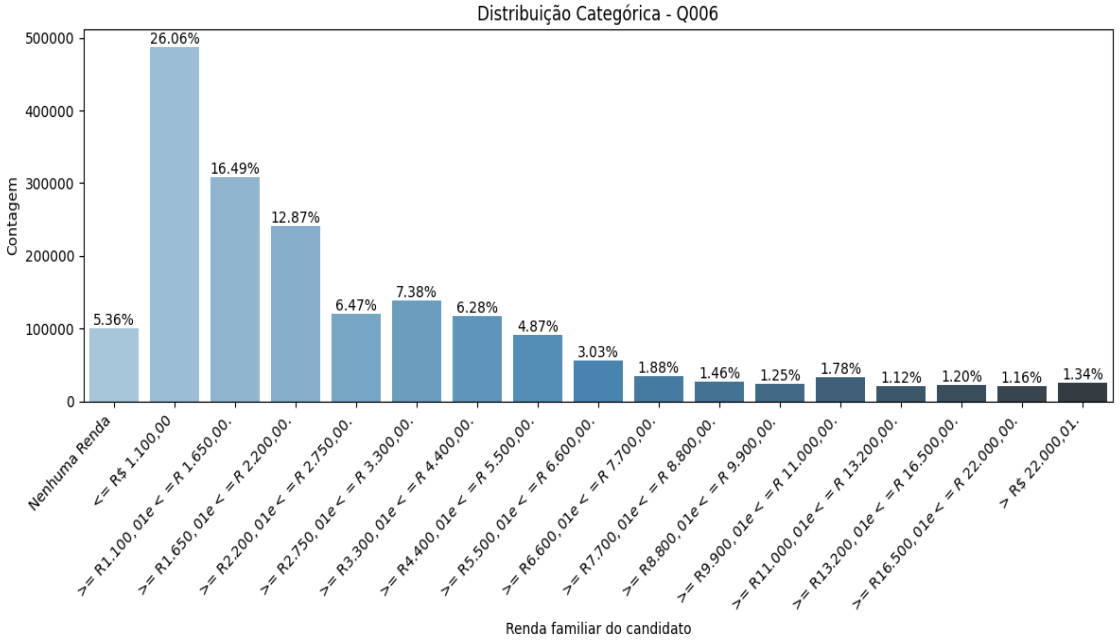

Figura 4 - Distribuição univariada da Questão 006- Autoria Própria

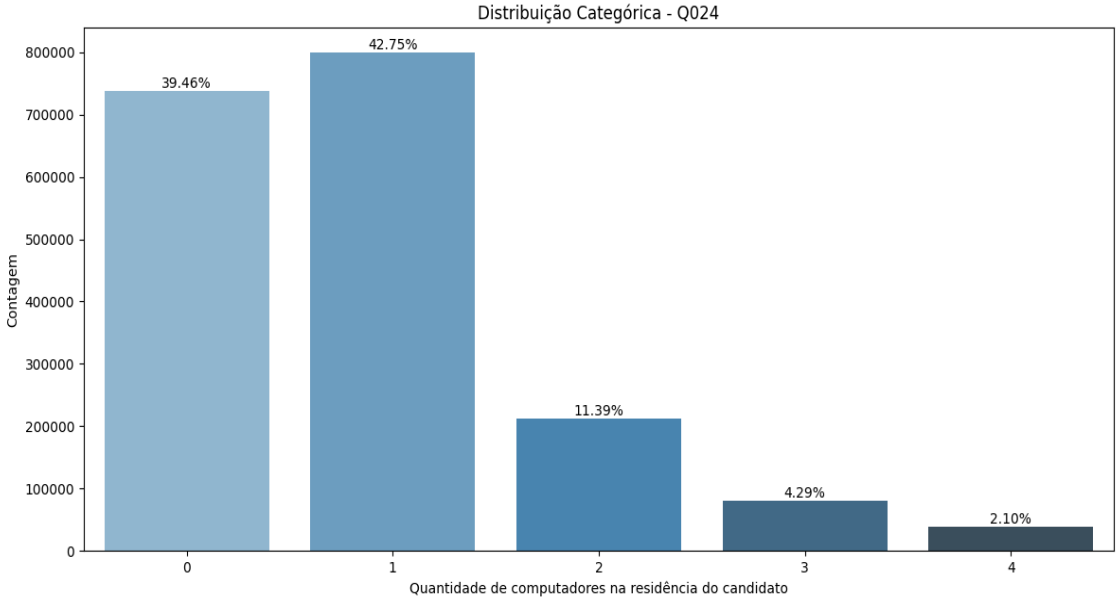

Figura 5 - Distribuição univariada Questão 024 - Autoria Própria

Observando a relação entre as questões socioeconômicas e a nota média dos candidatos, Figuras 6 e 7, podemos observar dados significativos sobre como esses fatores influenciam o desempenho dos estudantes. Em relação à questão da renda, fica evidente que quanto maior a renda familiar, melhor é o desempenho dos candidatos. O mesmo padrão é observado na questão do acesso a computadores, em que a presença de um maior número de computadores na residência está associada a um melhor desempenho na prova.

Essas análises preliminares destacam a importância de considerar os fatores socioeconômicos, como renda familiar e acesso à tecnologia, ao examinar os resultados e o desempenho dos candidatos no ENEM de 2021. Esses *insights* podem ser valiosos para compreender as disparidades existentes e fornecer o suporte adequado aos estudantes, visando melhorar a equidade no sistema educacional.

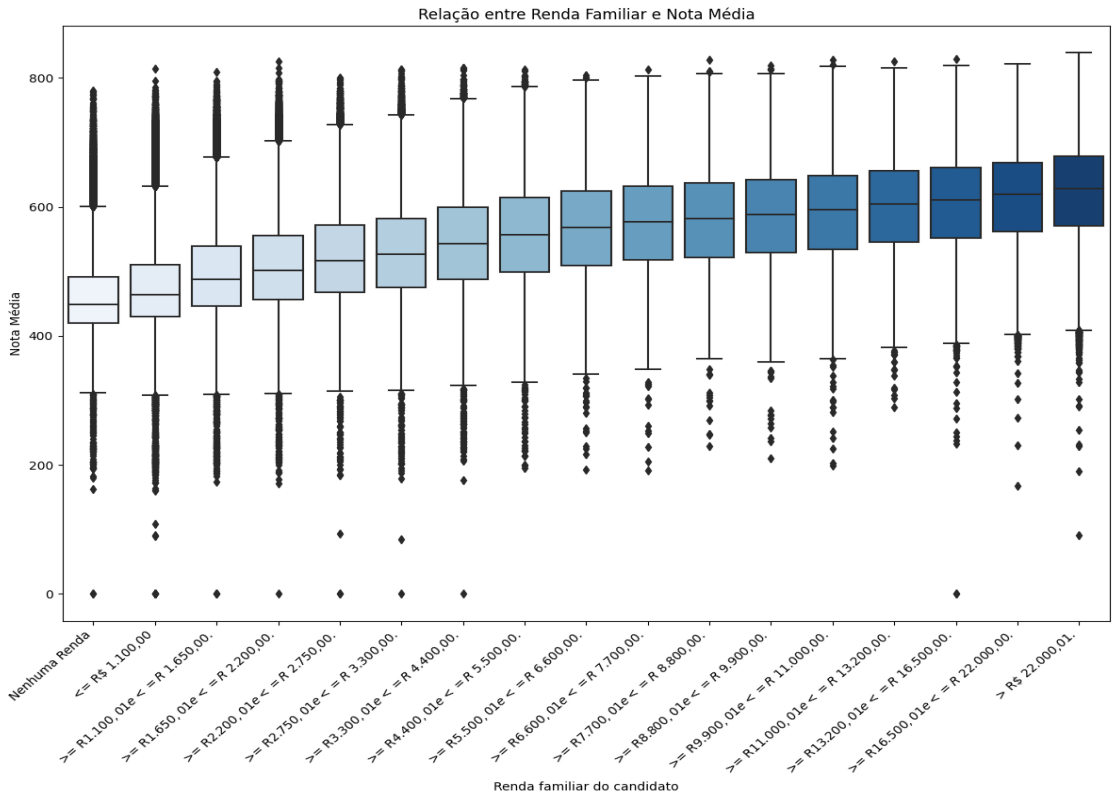

Figura 6 - Gráfico BoxPlot relacionando a questão de renda com a nota média - Autoria Própria

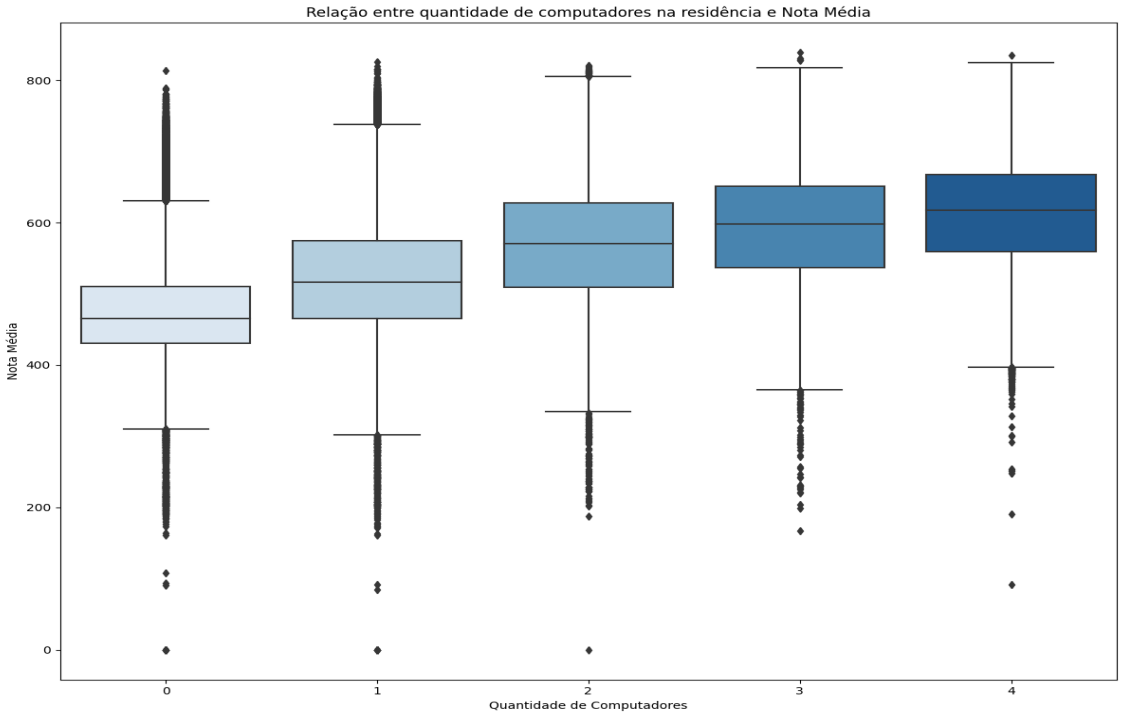

Figura 7 - Gráfico BoxPlot relacionando a questão de quantidade de computadores com a nota média - Autoria Própria

5. CONCLUSÃO

Através da construção dos modelos preditivos, foi possível identificar as questões socioeconômicas que exerceram maior influência no desempenho dos candidatos do ENEM no ano de 2021. Os resultados obtidos confirmaram que a renda familiar e a presença de computador na residência foram os principais fatores que impactaram o desempenho dos candidatos, representando aproximadamente 70% de influência no modelo. Esses resultados estão alinhados com referências utilizadas e com estudos anteriores realizados com dados de outros anos do ENEM. A consistência desses achados reforça a importância desses fatores socioeconômicos no desempenho dos estudantes em avaliações educacionais de grande escala.

Para trabalhos futuros, recomenda-se aprimorar o modelo, buscando parâmetros otimizados, bem como a exclusão de *features* que apresentem pouca relevância para o modelo preditivo. Além disso, a melhoria da arquitetura geral do modelo de Redes Neurais pode ser explorada, visando aumentar sua precisão e capacidade de generalização.

No contexto do objetivo deste relatório, o modelo utilizado mostrou-se adequado para investigar a influência das questões socioeconômicas no desempenho dos candidatos do ENEM 2021.

REFERÊNCIAS

- CASSIANI, S.; SILVA, H. D.; PIERSON, A. OLHARES PARA O ENEM NA EDUCAÇÃO CIENTÍFICA E TECNOLÓGICA. JUNQUEIRA & MARIN, 2016. ISBN 9788582030257.
- FARIAS, E. R.; MARCIO. USO DE DATA SCIENCE NA ANÁLISE DAS PROVAS DO ENEM. Disponível em: <[https://books.google.com.br/books?id=q7R2DwAAQBAJ](https://dspace.bc.uepb.edu.br/xmlui/bitstream/handle/123456789/26068/PDF%20-%20Marcio%20Edglaiton%20Rosa%20Farias?sequence=1&isAllowed=y)>.
- ROSAL, I. L. Iury Rosal. Disponível em: . Acesso em: 11 dez. 2022.
- gustavomccoelho. Prevendo Desempenho no ENEM através de Fatores Socioeconômicos e Métodos de Aprendizado de Máquina. Disponível em: . Acesso em: 11 dez. 2022.