DESIGUALDADES EDUCACIONAIS NO ENEM: UMA PERSPECTIVA BASEADA EM VARIÁVEIS SOCIOECONÔMICAS E APRENDIZAGEM DE MÁQUINA

https://doi.org/10.5902/2318133893251

Marcelo de Souza[1]
Daniel Larion Klug[2]

Resumo
O Exame Nacional do Ensino Médio representa um importante mecanismo de acesso ao ensino superior no Brasil. No âmbito desse estudo investigou-se a relação entre variáveis socioeconômicas e o desempenho dos estudantes no exame, utilizando técnicas de aprendizado de máquina para identificar padrões significativos. A pesquisa teve como objetivos desenvolver modelos preditivos baseados em random forest para classificar o desempenho dos estudantes; identificar as variáveis socioeconômicas mais relevantes e analisar seu impacto nos resultados, visando a subsidiar políticas educacionais mais equitativas. Foram adotados os microdados do Enem 2023, submetidos a uma etapa de pré-processamento que envolveu técnicas de one-hot encoding para tratamento de algumas variáveis e Smote para balanceamento. Foram construídos dez modelos de random forest, com o ajuste de hiperparâmetros via busca aleatória. O desempenho foi avaliado por métricas como acurácia, precisão, recall e F1-score, além da análise de importância das variáveis. Os modelos apresentaram desempenho satisfatório, com acurácias em torno de 94% e precisão de até 99%. A escolaridade e ocupação dos pais, junto com a renda familiar, emergiram como os principais preditores. Estudantes com pais mais escolarizados e em profissões estratégicas tiveram probabilidade três vezes maior de alto desempenho, enquanto aqueles de famílias de baixa renda apresentaram maior tendência a desempenho insatisfatório. Os resultados evidenciam a influência de fatores socioeconômicos no desempenho educacional, reforçando a necessidade de políticas públicas apropriadas. A eficácia dos modelos comprova sua utilidade para diagnósticos educacionais.
Palavras-chave: microdados do Enem; random forest; aprendizagem de máquina.

EDUCATIONAL INEQUALITIES IN ENEM: A PERSPECTIVE BASED ON SOCIOECONOMIC VARIABLES AND MACHINE LEARNING

Abstract
The National High School Exam serves as an important gateway to higher education in Brazil. This study examines the relationship between socioeconomic variables and student performance on the

[1] Universidade do Estado de Santa Catarina, Ibirama, Santa Catarina, Brasil. E-mail: marcelo.desouza@udesc.br. Orcid: https://orcid.org/0000-0002-0786-2127.
[2] Universidade do Estado de Santa Catarina, Ibirama, Santa Catarina, Brasil. E-mail: daniel.klug@edu.udesc.br. Orcid: https://orcid.org/0009-0004-1319-1181.

Critérios de autoria: os autores, coletivamente, realizaram a concepção, criação e consolidação do artigo.

Recebido em 14 de agosto de 2025. Aceito em 7 de outubro de 2025.

exam, employing machine learning techniques to identify significant patterns. The research has three main objectives to develop predictive models based on random forest to classify student performance to identify the most relevant socioeconomic variables and to analyze their impact on results, aiming to inform more equitable educational policies. The study used Enem 2023 microdata, which underwent preprocessing including one-hot encoding for certain variables and Smote for balancing. Ten random forest models were built, with hyperparameter tuning via random search. Performance was evaluated using metrics such as accuracy, precision, recall, and F1-score, along with variable importance analysis. The models demonstrated satisfactory performance, with accuracy around 94% and precision up to 99%. Parental education level, occupation, and family income emerged as key predictors. Students with more educated parents in strategic professions were three times more likely to achieve high performance, while those from low-income families showed greater tendency toward unsatisfactory results. The findings highlight the influence of socioeconomic factors on educational performance, underscoring the need for appropriate public policies. The models' effectiveness confirms their utility for educational diagnostics.
Key-words: Enem data; random forest; machine learning.

Introdução

O Exame Nacional do Ensino Médio – Enem – é uma prova realizada pelo Ministério da Educação para avaliar o desempenho escolar de estudantes concluintes da educação básica. Desde 2009 a nota do Enem passou a ser usada para acessar o Sistema de Seleção Unificada – Sisu – e o Programa Universidade para Todos – Prouni. Os participantes do Enem também podem pleitear financiamento estudantil em programas governamentais, como o Fundo de Financiamento Estudantil.

O MEC (2023) disponibiliza os microdados do Enem, com várias informações reunidas em uma plataforma Web aberta ao público geral. Através desses dados, é possível analisar indicadores nacionais dos candidatos do exame. Há quatro áreas de conhecimento a serem examinadas: Ciências da Natureza, Linguagens e Códigos, Ciências Humanas e Matemática. Os candidatos são avaliados por 45 questões de cada área, além de uma redação dissertativa-argumentativa, elaborada a partir de uma situação-problema apresentada.

Os microdados do Enem são ricos em informações. Além das respostas às questões e do desempenho dos candidatos, são apresentados dados socioeconômicos, obtidos pelo preenchimento de um questionário na ficha de inscrição. Esses dados têm valor informativo, pois podem revelar padrões e tendências que ajudam a compreender os fatores que afetam o desempenho dos candidatos. No entanto, parte deste potencial informativo permanece inexplorado, limitando as possibilidades de examinar a relação entre o desempenho e o contexto socioeconômico dos candidatos. Uma análise aprofundada pode revelar a influência de alguns fatores no desempenho do candidato, como a renda familiar, a escolaridade dos pais e o acesso a recursos educacionais.

Este trabalho apresenta um estudo do desempenho dos candidatos do Enem com base nos seus dados socioeconômicos. A pesquisa explorou modelos de aprendizagem de máquina para predição do desempenho dos candidatos, permitindo não só predizer candidatos com alto ou baixo desempenhos, mas identificar as variáveis socioeconômicas associadas a esses resultados. Em particular, são explorados modelos de random forest, que se mostraram eficazes para a tarefa de interesse, atingindo valores superiores a 90% para diversas medidas de desempenho, como acurácia, precisão e recall.

A relevância deste estudo está em fornecer uma compreensão mais profunda dos fatores socioeconômicos que impactam no desempenho dos candidatos no Enem. São fornecidas ferramentas capazes de identificar tendências e padrões nessas variáveis, e sua relação com as notas obtidas. Tais achados têm o potencial de contribuir com políticas educacionais mais eficientes e inclusivas. Ao identificar os fatores que mais influenciam no desempenho nas provas, é possível direcionar esforços para mitigar desigualdades, promovendo iniciativas que potencializem o sucesso dos estudantes de diferentes contextos socioeconômicos. A solução proposta responde a uma demanda educacional e social, contribuindo para um cenário de maior igualdade na educação brasileira.

Metodologia

O trabalho foi estruturado em quatro etapas. A primeira consistiu na coleta dos dados e sua análise inicial, com o objetivo de entender as informações disponíveis e sua estrutura. Na segunda etapa, foram conduzidas tarefas de pré-processamento dos dados, de modo a prepará-los para a aplicação de modelos preditivos. Foram criadas novas variáveis com informações de interesse, enquanto outras variáveis foram transformadas. A terceira etapa consistiu na construção e avaliação dos modelos preditivos. Finalmente, a última etapa explorou os modelos construídos, extraindo as variáveis com maior contribuição para a predição. A partir desses resultados, foi realizada uma análise exploratória dos dados, identificando a relação entre as variáveis e o desempenho dos candidatos.

Coleta e pré-processamento dos dados

Foram coletados dados da edição de 2023 do Enem realizado no estado de Santa Catarina. Esses dados foram divididos em três grupos: dados gerais do candidato; desempenho do candidato; dados socioeconômicos. Os dados gerais do candidato são faixa etária, gênero, estado civil, nacionalidade, situação de conclusão do Ensino Médio – concluído, conclusão prevista para o ano atual, conclusão prevista para após o ano atual ou não concluído nem cursando –, ano de conclusão do Ensino Médio, tipo de escola do Ensino Médio, pública ou privada, tipo de ensino do Ensino Médio, regular ou especial, e treineiro, sim ou não. Também são incluídos nesse grupo os dados relacionados ao exame prestado, com o município e o Estado de realização da prova. Os dados de desempenho do candidato são as cinco notas obtidas por ele em cada uma das quatro áreas – Ciências da Natureza, Linguagens e Códigos, Ciências Humanas e Matemática – e na redação.

Os dados socioeconômicos são compostos pelas respostas do candidato a um questionário aplicado no momento da inscrição no Enem. Detalhes das questões são apresentados na tabela 1. Esse grupo reúne dados sobre o grau de instrução e ocupação dos pais ou responsáveis do candidato, o tamanho da família e a renda familiar, bem como dados relacionados à estrutura da residência e ao acesso do candidato a recursos, como carro, computador e Internet. O uso dessas informações para predizer o desempenho do candidato é relevante, dado que a situação socioeconômica pode interferir no desempenho dos estudantes no exame.

Algumas variáveis relacionadas ao questionário socioeconômico foram ajustadas para converter alternativas categóricas ou alfanuméricas em valores numéricos representativos. Por exemplo, na variável de faixa etária idade, o ajuste atribui o valor máximo da faixa etária selecionada para validação. Da mesma forma, na variável que indica a faixa de renda

familiar, é utilizado o valor máximo da faixa declarada. Esse procedimento também é aplicado em questões relacionadas à quantidade de itens que a família possui em casa, garantindo uma representação numérica consistente dos dados.

Para questões cujas respostas não representam valores numéricos específicos, como gênero, estado civil ou perguntas com respostas do tipo sim ou não, aplica-se a técnica de One Hot Encoding (Seger, 2018). Esse método converte cada alternativa numa variável binária, atribuindo os valores 0 e 1 para indicar se uma opção foi selecionada ou não, respectivamente. Isso permite que variáveis categóricas sejam representadas de forma compatível com os algoritmos de aprendizagem de máquina. Finalmente, variáveis numéricas foram normalizadas, de tal forma que seus valores tenham média igual a 0 e desvio padrão igual a 1.

Quadro 1 –
Dados socioeconômicos oriundos do questionário com os candidatos.

Questão	Descrição
Q01	Até que série seu pai, ou o homem responsável por você, estudou?
Q02	Até que série sua mãe, ou a mulher responsável por você, estudou?
Q03	Qual a ocupação do seu pai ou do homem responsável por você?
Q04	Qual a ocupação da sua mãe ou da mulher responsável por você?
Q05	Incluindo você, quantas pessoas moram atualmente em sua residência?
Q06	Qual é a renda mensal de sua família?
Q07	Em sua residência trabalha empregado doméstico?
Q08	Na sua residência tem banheiro?
Q09	Na sua residência tem quartos para dormir?
Q10	Na sua residência tem carro?
Q11	Na sua residência tem motocicleta?
Q12	Na sua residência tem geladeira?
Q13	Na sua residência tem freezer (independente ou segunda porta da geladeira)?
Q14	Na sua residência tem máquina de lavar roupa?
Q15	Na sua residência tem máquina de secar roupa?
Q16	Na sua residência tem forno micro-ondas?
Q17	Na sua residência tem máquina de lavar louça?
Q18	Na sua residência tem aspirador de pó?
Q19	Na sua residência tem televisão em cores?
Q20	Na sua residência tem aparelho de DVD?
Q21	Na sua residência tem TV por assinatura?
Q22	Na sua residência tem telefone celular?
Q23	Na sua residência tem telefone fixo?
Q24	Na sua residência tem computador?
Q25	Na sua residência tem acesso à Internet?

Fonte: autores (2025).

Além disso, foram criadas novas variáveis associadas a cada prova e à redação. Essa variável indica se o candidato obteve alto desempenho, baixo desempenho ou desempenho regular para aquela nota. O alto desempenho é conferido aos candidatos com nota entre as 10% melhores. O baixo desempenho é conferido aos candidatos com nota entre as 10% piores. Os demais candidatos são considerados de desempenho regular.

No cenário proposto, existe um desbalanceamento das classes, pois a maioria dos candidatos apresentam desempenho regular. Essa situação pode introduzir vieses nos modelos preditivos, que podem favorecer a predição de desempenho regular em detrimento dos demais, pelo simples fato de ser a classe mais frequente no conjunto de dados. Para tratar esse problema, foram aplicadas técnicas de rebalanceamento de dados (He; Garcia, 2009; Krawczyk, 2016). Para tanto, foi adotado o método Smote (Chawla et al., 2002; Fernández et al., 2018), que cria novas amostras das classes menos frequentes – classes minoritárias –, garantindo que a base de dados contenha o mesmo número de exemplos para todas as classes.

Construção e avaliação dos modelos preditivos

Este trabalho trata da predição do desempenho do candidato de forma independente para cada nota, isto é, prediz o desempenho específico em cada prova e na redação. Neste sentido, foram criados diferentes modelos de classificação. Para cada nota, há um modelo que classifica o candidato como alto desempenho ou não, e um segundo modelo que classifica o candidato como baixo desempenho ou não. Como existem cinco notas, foram elaborados dez modelos de classificação. Todos os modelos são baseados em random forest (Breiman, 2001). Esses modelos usam os dados socioeconômicos e de desempenho, excluindo os dados gerais do candidato e local de prova.

Para a elaboração desses modelos, os dados foram divididos nos conjuntos de treinamento e teste, contendo dois terços e um terço dos dados, respectivamente. O conjunto de treinamento é usado para a construção dos modelos. Nesse processo, é feito o ajuste dos hiperparâmetros dos modelos de random forest, usando o algoritmo de busca aleatória (Bergstra e Bengio, 2012) com validação cruzada com cinco partições. Os hiperparâmetros ajustados foram o número de árvores, com possíveis valores {50, 100, 150, 200}; profundidade máxima das árvores, com possíveis valores {0, 10, 20, 30}; número mínimo de amostras para dividir um nó, com possíveis valores {2, 5, 10}; número mínimo de amostras para formar uma folha, com possíveis valores {1, 2, 4}; amostragem com reposição, com possíveis valores {verdadeiro, falso}. Ao final desse processo, foi selecionado o modelo que apresentou a melhor acurácia. Esse modelo é então avaliado no conjunto de teste, medindo seu desempenho final.

Análise das variáveis mais importantes

Uma vez elaborados e validados os modelos preditivos, eles fornecem informações relevantes sobre a tarefa de interesse e o conjunto de dados. Para avaliar as variáveis mais relevantes para classificar o desempenho do candidato, ou seja, se ele possui desempenho alto, baixo ou regular, as dez variáveis mais importantes de cada modelo são identificadas.

Os modelos de random forest calculam a importância de cada variável para o modelo pela sua capacidade de discriminar os dados. Para definir a importância geral de uma variável, é calculada a média da sua importância para os dez modelos construídos.

Para verificar a relação entre as variáveis identificadas como mais importantes e o desempenho dos candidatos, foram geradas visualizações dos dados, mostrando o percentual de candidatos com desempenho alto, baixo e regular para os diferentes valores possíveis dessas variáveis. Dessa forma, é possível identificar relações entre a variável socioeconômica e o desempenho médio dos candidatos.

Resultados e discussão

Os modelos que predizem alto desempenho foram avaliados para cada nota. Ou seja, para predizer se candidatos apresentam alto desempenho em cada uma das provas e na redação. Os resultados foram avaliados analisando métricas de acurácia, precisão, recall e F1-score. A tabela 1 mostra o desempenho preditivo desses modelos. Pode-se observar que os valores de precisão estão entre 96% e 98%, demonstrando que a maior parte das amostras cuja predição foi positiva, isto é, cuja predição indica alto desempenho, é, de fato, positiva. Logo, entre 2% e 4% de amostras com essa classificação se tratam de falsos positivos. Para os valores de recall, a avaliação indica que o modelo conseguiu identificar entre 92% e 93% das amostras positivas. Finalmente, um valor entre 94% e 95% para a métrica F1-score, tanto para o alto desempenho quanto para o não alto desempenho mostra um equilíbrio adequado entre precisão e recall. Com a acurácia dos modelos entre 94% e 95%, o modelo se mostra eficaz em predizer candidatos com alto desempenho.

Tabela 1 –
Avaliação dos modelos preditivos de alto desempenho.

Prova	Alto desempenho			Não alto desempenho			Acurácia
	Precisão	Recall	F1-score	Precisão	Recall	F1-score	
Ciências da Natureza	0,96	0,93	0,94	0,93	0,97	0,95	0,95
Ciências Humanas	0,97	0.92	0,94	0,92	0,97	0,95	0,95
Linguagens e Códigos	0,97	0,92	0,94	0,92	0,97	0,94	0,94
Matemática	0,97	0,93	0,95	0,93	0,97	0,95	0,95
Redação	0,98	0,92	0,95	0,92	0,98	0,95	0,95

Fonte: autores (2025).

Os resultados do modelo para predição de baixo desempenho são apresentados na tabela 3. De modo similar, esses modelos se mostraram eficazes na tarefa proposta, com desempenho preditivo satisfatório. A precisão varia entre 98% e 99%, indicando que a maioria das amostras identificadas como positivas realmente correspondem a candidatos de baixo desempenho, com uma margem de apenas 1% a 2% de falsos positivos. Já o recall, situado entre 89% e 91%, revela que o modelo foi capaz de capturar a maioria das

amostras positivas, identificando com sucesso candidatos de baixo desempenho. Um recall menor que o modelo de predição de alto desempenho indica que o modelo apresentou maior taxa de falsos positivos.

Tabela 2 –

Avaliação dos modelos preditivos de baixo desempenho.

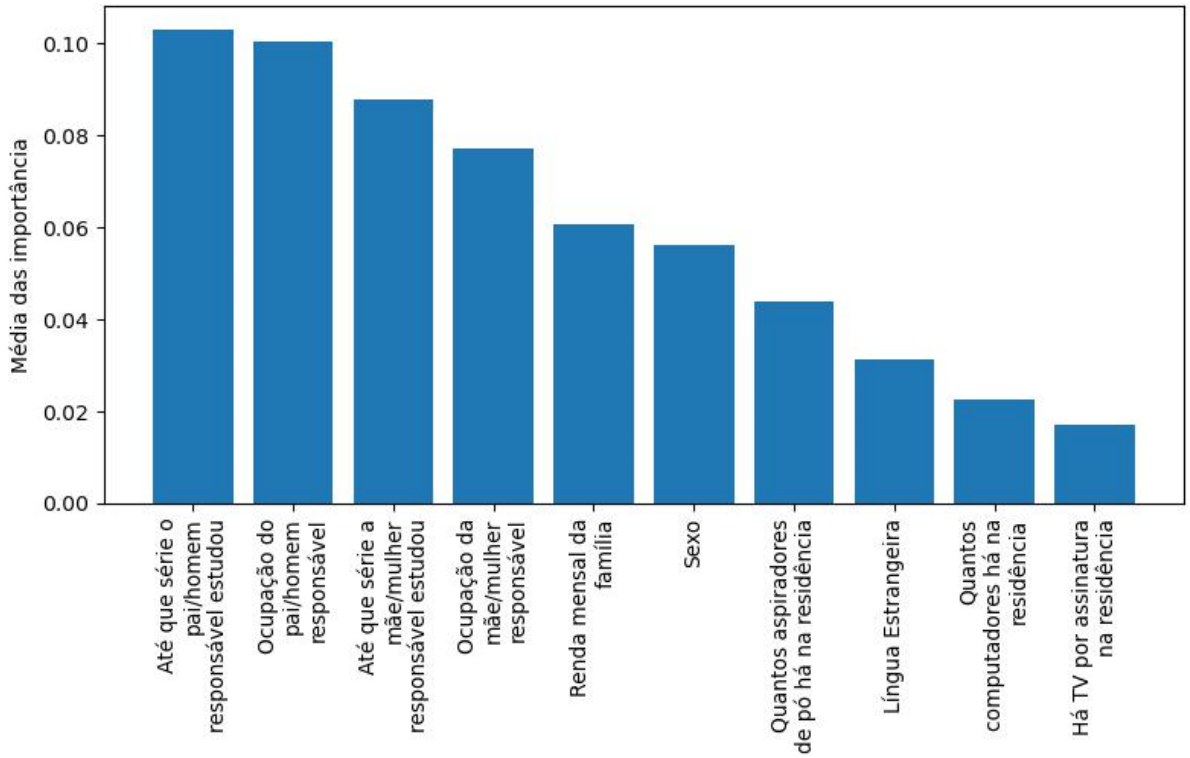
Prova	Alto desempenho			Não alto desempenho			Acurácia
	Precisão	Recall	F1-score	Precisão	Recall	F1-score	
Ciências da Natureza	0,98	0,90	0,94	0,91	0,99	0,94	0,95
Ciências Humanas	0,98	0,91	0,94	0,91	0,98	0,94	0,95
Linguagens e Códigos	0,99	0,91	0,94	0,91	0,99	0,95	0,95
Matemática	0,98	0,91	0,94	0,91	0,98	0,95	0,94
Redação	0,98	0,89	0,94	0,90	0,99	0,94	0,94

Fonte: autores (2025).

Para o F1-score os valores estão entre 94% e 95%, tanto para o baixo desempenho, quanto para o não baixo desempenho, demonstrando um equilíbrio satisfatório entre precisão e recall. Por fim, a acurácia geral do modelo, também entre 94% e 95%, reforça sua capacidade de prever com confiabilidade candidatos com baixo desempenho, consolidando-o como uma ferramenta eficaz para essa tarefa.

As dez variáveis com maior importância média são apresentadas na figura 1, juntamente com seu valor de importância. Pode-se perceber que as variáveis mais importantes para os modelos construídos são aquelas relacionadas ao grau de instrução e ocupação dos pais do candidato, bem como à renda mensal da família, o que corresponde às cinco variáveis mais importantes. Esses achados sugerem que esses fatores possuem maior correlação com o desempenho final do estudante no Enem, que será analisada nas próximas análises.

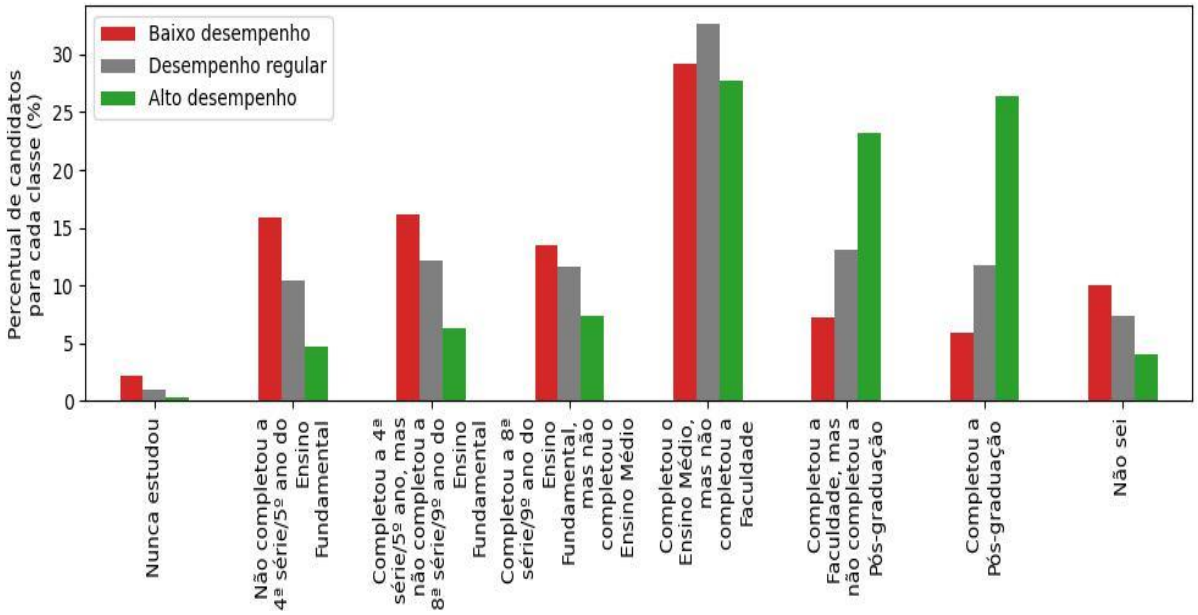
Figura 1 –
Conjunto das dez variáveis mais importantes do estudo.



Fonte: autores (2025).

A figura 2 apresenta a comparação entre o desempenho apresentado pelos candidatos na prova de ciências humanas e o valor da variável que representa o grau de instrução do pai: variável apontada como mais importante para predição do desempenho dos candidatos. Pode-se observar que para menores graus de instrução do pai, o percentual de candidatos de baixo desempenho tende a ser maior. Para graus de instrução maiores, como ensino superior ou pós-graduação completos, o percentual de candidatos com alto desempenho tende a ser maior. Esses resultados sugerem que o desempenho do candidato é influenciado pelo nível de formação do pai.

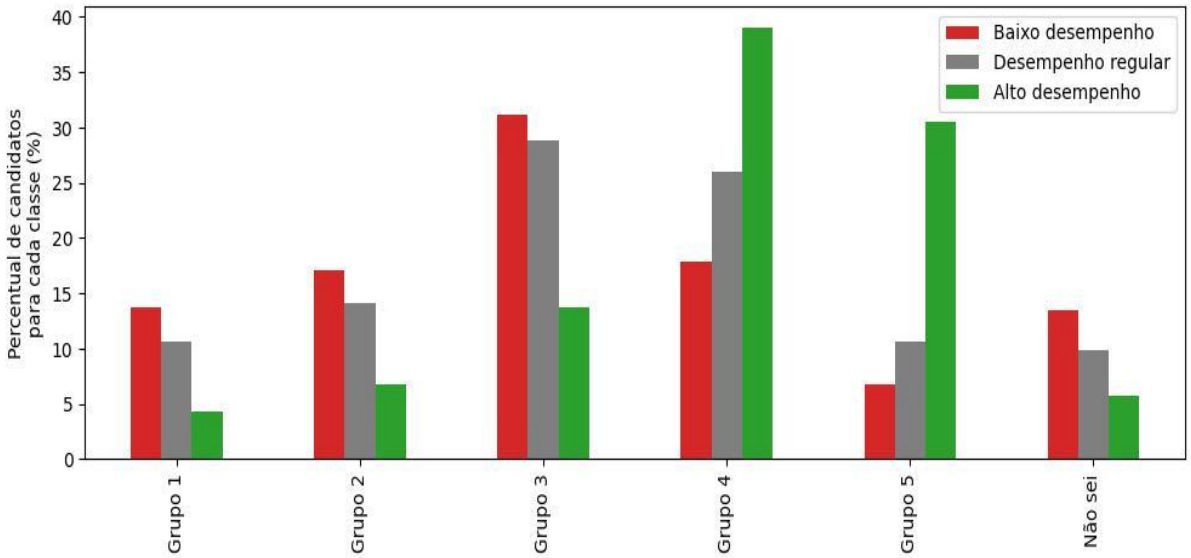
Figura 2 –
Desempenho dos candidatos em função do grau de instrução do pai; prova de ciências humanas.



Fonte: autores (2025).

A figura 3 compara o desempenho dos candidatos na prova de ciências da natureza com a ocupação do pai. Essa variável agrupa as respostas em cinco grupos. Apesar de um grupo não representar que o pai ou responsável tem maior estudo ou maior renda, os grupos estão divididos onde o grupo 1 inicia com funções mais operacionais e braçais – lavrador, agricultor, pescador –, e o grupo 5 com atividades mais estratégicas ou de liderança: médico, engenheiro, diretor de empresa.

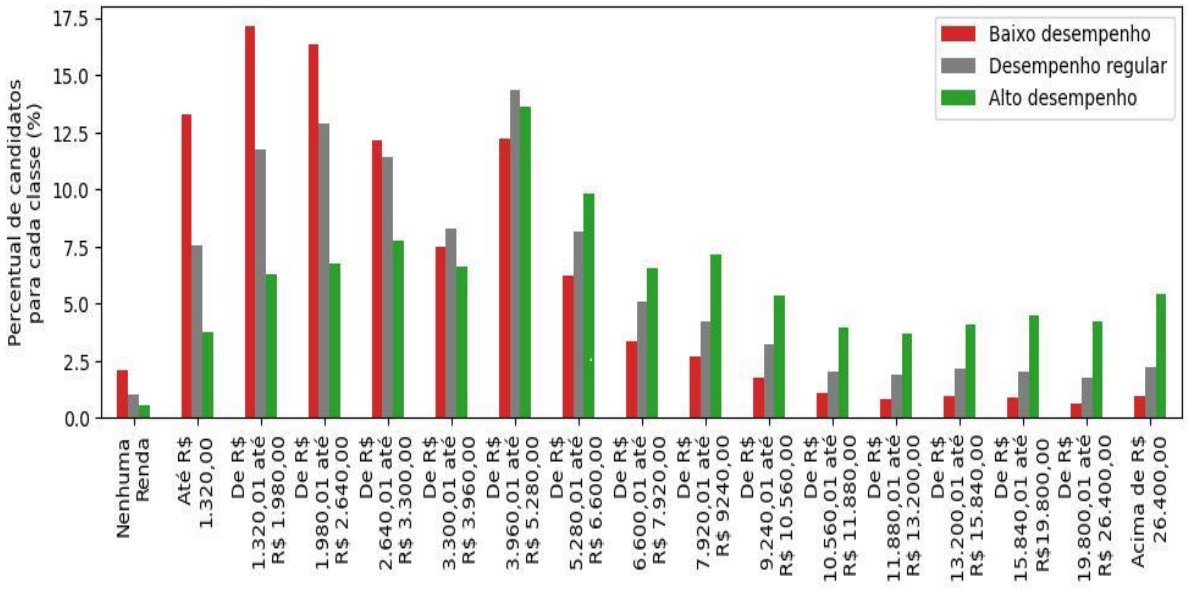
Figura 3 –
Desempenho dos candidatos em função da ocupação do pai; prova de ciências da natureza.



Fonte: autores (2025).

Pela análise da figura 3, percebe-se que candidatos cuja ocupação do pai se enquadra nos grupos 1 a 3 tendem a apresentar maiores percentuais para a classe de baixo desempenho. Quando a ocupação do pai está nos grupos 4 ou 5, os candidatos tendem a ser classificados como alto desempenho. Ou seja, ocupações do pai relacionadas a potenciais maiores experiência, nível de formação ou renda estão associadas a melhores desempenhos médios do candidato. A relação entre o grau de instrução e ocupação da mãe e o desempenho do candidato apresenta comportamento similar, porém com menor intensidade em comparação com as variáveis associadas ao pai.

Figura 4 –
Desempenho dos candidatos em função da renda familiar; redação.



Fonte: autores (2025).

Finalmente, a figura 4 mostra a relação entre a renda familiar dos candidatos e seu desempenho na redação. Percebe-se que as distribuições dos percentuais de candidatos com baixo e alto desempenhos se aproximam de distribuições normais. No entanto, a média é menor para o grupo de candidatos de baixo desempenho, em relação ao grupo de candidatos de alto desempenho. Ou seja, baixo desempenho está associado a menores valores de renda, enquanto alto desempenho está associado a maiores valores de renda.

Considerações finais

Este trabalho explora técnicas de ciência de dados e aprendizagem de máquina para estudar o desempenho dos estudantes no Enem, e sua relação com os dados socioeconômicos desses estudantes. São construídos modelos baseados em random forest para classificar o desempenho dos estudantes em alto, baixo ou regular, conforme os dados socioeconômicos apresentados. Esses modelos se mostraram eficazes para essa tarefa preditiva, apresentando valores de acurácia, precisão, recall e F1-score superiores a 90%, o que indica um desempenho preditivo satisfatório. Esses resultados são consistentes na predição dos desempenhos dos estudantes para as diferentes provas que compõem o Enem, bem como para a redação.

Os modelos construídos foram explorados mais a fundo e identificadas as variáveis que melhor explicam as predições realizadas. O desempenho dos estudantes foi analisado em função das cinco variáveis mais importantes. Foi possível perceber que o desempenho dos estudantes tende a ser melhor quando o grau de instrução dos seus pais é maior, quando a ocupação dos seus pais está associada a atividades mais estratégicas ou de liderança e, finalmente, quando sua renda familiar é maior.

Esses achados destacam a relevância dos fatores socioeconômicos no desempenho educacional, oferecendo diretrizes para a formulação de políticas públicas e estratégias pedagógicas para melhoria da educação. É importante identificar variáveis que impactam o desempenho acadêmico, como renda familiar, ocupação e grau de instrução dos pais, pois auxilia no planejamento de intervenções direcionadas a mitigar desigualdades e promover equidade educacional. Além disso, os resultados reforçam o papel da análise preditiva como uma ferramenta para compreender dinâmicas sociais complexas, auxiliando gestores e educadores na tomada de decisões baseadas em dados.

Este estudo pode ser estendido em diferentes direções, das quais destacam-se duas. Em primeiro lugar, outras variáveis presentes nos microdados do Enem podem ser incorporadas na elaboração dos modelos e, por consequência, na predição do desempenho dos estudantes. Por exemplo, o tipo de escola em que o estudante concluiu o Ensino Médio ou o ano de conclusão. Em segundo lugar, uma investigação mais aprofundada da influência das variáveis analisadas no desempenho do candidato pode fornecer um entendimento mais adequado da situação educacional do país.

Referências

BERGSTRA, James; BENGIO, Yoshua. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, Brookline, v. 13, n. 2, 2012, p. 281-305.

BREIMAN, Leo. Random forests. Machine Learning, Berlim, v. 45, 2001, p. 5-32.

CHAWLA, Nitesh V; BOWYER, Kevin W; HALL, Lawrence O; KEGELMEYER, W. Philip. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, El Segundo, v. 16, 2002, p. 321-357.

MEC. *Enem*: Exame Nacional do Ensino Médio 2023. Disponível em: https://www.gov.br/inep/pt-br/areas-de-atuacao/avaliacao-e-exames-educacionais/enem. Acesso em: 18 set. 2023.

SEGER, Christian. *An investigation of categorical variable encoding techniques in machine learning*: binary versus one-hot and feature hashing. KTH Royal Institute of Technology, School of Electrical Engineering and Computer Science: Stockholm, Sweden, 2018.

HE, Haibo; GARCIA, Edwardo. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, Los Alamitos, v. 21, n. 9, 2009, p. 1263-1284.

KRAWCZYK, Bartosz. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, Heidelberg, v. 5, n. 4, 2016, p. 221-232.

FERNÁNDEZ, Alberto; GARCIA, Salvador; HERRERA, Francisco; CHAWLA, Nitesh. SMOTE for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary. *Journal of Artificial Intelligence Research*, El Segundo, v. 61, 2018, p. 863-905.