MECAI 2024 III Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Influências socioeconômicas e geográficas no desempenho do ENEM 2023: Um estudo estatístico e de aprendizado de máquina

Lucas de Carvalho Leitão Maretti[1]
ICMC-USP
Paulino Ribeiro Villas Boas[2]
ICMC-USP

1 Introdução

No Brasil, o Exame Nacional do Ensino Médio (Enem) é uma das principais ferramentas de avaliação do nível de aprendizado dos estudantes de ensino médio, podendo ser utilizado tanto para acesso ao ensino superior quanto para medir a qualidade do ensino ofertado. A relação entre desempenho acadêmico e fatores socioeconômicos tem sido amplamente discutida na literatura, evidenciando que condições sociais, econômicas e culturais afetam diretamente as oportunidades educacionais dos estudantes. [1]

Este trabalho tem como objetivo utilizar técnicas estatísticas como análise exploratória, aprendizado de máquina e inferência para mineração de dados e extração de informações de grandes bases de dados públicas usando como objeto para este estudo os microdados do Enem 2023, tendo como objetivo identificar os principais determinantes socioeconômicos e geográficos que influenciam o desempenho dos estudantes na prova.

2 Metodologia

O *dataset* dos microdados do Enem 2023 compreende cerca de 2.2 milhões de observações, em que se separou 20% desse conjunto antes de fazer qualquer análise como conjunto de teste. Quando do treinamento dos modelos, realizou-se nova separação na proporção 70:30 para dados de treinamento e validação. De acordo com o site Guia da Carreira [2], estudantes que conseguem notas maiores do que 650 na média geral do Enem têm maiores chances de competir por

[1]lucas.maretti@usp.br
[2]paulino.villas-boas@usp.br

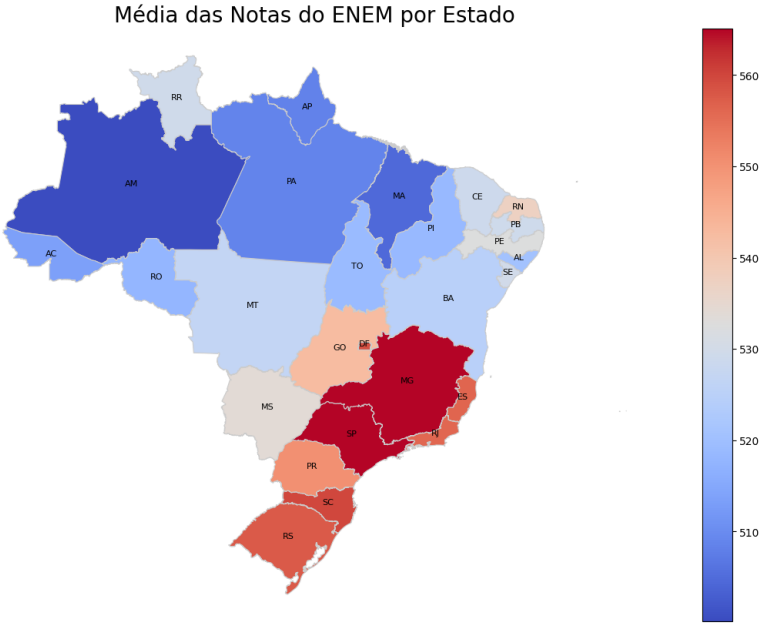
vagas em universidades federais brasileiras. Com base nisso, para aplicação de técnicas de aprendizado de máquina o problema foi modelado como um problema de classificação binária em que a variável *target* recebe 1 caso o aluno atingisse a nota de corte de 650 e 0 do contrário. Com isso, o problema tornou-se de classificação desbalanceada, com 87% dos dados pertencentes à classe 0 e 13% à classe 1. Para lidar com o desbalanceamento optou-se por utilizar modelos tipo *strong learners* conforme designação usada em [3]. Estes modelos caracterizam-se por possuírem hiperparâmetros capazes de lidar com desbalanceamento de classes sem necessidade de criação de amostras sintéticas. Este estudo avaliou o desempenho dos modelos Regressão Logística (baseline), XGBoost, LightGBM e Random Forest, treinados com validação cruzada *5-fold* para determinação dos hiperparâmetros ótimos. A métrica principal de análise foi a auprc (area under precision-recall curve), recomendada para problemas com classes desbalanceadas.

3 Resultados e Discussão

3.1 Análise Exploratória

A análise exploratória dos dados consistiu na utilização de técnicas de visualização e estatísticas, além da aplicação de *feature engineering* para a criação de novas variáveis capazes de ilustrar a relação entre fatores socioeconômicos e o desempenho acadêmico dos estudantes. Um exemplo dessa análise é apresentado na Figura 1, onde foi plotado um mapa do Brasil com o desempenho médio dos estudantes em cada estado. Esse gráfico permite identificar discrepâncias geográficas em relação às notas, evidenciando variações regionais no desempenho acadêmico.

Figura 1: Mapa das notas do Enem por Estado

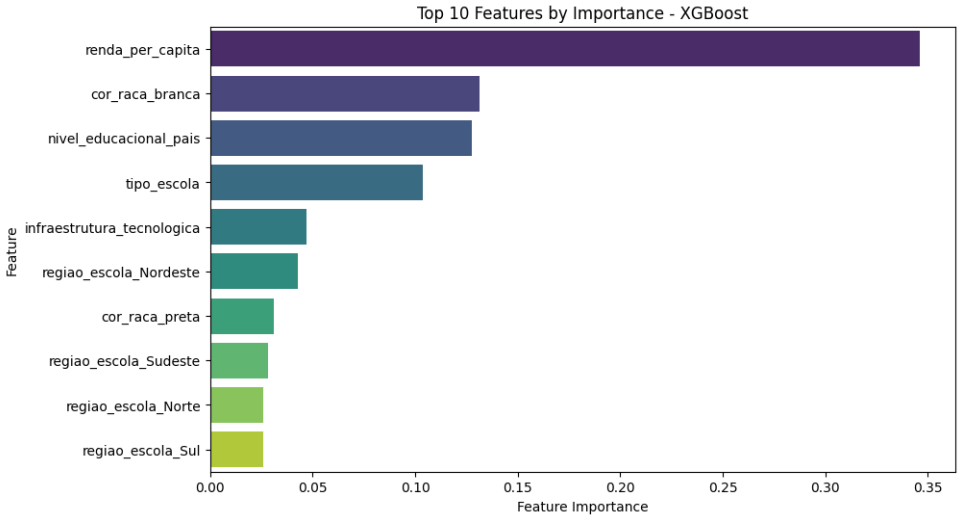


Fonte: Autor

3.2 Aprendizado de Máquina

O modelo XGBoost foi o que apresentou melhor desempenho na métrica selecionada em comparação com o modelo baseline. Além da capacidade preditiva, este estudo avaliou a capacidade de inferência dos modelos, ou seja, na capacidade destes de responder à perguntas como: ”quais preditores estão mais associados com a variável resposta?” [4]. Sendo assim, após treinamento analisou-se as *feature importances* do modelo XGBoost para entender quais variáveis apresentavam maior associação com o target, apresentado na Figura 2 abaixo.

Figura 2: Resultados de feature importance para o modelo XGBoost



Fonte: Autor

Como se observa da Figura 2, a renda per capita dos alunos é o fator fundamental para um melhor desempenho preditivo do modelo. Ressalta-se que para os demais modelos estudados esta mesma variável também apareceu como a mais influente no desempenho preditivo. Apesar de isto não significar que se pode atribuir relação de causalidade entre renda per capita e desempenho acadêmico, o resultado observado mostra que existe valor em se utilizar técnicas de aprendizado de máquina em grandes conjuntos de dados como forma de minerar padrões nos dados que seriam de difícil observação de outra forma. E, a partir disto, estudos mais direcionados podem ser feitos para melhorar entender esses padrões identificados.

4 Conclusão e próximos passos

Este estudo buscou aplicar técnicas de aprendizado de máquina e estatísticas para explorar os microdados do ENEM 2023 e identificar os principais fatores socioeconômicos e geográficos que influenciam o desempenho acadêmico dos estudantes. A partir da modelagem do problema como uma tarefa de classificação binária, utilizando modelos como XGBoost, LightGBM, e Random Forest, foi possível extrair informações valiosas sobre o impacto de fatores externos no desempenho escolar.

A análise de importância das variáveis revelou que a renda per capita foi o fator mais relevante para a predição do desempenho acadêmico dos estudantes, sendo observada em todos os modelos avaliados. Embora não se possa inferir causalidade direta entre renda e desempenho, os resultados indicam que a condição socioeconômica exerce um papel determinante no acesso a melhores oportunidades educacionais. Além disso, a análise exploratória dos dados destacou disparidades regionais no Brasil, sugerindo que políticas públicas voltadas para a redução dessas desigualdades poderiam melhorar os resultados educacionais em áreas mais vulneráveis.

As técnicas de aprendizado de máquina mostraram-se ferramentas poderosas para a mineração de grandes volumes de dados, permitindo identificar padrões que poderiam passar despercebidos em análises tradicionais. No entanto, é importante ressaltar que os resultados aqui apresentados devem ser complementados por estudos mais aprofundados que explorem as causas subjacentes dessas associações.

Como próximos passos, recomenda-se o desenvolvimento de modelos preditivos que levem em consideração não apenas os fatores socioeconômicos, mas também aspectos relacionados à qualidade do ensino e ao contexto familiar dos alunos. Além disso, estudos adicionais poderiam se concentrar em avaliar a eficácia de políticas educacionais implementadas para mitigar as desigualdades observadas, bem como avaliar relação de causalidade entre desempenho dos estudantes e variáveis socioeconômicas usando técnicas de inferência causal, por exemplo.

Referências

[1] LIMA, Priscila da Silva Neves et al. Análise de dados do Enade e Enem: uma revisão sistemática da literatura. Avaliação: Revista da Avaliação da Educação Superior (Campinas), v. 24, p. 89-107, 2019.

[2] https://www.guiadacarreira.com.br/blog/pontos-do-enem-para-cada-curso

[3] Elor, Yotam, and Hadar Averbuch-Elor. ”To SMOTE, or not to SMOTE?.”arXiv preprint arXiv:2201.08528 (2022).

[4] James, Gareth, et al. An introduction to statistical learning. Vol. 112. New York: springer, 2013.