

O PESO DA ORIGEM: MODELAGEM PREDITIVA DE INTELIGÊNCIA ARTIFICIAL DOS DETERMINANTES SOCIOECONÔMICOS E ESTRUTURAIS NO ENEM

Gabriel Sartori Mosquer - Centro Universitário SENAI/SC
gabriel_mosquer@estudante.sesisenai.org.br
Fernando Ramos Lengler - Centro Universitário SENAI/SC
Willian Daniel de Mattos - Centro Universitário SENAI/SC

INTRODUÇÃO: A desigualdade educacional brasileira é um fenômeno complexo, no qual o contexto de vida do estudante exerce influência determinante sobre o desempenho, frequentemente obscurecendo o mérito individual. O presente trabalho tem como objetivo desenvolver um modelo de *Machine Learning* baseado em *Gradient Boosting* capaz de estimar a nota média do ENEM, utilizando exclusivamente variáveis socioeconômicas, demográficas e de infraestrutura escolar. Esta abordagem visa preencher uma lacuna analítica: mensurar estatisticamente, através de métricas de erro (RMSE) e explicabilidade (SHAP), o peso das variáveis determinantes estruturais no rendimento discente, isolando-os de históricos escolares pregressos. **METODOLOGIA:** A premissa metodológica é dissociar a previsão de notas pregressas para revelar o impacto puramente estrutural dessas condicionantes. A pesquisa segue o ciclo CRISP-DM (*Cross-Industry Standard Process for Data Mining*) e utiliza os microdados públicos do INEP (2020-2022). A etapa de *Data Preparation* foi concluída, resultando em um *dataset* analítico otimizado com engenharia de atributos focada em índices de infraestrutura e conforto. Na fase atual de modelagem, algoritmos de estado da arte, como LightGBM e XGBoost, são treinados e submetidos à otimização bayesiana de hiperparâmetros para maximizar a precisão preditiva. **RESULTADOS:** Os resultados parciais consolidam a estruturação de modelos preditivos robustos, focados na minimização do erro quadrático médio. O impacto central da análise reside na aplicação de Inteligência Artificial Explicável (XAI), via valores SHAP (*SHapley Additive exPlanations*). Esta técnica permite traduzir as previsões do modelo em *insights* sociológicos, hierarquizando quantitativamente quais fatores — como escolaridade materna, acesso à tecnologia ou tipo de escola — exercem maior pressão, positiva ou negativa, sobre a nota final. **CONCLUSÕES:** A intervenção demonstra a utilidade prática da Ciência de Dados no diagnóstico educacional. O diferencial do estudo reside na capacidade de fornecer métricas acionáveis que transcendem a simples correlação, oferecendo uma leitura causal da desigualdade. Espera-se que a consolidação destes modelos sirva de subsídio técnico para tomadas de decisão, evidenciando cientificamente onde a estrutura social e escolar atua como gargalo para o desempenho do estudante brasileiro.

Palavras-chave: Machine Learning; Inteligência Artificial; Desigualdade Educacional; ENEM; Mineração de Dados.