



PREDIÇÃO DE DESEMPENHO ACADÊMICO E RESILIÊNCIA EDUCACIONAL EM DADOS PÚBLICOS DO ENSINO MÉDIO

Jhôwfrann dos Santos Pinheiro Silva¹

¹UFMA, São Luís, Brasil (jhowfrann@gmail.com)

Resumo: Este estudo investiga fatores associados ao desempenho acadêmico e à resiliência educacional no ensino médio com dados públicos de Matemática e Português. Foram aplicados pré-processamento, engenharia de atributos, regressão, classificação e SHAP. O melhor modelo atingiu $R^2=0,419$ e acurácia=0,770, com destaque para reprovações anteriores, tendência das notas e faltas.

Palavras-chave: resiliência educacional; aprendizado de máquina; desempenho acadêmico; interpretabilidade; SHAP

INTRODUÇÃO

A predição do desempenho escolar é relevante porque o conjunto Student Performance reúne variáveis demográficas, sociais e escolares, sendo amplamente utilizado em tarefas de classificação e regressão. Neste estudo, observou-se que as variáveis G1, G2 e G3 exigem cuidado metodológico para evitar vazamento de informação. Em dados tabulares com relações não lineares, modelos como Random Forest e XGBoost são apropriados, enquanto SHAP permite interpretar a contribuição de cada variável nas previsões (Breiman, 2001; Chen e Guestrin, 2016; Lundberg e Lee, 2017).

Neste contexto, o objetivo deste trabalho foi analisar fatores associados ao desempenho acadêmico e identificar indicadores de resiliência educacional em estudantes do ensino médio, com ênfase em interpretabilidade e em resultados experimentais realistas.

MATERIAL E MÉTODOS

Foi utilizado o conjunto Student Performance, composto por registros de estudantes de duas escolas portuguesas, com informações de Matemática e Língua Portuguesa, originalmente descrito por Cortez e Silva (2008). O conjunto foi disponibilizado publicamente e contém variáveis demográficas, sociais e escolares, além das notas G1, G2 e G3, empregadas em tarefas de classificação binária, classificação multiclasse e regressão no estudo original.

As variáveis foram combinadas em um único banco e submetidas a pré-processamento com codificação categórica e padronização numérica, preservando a disciplina como atributo explicativo. Foram criadas variáveis derivadas sem uso direto da nota final como preditor, a fim de evitar vazamento de informação. A

resposta da regressão foi a nota final G3, e a resposta da classificação foi a aprovação binária.

Para modelagem, foram treinados Ridge Regression, Random Forest e XGBoost na regressão, além de Random Forest e XGBoost na classificação. A avaliação considerou R^2 , erro quadrático médio e acurácia, e a interpretabilidade foi obtida por SHAP (Hoerl e Kennard, 1970; Breiman, 2001; Chen e Guestrin, 2016; Lundberg e Lee, 2017). O uso de Ridge é adequado em problemas com colinearidade, enquanto Random Forest e XGBoost são apropriados para relações não lineares e interação entre variáveis.

RESULTADOS E DISCUSSÃO

Os melhores resultados foram obtidos pelo Random Forest Regressor, com $R^2=0,419$, e pelo Random Forest Classifier, com acurácia de 0,770. Esses valores refletem a complexidade da predição acadêmica quando se evita o uso de variáveis de histórico imediato como preditoras, tornando a tarefa metodologicamente mais realista.

Na análise de SHAP, as variáveis com maior influência foram failures, grade_trend, absences, subject_Portuguese e risk_score. O histórico de reprovações apareceu como o principal fator associado ao resultado, enquanto a tendência entre G2 e G1 foi interpretada como um indicador de evolução ao longo do período letivo. O número de faltas também teve impacto relevante, reforçando seu papel como indicador de risco acadêmico. A diferença entre disciplinas mostrou que o contexto curricular influencia o desempenho, o que amplia a interpretação educacional dos resultados. A escolha de SHAP foi adequada porque esse método atribui importância local e global às variáveis, facilitando a leitura do modelo (Lundberg e Lee, 2017).



O percentual de estudantes classificados como resilientes foi de aproximadamente 24%, proporção que permitiu analisar o grupo sem torná-lo excessivamente raro nem artificialmente amplo. Assim, os resultados sugerem que a resiliência educacional não depende apenas de condições iniciais, mas também de trajetórias de recuperação ao longo do tempo.

CONCLUSÃO

O estudo mostrou que o desempenho acadêmico pode ser modelado com qualidade moderada em dados públicos do repositório UCI, sem recorrer a vazamento de informação. Os resultados apontam histórico de reprovações, faltas, tendência de evolução e contexto da disciplina como fatores centrais para compreender a resiliência educacional. O uso de SHAP contribuiu para interpretar as previsões e fortalecer a validade científica da análise.

AGRADECIMENTOS

Agradeço ao apoio institucional, ao acesso aos dados públicos utilizados neste estudo e à minha esposa, pelo apoio e incentivo durante o desenvolvimento deste trabalho.

REFERÊNCIAS

- BREIMAN, Leo. Random forests. *Machine Learning*, v. 45, n. 1, p. 5-32, 2001.
- CHEN, Tianqi; GUESTRIN, Carlos. XGBoost: a scalable tree boosting system. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. San Francisco: ACM, 2016. p. 785-794.
- CORTEZ, Paulo; SILVA, Alice Maria Gonçalves. Using data mining to predict secondary school student performance. In: BRITO, A.; TEIXEIRA, J. (ed.). *Proceedings of the 5th Future Business Technology Conference (FUBUTEC 2008)*. Porto: EUROSIS, 2008. p. 5-12.
- HOERL, Arthur E.; KENNARD, Robert W. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, v. 12, n. 1, p. 55-67, 1970.
- LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. In: GUYON, I.; VON LUXBURG, U.; BENGIO, S.; WALLACH, H.; FERGUS, R.; VISHWANATHAN, S.; GARNETT, R. (ed.). *Advances in Neural Information Processing Systems 30*. Red Hook: Curran Associates, 2017. p. 4765-4774.