AcolheEdu: Uma Solução de Triagem Escolar para Risco Psicossocial com Aprendizado de Máquina

Beatriz Brum[1], Iasmin Dembinski[2], Cristhiano Vasconcellos[1], Danísio Trindade[1], Carlos Santos[1]

[1] Instituto Federal de Educação, Ciência e Tecnologia Farroupilha (IFFar)
Alegrete – RS – Brasil

[2]Universidade Federal do Pampa (UNIPAMPA)
Alegrete – RS – Brasil

beatriz.06059@aluno.iffar.edu.br
iasminsilva.aluno@unipampa.edu.br
cristhiano.vasconcellos@iffarroupilha.edu.br
danisio.trindade@iffarroupilha.edu.br
carlos.santos@iffarroupilha.edu.br

Abstract. *AcolheEdu presents a data-driven screening approach for early identification of students potentially vulnerable to psychological distress. Using anonymized PeNSE 2019 microdata, a calibrated HistGradientBoosting model estimates a risk score from self-reported psychosocial factors. It achieved ROC-AUC 0.86 and PR-AUC 0.77 in stratified cross-validation, and ROC-AUC 0.859 on a holdout test. With a 0.30 threshold, Recall reached 80.1% with 61.7% precision. A navigable Figma prototype is available, with FastAPI REST integration planned.*

Resumo. *O AcolheEdu propõe uma triagem escolar orientada a dados para identificar precocemente estudantes potencialmente vulneráveis a sofrimento psíquico. Com microdados públicos e anonimizados da PeNSE 2019, um HistGradientBoosting calibrado estima um risk score a partir de fatores psicossociais autorreferidos. O modelo obteve ROC-AUC 0,86 e PR-AUC 0,77 na validação cruzada e ROC-AUC 0,859 no holdout. Com limiar 0,30, atingiu Recall de 80,1% e Precision de 61,7%. O fluxo atual é um protótipo navegável no Figma, com integração via API (FastAPI) prevista.*

1. Contexto

A saúde mental no ambiente escolar tem se consolidado como um desafio relevante no Brasil. Dados da Pesquisa Nacional de Saúde do Escolar indicam elevada prevalência de sentimentos de tristeza e solidão entre adolescentes [IBGE 2021]. No âmbito institucional, equipes pedagógicas e núcleos de apoio psicopedagógico enfrentam limitações de recursos humanos associadas a alta demanda por atendimentos, o que dificulta a identificação precoce de estudantes em sofrimento psíquico. Nesse cenário, as intervenções tendem a ocorrer de forma predominantemente reativa, quando os sinais já se apresentam em maior gravidade, o que pode estar associado a riscos como evasão escolar, queda do desempenho acadêmico e agravamento do sofrimento psíquico [Andifes 2019].

Diante desse contexto, o AcolheEdu foi concebido como uma solução de apoio à atuação de núcleos de apoio psicopedagógico e pedagógico, por meio de uma triagem automatizada de alta sensibilidade baseada em dados públicos. O Instituto Federal Farroupilha (IFFar), no contexto do ensino médio integrado, constitui o cenário inicial de adoção da solução como projeto piloto, com governança do fluxo atribuída ao núcleo de apoio do Câmpus Alegrete.

A proposta integra dados públicos, um modelo preditivo calibrado e um fluxo inicial de triagem escolar. Assim, este estudo caracteriza-se como uma prova de conceito (PoC), com o objetivo de avaliar a viabilidade técnica da solução como etapa preliminar à adoção institucional.

2. Processo adotado

O processo seguiu a metodologia CRISP-DM [Shearer 2000, Martinez-Plumed et al. 2021], com quatro etapas principais: (i) preparação de dados, (ii) modelagem e calibração, (iii) definição de limiar operacional e (iv) prototipagem do fluxo de uso.

Utilizamos microdados da PeNSE 2019, considerando 42.422 registros de estudantes do Ensino Médio. A variável-alvo foi operacionalizada como um *proxy* de risco baseado em autorrelato, uma vez que a base não contém diagnóstico clínico. Para a triagem inicial, priorizou-se o item de tristeza frequente ou constante nos últimos 30 dias; adicionalmente, o item "vida não vale a pena" foi considerado na contextualização metodológica do construto de risco psicossocial. Foram selecionadas 46 variáveis preditoras distribuídas em blocos temáticos, incluindo sono, violência/bullying, apoio social, atividade física, uso de substâncias, rotina, telas/internet e contexto escolar. Variáveis diretamente correlatas ao bloco de saúde mental foram excluídas, buscando reduzir vazamento semântico.

Comparamos Regressão Logística, adotada como *baseline* interpretável, e HistGradientBoosting, com implementação na biblioteca *Scikit-learn*. O modelo final foi calibrado via *Platt scaling* para melhorar a confiabilidade das probabilidades em um uso baseado em limiar [Niculescu-Mizil and Caruana 2005]. Para a triagem inicial, priorizamos *Recall*, dada a gravidade de falsos negativos. Avaliamos múltiplos limiares e selecionamos 0,30 como ponto de operação com sensibilidade elevada e *Precision* ainda aceitável.

O fluxo de produto pretendido envolve um aplicativo móvel que: (i) coleta respostas de um questionário, (ii) consome uma API (FastAPI) e (iii) retorna *risk score*, faixa de risco e recomendações para apoiar a gestão escolar. Nesta fase, entretanto, o protótipo está implementado como um fluxo navegável no Figma, sem integração funcional com API; a implementação da API e a integração ponta a ponta fazem parte do plano de evolução do sistema. Nesta etapa, utilizamos exclusivamente microdados públicos e já anonimizados da Pesquisa Nacional de Saúde do Escolar (PeNSE 2019), disponibilizados pelo IBGE[1]. Por se tratar de uma prova de conceito, a motivação foi reduzir incertezas técnicas de pipeline, calibração, definição de limiares e apresentação dos resultados antes de avançar para etapas de campo com atores institucionais e procedimentos formais de validação.

[1] https://www.ibge.gov.br/estatisticas/sociais/saude/9134-pesquisa-nacional-de-saude-do-escolar.html

Como próximos passos, planejamos conduzir elicitação e validação exploratória com atores da escola, como psicólogos, equipes de apoio e gestão, por meio de entrevistas semiestruturadas e/ou oficinas de co-design, visando refinar requisitos, linguagem das recomendações e fluxos de encaminhamento. Também estruturaremos um piloto institucional com termos de consentimento aplicáveis e apreciação ética em Comitê de Ética em Pesquisa (CEP), quando cabível, antes de qualquer coleta ou uso de dados não públicos.

3. Solução Proposta

A solução integra um modelo preditivo calibrado a um fluxo operacional de triagem, com atenção à privacidade e ao uso responsável dos resultados.

A arquitetura alvo prevê uma API REST (FastAPI) hospedando o pipeline de pré-processamento e o modelo calibrado, expondo um *endpoint* `/predict`. O aplicativo móvel coletará respostas de um questionário derivado de itens compatíveis com a lógica dos fatores utilizados no treinamento, enviará essas respostas à API e receberá: (i) probabilidade calibrada, (ii) faixa de risco (baixo/moderado/alto) e (iii) recomendações acionáveis, como orientações relacionadas a sono, bullying, suporte social e encaminhamento institucional. No momento, o fluxo de interface e navegação está prototipado no Figma, e o modelo é avaliado *offline*; a integração funcional app–API está prevista para a próxima etapa de implementação. Nesta prova de conceito, não há armazenamento de dados pessoais de estudantes, pois foram utilizados apenas microdados públicos anonimizados. Em eventual implantação institucional, a solução deverá observar princípios da LGPD, controle de acesso e governança de dados [Brasil 2018]. O sistema não tem finalidade diagnóstica; seus resultados destinam-se exclusivamente ao apoio à triagem institucional. A Figura 1 apresenta as principais telas de protótipo do aplicativo AcolheEdu.

Figura 1. Protótipos do aplicativo AcolheEdu: tela inicial, login, menu do estudante, questionário de avaliação de risco e painel de monitoramento do setor de apoio.

4. Resultados parciais

A Tabela 1 resume os principais resultados parciais obtidos pelo modelo.

Tabela 1. Desempenho do modelo calibrado.

Cenário	ROC-AUC	PR-AUC / Precision / Recall
Validação cruzada (5-fold)	0,86	PR-AUC = 0,77
Teste (*holdout*)	0,859	Precision = 0,617; Recall = 0,801 (thr = 0,30)

O desempenho observado (ROC-AUC $\approx$ 0,86 e PR-AUC $\approx$ 0,77 em validação cruzada; ROC-AUC $\approx$ 0,86 no *holdout*) sugere boa separação entre casos de maior e menor risco, mesmo com um alvo *proxy*. No ponto de operação 0,30, a sensibilidade (*Recall* $\approx$ 80%) é compatível com a lógica de triagem inicial, na qual se busca reduzir falsos negativos, ainda que com volume moderado de falsos positivos.

O AcolheEdu viabiliza uma triagem inicial escalável e de baixo custo, com probabilidade calibrada para apoiar a definição de limiares e estimar volume de encaminhamentos. A interpretabilidade futura, por exemplo com SHAP, pode subsidiar comunicação mais responsável sobre fatores associados ao risco, sem pretensão diagnóstica.

Como limitação, destaca-se que o *risk score* é um *proxy* por autorrelato e não substitui avaliação humana especializada; seu uso deve ser entendido como apoio à triagem e priorização de acolhimento. Além disso, como os dados são transversais (PeNSE 2019), pode haver *dataset shift*, exigindo calibração e validação local antes de uso continuado. Viés de autorrelato pode gerar falsos positivos e falsos negativos, demandando revisão humana e monitoramento da taxa de encaminhamentos conforme a capacidade institucional.

Para facilitar a avaliação da proposta e a visualização do fluxo de produto, disponibilizamos os seguintes materiais complementares:
Protótipo do AcolheEdu (Figma): `https://bit.ly/4rd3XSm`
Vídeo pitch do AcolheEdu: `https://bit.ly/3MsKvBT`
Modelo de negócio do AcolheEdu: `https://bit.ly/4amwYFp`

5. Conclusão e Trabalhos Futuros

O AcolheEdu representa uma proposta de triagem escolar baseada em dados, e os resultados indicam viabilidade técnica para uma triagem inicial de alta sensibilidade no contexto de saúde mental escolar, usando microdados públicos e anonimizados da PeNSE 2019 e um modelo calibrado (ROC-AUC 0,859 no *holdout*; com limiar 0,30, *Recall* 80,1% e *Precision* 61,7%). Nesta fase, o fluxo de produto está materializado como protótipo navegável no Figma, visando validar requisitos e fluxo de uso antes da implementação funcional.

Como trabalhos futuros, pretende-se conduzir um piloto institucional no IFFar – Câmpus Alegrete, em articulação com o núcleo de apoio aos discentes, para refinar questionário, recomendações e limiares conforme a capacidade operacional, além de estabelecer salvaguardas de privacidade, termos de consentimento e submissão ética ao CEP, quando aplicável.

Agradecimentos

Este estudo foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) – Brasil.

Referências

Andifes (2019). V pesquisa nacional de perfil socioeconômico e cultural dos(as) graduandos(as) das IFES – 2018. Technical report, Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (Andifes), Brasília.

Brasil (2018). Lei geral de proteção de dados pessoais (LGPD). Lei nº 13.709, de 14 de agosto de 2018.

IBGE (2021). Pesquisa nacional de saúde do escolar: 2019 (PeNSE 2019). Technical report, Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro.

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., and Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, 33(8):3048–3061.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning (ICML '05)*, pages 625–632.

Shearer, C. (2000). The crisp-dm model: the new blueprint for data mining. *Journal of Data Warehousing*, 5(4):13–22.