Aprendizado de Máquina na classificação do risco de uso frequente de substâncias em adolescentes na região Centro-Oeste

Thalles P. L. Mota[1], **Claudia A. Martins**[1]

[1]Instituto de Computação – Universidade Federal de Mato Grosso (UFMT)
CEP 78088-800 – Boa Esperança – Cuiabá - MT – Brasil

Abstract. *This study investigates the potential of Machine Learning in classifying substance use risk in adolescents using PeNSE-2019 data. Regional validation in Mato Grosso revealed that optimized XGBoost models achieved 0.80 Recall and 0.90 AUROC. Interpretability via SHAP highlighted the influence of peer groups and the protective role of family, validating the technology for identifying latent vulnerabilities and supporting primary prevention policies.*

Resumo. *Este estudo investiga o potencial do Aprendizado de Máquina na classificação do risco de uso de substâncias em adolescentes utilizando dados da PeNSE 2019. A validação regional em Mato Grosso revelou que modelos otimizados de XGBoost atingiram Recall de 0.80 e AUROC de 0.90. A interpretabilidade via SHAP evidenciou a influência do grupo de pares e o papel protetor da família, validando a tecnologia para a identificação de vulnerabilidades latentes e suporte à prevenção primária.*

1. Introdução

A adolescência é um período pivotal na transição da infância para a vida adulta, marcado por intensas mudanças nos processos de desenvolvimento biopsicossocial, e, nesse contexto de transformação, evidencia-se uma vulnerabilidade, que, associada à tendência à experimentação de novos comportamentos e combinada ao processo de consolidação de traços da personalidade, pode levar a eventos que impactam profundamente tais indivíduos ao longo da sua vida [Arcadepani et al. 2023] [dos Reis et al. 2018]. Exemplo dessa vulnerabilidade torna-se visível quando dados apontam que tal estágio do desenvolvimento relaciona-se diretamente com desfechos como o suicídio, que representa a segunda principal causa de morte global na faixa etária de 15 a 24 anos [Gerstner et al. 2018]. Nesse cenário, abordagens científicas diversas têm sido propostas para entender, medir e monitorar a saúde e eventuais riscos dessa população-alvo. Essas pesquisas são de fundamental importância para que, futuramente, políticas públicas de prevenção e proteção possam ser arquitetadas e implementadas de forma a usar ferramentas tecnológicas eficientes [dos Reis et al. 2018].

Nesse contexto de saúde, um tema recorrente e que motiva diversos pesquisadores a propor abordagens metodológicas é o tema do uso de substâncias tóxicas. Essas pesquisas se dão em variedade, variabilidade e volume de dados, em dimensão regional, nacional e internacional. Assim, pesquisas publicadas apontam uma variedade de metodologias que envolvem Análise de *Cluster* para identificação de perfis de risco de uso de substâncias em adolescentes [Martins et al. 2021]. Outros pesquisadores pautam sua

investigação com Aprendizado de Máquina (AM), propondo validação internacional para predição do uso de substâncias em adolescentes [Kim et al. 2025]. Mais que isso, pesquisadores utilizaram *Large Language Models (LLMs)* para classificar emoções associadas ao uso de drogas em dados de *webforums* que são frequentados por adolescentes em postagens que *(i)* envolvem ou *(ii)* não o assunto drogas, e depois fizeram classificações das emoções presentes nos comentários com AM e evidenciaram emoções associadas ao uso de drogas [Zhang et al. 2025].

Essas pesquisas ilustram os esforços de pesquisadores a nível global em explorar possibilidades tecnológicas para ampliar o potencial de acervo tecnológico útil em contexto de prevenção e identificação de risco precoce, e configuram um campo de computação aplicada em plena expansão. Nesse sentido, o presente estudo objetiva explorar o problema do uso frequente de substâncias por adolescentes, buscando implementar classificadores eficientes, e, mais do que isso, interpretáveis do ponto de vista epistemológico, validando o potencial da tecnologia no cenário regional e propondo caminhos futuros para pesquisas.

2. Metodologia

A presente investigação exploratória caracteriza-se como uma análise quantitativa fundamentada em Aprendizado de Máquina Supervisionado de classificação binária. Utilizou-se dados da Pesquisa Nacional de Saúde do Escolar (PeNSE[1]) de 2019. Os dados foram filtrados para a manutenção exclusiva de registros de escolares com idade entre 13 e 18 anos da região Centro-Oeste. Em seguida, realizou-se um processo pré-processamento de dados, no qual mapeou-se as variáveis em ordinais, nominais e condicionais. Situações de resposta condicionadas por pulos tiveram seus valores convertidos para que não houvesse perda severa nos dados. O Rótulo foi definido como o uso frequente de substâncias e engloba o uso por 3 vezes ou mais nos últimos 30 dias, tanto para cigarro quanto para drogas ilícitas, como maconha, cocaína, entre outras.

Tendo em vista a convergência epidemiológica e capacidade de generalização dos modelos, estruturou-se uma estratégia de Validação Geográfica. Assim, os dados foram subdivididos em: (*i*) Conjunto de Treinamento e Otimização, que engloba dados dos estados de Mato Grosso do Sul (MS), Goiás (GO) e Distrito Federal (DF), que foi utilizado para a seleção de variáveis e otimização de hiperparâmetros, e o (*ii*) Conjunto de Validação Regional, composto exclusivamente pelos dados do estado de Mato Grosso (MT). Os dados de validação foram mantidos isolados durante todo o processo de modelagem, tendo sido utilizados pontualmente conforme delineado. Essa validação visa submeter os classificadores às diferentes taxas de prevalência dos rótulos. Com isso, buscou-se simular um proceso de validação mais robusto, no qual o modelo lida com dados ligeiramente divergentes dos que foram utilizados no treinamento, possibilitando uma análise mais efetiva da capacidade de generalização.

A modelagem preditiva explorou algoritmos *Random Forest* e *XGBoost*. A otimização global de hiperparâmetros e a seleção de atributos foram realizadas por meio do *Optuna*, que utilizou o estimador *Tree-structured Parzen Estimator* (*TPE*) ao longo de 300

[1]Instituto Brasileiro de Geografia e Estatística (IBGE). Disponível em: `https://www.ibge.gov.br/estatisticas/sociais/populacao/9134-pesquisa-nacional-de-saude-do-escolar.html`. Acesso 13 de Janeiro de 2026.

iterações. O processo priorizou a maximização do *Recall*, visando minimizar falsos negativos na identificação de riscos latentes. A convergência e estabilidade das descobertas foram exploradas via análise de (*SHapley Additive exPlanations (SHAP)*).

3. Resultados e Discussão

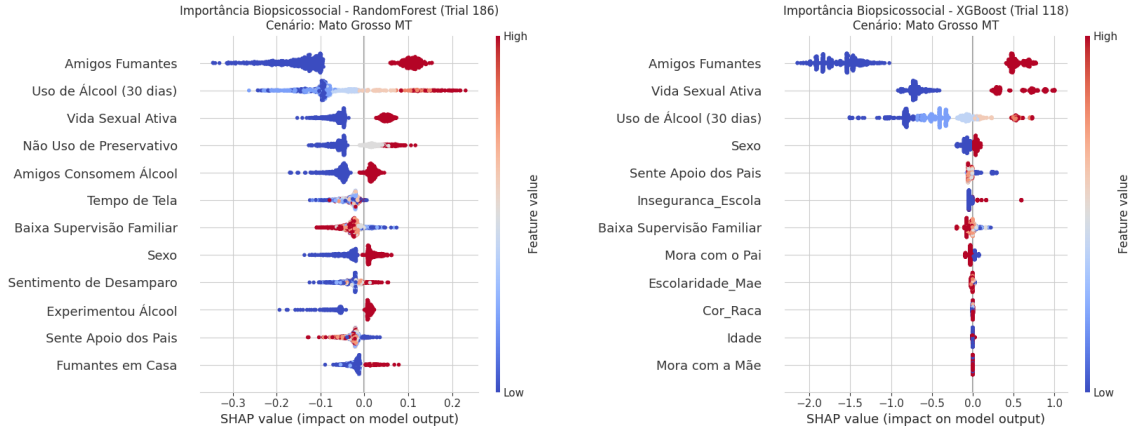
A análise inicial das prevalências no conjunto de treino contou com 10.348 registros e revelou uma prevalência de 5,3% para o rótulo, o que evidencia um desbalanceamento severo entre as classes. Esse cenário de escassez de eventos positivos representa um desafio técnico significativo para os modelos, exigindo que os mesmos capturem padrões complexos para evitar a negligência de casos reais. O conjunto de Validação contou com 3.277 registros e uma prevalência de 5,6% para o rótulo, conforme verificado posteriormente.

Quanto ao desempenho preditivo, pode-se observar na Tabela 1 que os modelos demonstraram uma robustez considerável ao operarem em dados inéditos. Ambos os classificadores atingiram valores de *AUROC* próximos a 0,90, o que indica uma alta capacidade discriminatória da tecnologia entre adolescentes de baixo e alto risco. Observou-se um *trade-off* estrutural entre as métricas de *Recall* e Precisão: enquanto o *Recall* manteve-se elevado, atingindo patamares de 0,80, a Precisão oscilou em níveis consideravelmente inferiores, situando-se entre 0,20 e 0,27. Na prática, essa configuração assegura que a ferramenta seja eficaz em não omitir casos de risco real, embora gere um volume significativo de alarmes falsos. Contudo, interpretados sob a ótica da vulnerabilidade potencial, esses sinais de risco podem representar algum risco real. Esses resultados corroboram com os achados de [Kim et al. 2025] em que os modelos são eficientes e encontram padrões generalizáveis para outros contextos sociodemográficos.

Tabela 1. Comparativo de desempenho entre os modelos classificadores para o cenário de Mato Grosso (MT).

Trial_ID	K_Variáveis	*AUROC*	Precisão	*Recall*
		XGBoost		
92	13	0.890	0.235	0.798
110	13	0.890	0.234	0.798
111	13	0.890	0.233	0.809
113	13	0.890	0.235	0.798
114	13	0.891	0.232	0.804
118	13	0.892	0.233	0.809
119	13	0.891	0.234	0.798
121	13	0.891	0.234	0.798
131	13	0.891	0.235	0.798
132	13	0.892	0.233	0.809
		Random Forest		
2	11	0.885	0.254	0.706
8	15	0.891	0.221	0.809
26	10	0.894	0.250	0.733
35	9	0.852	0.200	0.695
38	12	0.893	0.271	0.706
46	10	0.893	0.251	0.722
59	10	0.875	0.231	0.701
129	13	0.899	0.253	0.739
157	13	0.890	0.242	0.782
186	13	0.901	0.262	0.744

Na comparação entre as arquiteturas algorítmicas, o modelo baseado em *XGBoost* apresentou uma estabilidade técnica superior ao *Random Forest*. O *XGBoost* lidou de forma mais resiliente com o desbalanceamento das classes, mantendo métricas de *Recall* mais constantes e minimizando a negligênica de identificação de risco, o que o consolida como o estimador mais apto para sistemas de triagem em saúde pública. Isso dialoga com [Han et al. 2019] que, ao predizer uso de opióides, reportou desafios semelhantes para *Recall* e Precisão em modelos baseados em *XGBoost*.



(a) Modelo baseado em *Random Forest*. **(b) Modelo baseado em *XGBoost*.**

Figura 1. Análise de Estabilidade *SHAP* para os conjuntos de validação (MT).

A validação epistemológica via *SHAP* assegura que a decisão algorítmica respeite evidências empíricas [Aishwarya et al. 2025]. A análise de interpretabilidade via *SHAP* confirmou que a lógica decisória dos algoritmos está alinhada à epistemologia. O fator de socialização (amigos fumantes) isolou-se como o preditor de maior magnitude para a classificação de risco, seguido pelo (uso recente de álcool) e pela (vida sexual ativa). Adicionalmente, a tecnologia foi capaz de capturar dimensões protetivas, identificando que algumas variáveis (supervisão familiar e apoio dos pais) exercem uma contribuição reduzindo as probabilidades de risco, conforme evidenciado em outros artigos publicados [dos Reis et al. 2018] [Martins et al. 2021].

4. Considerações Finais

Este estudo explorou a viabilidade técnica de utilizar o AM para classificar o risco de uso frequente de substâncias entre adolescentes, apresentando uma alternativa robusta aos métodos estatísticos tradicionais em contexto de severo desbalanceamento. A validação regional em Mato Grosso confirmou que perfis de vulnerabilidade podem ser detectados de forma automatizada e com poucas questões. Por fim, a análise enfrentou limitações intrínsecas aos dados secundários, como a característica transversal dos dados e a potencial subnotificação, que pode distorcer a interpretação dos falsos positivos e, consequentemente, o desempenho dos modelos. Pesquisas futuras devem focar na validação em outros contextos regionais brasileiros, envolver especialistas da psicologia e visar avaliar a estabilidade desses modelos, bem como direcionar para uma finalidade prática real.

O presente trabalho foi realizado com o apoio da Fundação de Amparo à Pesquisa do Estado de Mato Grosso (FAPEMAT) por meio de concessão de bolsa de pesquisa

Referências

Aishwarya, S., Siddalingaswamy, P. C., and Chadaga, K. (2025). Explainable artificial intelligence driven insights into smoking prediction using machine learning and clinical parameters. *Scientific Reports*, 15(24069):1–22. DOI:https://doi.org/10.1038/s41598-025-09409-w.

Arcadepani, F. B., Fernandes, A. G., Castaldelli-Maia, J. M., and Fidalgo, T. M. (2023). Violent behavior, perceived safety, and assault experiences among adolescents: results from the brazilian national adolescent school-based health survey. *Brazilian Journal of Psychiatry*, 45(1):5–10. DOI: https://doi.org/10.47626/1516-4446-2022-2623.

dos Reis, A. A. C., Malta, D. C., and Furtado, L. A. C. (2018). Desafios para as políticas públicas voltadas à adolescência e juventude a partir da pesquisa nacional de saúde do escolar (pense). *Ciência & Saúde Coletiva*, 23(9):2879–2890. DOI: https://doi.org/10.1590/1413-81232018239.14432018.

Gerstner, R. M. F., Soriano, I., Sanhueza, A., Caffe, S., and Kestel, D. (2018). Epidemiología del suicidio en adolescentes y jóvenes en ecuador. *Revista Panamericana de Salud Pública*, 42:1–7. DOI:https://doi.org/10.26633/RPSP.2018.100.

Han, D., Lee, S., and Seo, D. (2019). Using machine learning to predict opioid misuse among U.S. adolescents. *Preventive Medicine*, 130:105886. DOI:https://doi.org/10.1016/j.ypmed.2019.105886.

Kim, S., Kim, H., Kim, S., Lee, H., Hammoodi, A., Choi, Y., Kim, H. J., Smith, L., Kim, M. S., Fond, G., Boyer, L., Baik, S. W., Lee, H., Park, J., Kwon, R., Woo, S., and Yon, D. K. (2025). Machine learning-based prediction of substance use in adolescents in three independent worldwide cohorts: Algorithm development and validation study. *Journal of Medical Internet Research*, 27:e62805. DOI: https://doi.org/10.2196/62805.

Martins, D. D., Lucena, L. B., Damasceno, A. R. M. B., Siqueira, H., Tadano, Y. d. S., and Caldas, I. F. R. (2021). Clusterização do perfil de adolescentes escolares com predisposição ao uso de substância psicoativas. *Research, Society and Development*, 10(2):e37510212528. DOI: https://doi.org/10.33448/rsd-v10i2.12528.

Zhang, X., Zhu, J., Kenne, D. R., and Jin, R. (2025). Teenager substance use on reddit: Mixed methods computational analysis of frames and emotions. *Journal of Medical Internet Research*, 27:e59338. DOI: https://www.jmir.org/2025/1/e59338.