OPEN ACCESS

DOI: 10.53843/bms.v11i15.1148

Brazilian Medical Students Journal

ARTIGO ORIGINAL

DOI: 10.53843/bms.v11i15.1148

MACHINE LEARNING NA SAÚDE PÚBLICA: PREDIÇÃO DO USO DE DROGAS ILÍCITAS ENTRE ADOLESCENTES BRASILEIROS

MACHINE LEARNING IN PUBLIC HEALTH: PREDICTING ILLICIT DRUG USE AMONG BRAZILIAN ADOLESCENTS

Mateus Zani De Nadai [1]; *Kaliani Angelo Ramos* [2]; *Marcelo Zanon Silva* [1]; *Felipe Sampaio Almeida Cardoso* [1]; *Gustavo Caballero Gilardy Mantovani* [1]; *Danylo Figueredo Cezana* [1]; *William Rodrigues de Freitas* [3].

1. Centro Universitário do Espírito Santo (UNESC), Colatina, Espírito Santo/ES, Brasil - estudante de medicina
2. Universidade Federal de Ouro Preto (UFOP), Ouro Preto, Minas Gerais/MG, Brasil - estudante de medicina
3. Universidade Federal do Sul da Bahia (UFSB), Itabuna, Bahia/BA, Brasil - professor no centro de formação em ciências da saúde

* mateus.zani@hotmail.com

Editor Associado: Kleuber Arias Meireles Martins.

Recebido: 04/07/2025. **Aceito:** 12/02/2026. **Publicado:** 17/05/2026.

RESUMO

INTRODUÇÃO: Machine Learning (ML) é uma área da inteligência artificial que permite o desenvolvimento de algoritmos capazes de aprender a partir de dados, sem programação explícita. O avanço dessas técnicas tem possibilitado aplicações importantes na saúde, incluindo a predição de riscos e comportamentos. Este estudo propõe o desenvolvimento de um modelo preditivo para compreender e prever o consumo de drogas ilícitas entre adolescentes brasileiros de 13 a 17 anos, a partir de padrões sociodemográficos e comportamentais, com o objetivo de subsidiar políticas públicas mais eficazes na saúde coletiva. **METODOLOGIA:** Trata-se de um estudo observacional e transversal, baseado em dados de 165.838 estudantes da Pesquisa Nacional de Saúde Escolar (PeNSE) 2019. Foram comparados diferentes modelos de ML para identificar aquele com melhor desempenho na predição do uso de drogas ilícitas. As variáveis preditoras incluíram sexo, idade, planos futuros, uso de álcool e tabaco, condições de moradia, posse de bens, nível de escolaridade dos pais e hábitos familiares. **RESULTADOS:** Entre os modelos testados, a Regressão Logística apresentou a maior AUC-ROC (0,90), evidenciando melhor desempenho global. O Random Forest, entretanto, foi utilizado para avaliar a importância das variáveis devido à sua robustez interpretativa. Os principais fatores associados ao risco foram: uso de álcool, nível de escolaridade materna, apoio de colegas e pais, além do consumo de álcool pelos pais. **DISCUSSÃO:** Os achados confirmam o potencial da ML na identificação de padrões de risco, em consonância com estudos recentes e com os dados epidemiológicos nacionais. A inclusão de variáveis familiares e comportamentais reforça a relevância de estratégias preventivas voltadas ao ambiente escolar e doméstico. **CONCLUSÃO:** A aplicação de modelos de ML, especialmente a Regressão Logística, mostrou-se válida para prever o risco de consumo de drogas ilícitas em adolescentes. Esses resultados podem orientar políticas públicas direcionadas, priorizando fatores de risco modificáveis e otimizando o uso de recursos na saúde coletiva.

CC BY

DOI: 10.53843/bms.v11i15.1148

PALAVRAS-CHAVE: *Planejamento em Saúde; Aprendizagem de Máquina; Abuso Oral de Substâncias.*

ABSTRACT

INTRODUCTION: Machine Learning (ML) is a field of artificial intelligence that enables the development of algorithms capable of learning from data, without explicit programming. Advances in these techniques have enabled important applications in healthcare, including risk and behavior prediction. This study proposes the development of a predictive model to understand and predict illicit drug use among Brazilian teenagers aged 13 to 17, based on sociodemographic and behavioral patterns, with the aim of informing more effective public health policies. **METHODOLOGY:** This is an observational, cross-sectional study based on data from 165,838 students from the 2019 National School Health Survey (PeNSE). Different ML models were compared to identify the one with the best performance in predicting illicit drug use. Predictor variables included gender, age, future plans, alcohol and tobacco use, housing conditions, property ownership, parental educational backgrounds, and family habits. **RESULTS:** Among the models tested, Logistic Regression presented the highest AUC-ROC (0.90), demonstrating better overall performance. Random Forest, however, was used to assess the importance of variables due to its interpretive robustness. The main factors associated with risk were: alcohol use, maternal educational background, peer and parental support, and parental alcohol consumption. **DISCUSSION:** The findings confirm the potential of ML in identifying risk patterns, in line with recent studies and national epidemiological data. The inclusion of family and behavioral variables reinforces the relevance of preventive strategies targeted at the school and home environment. **CONCLUSION:** The application of ML models, especially Logistic Regression, proved to be valid for predicting the risk of illicit drug use in teenagers. These results can guide targeted public policies, prioritizing modifiable risk factors and optimizing the use of public health resources.

KEYWORDS: *Substance Abuse; Health Policy; Machine Learning*

INTRODUÇÃO

O consumo de substâncias psicoativas entre adolescentes constitui um desafio relevante de saúde pública, com impactos na saúde física, no desenvolvimento cognitivo, no bem-estar mental e na transição para a vida adulta [1]. Estima-se que 5,6% da população mundial entre 15 e 64 anos tenha experimentado drogas ao menos uma vez em 2016, sendo a prevalência geralmente maior entre jovens do que em adultos mais velhos [2].

No Brasil, dados da Pesquisa Nacional de Saúde Escolar (PeNSE) evidenciam crescimento progressivo da experimentação de substâncias. A proporção de adolescentes que relataram uso de bebidas alcoólicas aumentou de 52,9% em 2012 para 63,2% em 2019, enquanto a experimentação de drogas ilícitas passou de 8,2% em 2009 para 12,1% em 2019, com tendência média anual de crescimento de 4,5% [3]. Esses números reforçam a necessidade de abordagens inovadoras para identificação precoce e prevenção.

Diversos fatores têm sido associados ao consumo de substâncias na adolescência, incluindo sexo, idade, ambiente familiar, planos educacionais, posse de bens, hábitos alimentares, uso prévio de álcool e tabaco, além de comportamentos de pais e colegas e percepção de apoio social [1,2]. Estudos observacionais e revisões sistemáticas corroboram a relevância desses fatores, mas também apontam a limitação das intervenções preventivas tradicionais, que em geral apresentam benefícios modestos e pouco sustentados no tempo [1].

O Aprendizado de Máquina (Machine Learning — ML) é um ramo da inteligência artificial que se consolidou a partir da segunda metade do século XX e engloba métodos capazes de identificar padrões complexos em grandes conjuntos de dados, sem necessidade de programação explícita. Avanços recentes em capacidade computacional e disponibilidade de bases populacionais têm ampliado suas aplicações em saúde, desde previsão de risco cardiovascular e resposta a tratamentos

CC BY

DOI: 10.53843/bms.v11i15.1148

oncológicos até a estratificação de mortalidade em COVID-19 [9–11]. Um aspecto central para a adoção em saúde pública é a interpretabilidade dos modelos, permitindo traduzir resultados estatísticos em evidência útil para gestores e profissionais [8].

Embora fatores sociodemográficos e comportamentais associados ao uso de drogas na adolescência já sejam bem documentados, ainda há escassez de estudos que explorem o potencial preditivo do ML em bases populacionais brasileiras de larga escala, como a PeNSE. Trabalhos internacionais indicam que o ML é capaz de predizer comportamentos de risco entre adolescentes em diferentes contextos culturais [7], mas a aplicação a dados nacionais pode contribuir de forma inédita para compreender esse fenômeno e orientar políticas públicas mais eficazes.

Diante desse cenário, o presente estudo tem como objetivo avaliar o desempenho de diferentes modelos de ML na predição do consumo de drogas ilícitas entre adolescentes brasileiros de 13 a 17 anos, utilizando dados da PeNSE 2019, e identificar os fatores mais relevantes associados a esse comportamento, com vistas a subsidiar estratégias preventivas em saúde coletiva.

METODOLOGIA

Trata-se de estudo observacional, transversal, com dados secundários da Pesquisa Nacional de Saúde do Escolar (PeNSE) 2019, conduzida pelo Instituto Brasileiro de Geografia e Estatística (IBGE) em parceria com o Ministério da Saúde. A amostra foi composta por estudantes do 7º ano do ensino fundamental ao 3º ano do ensino médio, de escolas públicas e privadas de todo o território nacional, utilizando plano amostral complexo, por conglomerados em múltiplos estágios [3].

Para este estudo, foram incluídos adolescentes de 13 a 17 anos que responderam à questão referente ao uso de drogas ilícitas. Sendo excluídos participantes com idade <13 anos e aqueles com ausência de resposta na variável desfecho. O fluxograma de seleção da amostra está apresentado na **Figura 1**, indicando as etapas de exclusão e a composição final do banco analisado.

As variáveis preditoras foram selecionadas com base em disponibilidade na base de dados e em evidências prévias da literatura sobre fatores de risco e proteção associados ao consumo de drogas na adolescência [1,2,7]. Incluíram-se: sexo, idade, planos futuros, uso prévio de álcool e tabaco, coabitação com os pais, posse de bens, escolaridade materna, hábitos alimentares, número de amigos, percepção do convívio escolar, suporte parental e comportamento de uso de álcool pelos pais.

Valores faltantes em variáveis nominais foram codificados como “não informado”; variáveis ordinais foram convertidas em escala numérica; e variáveis contínuas com dados ausentes foram imputadas pela média como abordagem primária. Reconhecemos que essa técnica, embora simples e transparente, tende a reduzir variabilidade e atenuar associações; por isso, descrevemos explicitamente essa limitação e sugerimos, em análises futuras, imputação múltipla (MICE) ou imputação por vizinhos mais próximos (k-NN). O percentual de valores ausentes por variável encontra-se descrito na **Tabela 1**.

Para a Análise de colinearidade construiu-se a matriz de correlação de Pearson e foram removidas variáveis com coeficiente >0,70, a fim de reduzir redundância e risco de sobreajuste. As variáveis inicialmente analisadas e aquelas excluídas devido à colinearidade estão listadas na **Tabela 2**.

Na modelagem preditiva foram inicialmente explorados diferentes algoritmos de Aprendizado de Máquina: Regressão Logística, Random Forest, Extra Trees, Naive Bayes, K-Nearest Neighbors (KNN), Support Vector Machine (SVM) e XGBoost. Para apresentação final, priorizaram-se os modelos que melhor equilibraram desempenho, interpretabilidade e custo computacional:

- Regressão Logística, por aliar alto desempenho e interpretabilidade;
- Random Forest e Extra Trees, pela robustez e capacidade de ranking de importância de variáveis;
- Naive Bayes e KNN, como comparadores simples e tradicionalmente aplicados em epidemiologia.

O XGBoost, embora tenha apresentado desempenho comparável nos testes preliminares, foi mantido apenas em análises auxiliares devido à maior complexidade de ajuste e menor interpretabilidade clínica. Modelos de deep learning não foram empregados nesta versão, considerando o caráter tabular da base, o número limitado de variáveis com efeito robusto e a necessidade de interpretabilidade em saúde pública [8–11].

CC BY

DOI: 10.53843/bms.v11i15.1148

Todos os modelos foram treinados inicialmente com parâmetros padrão e, quando indicado, submetidos a ajuste de hiperparâmetros por grid search com validação cruzada. A validação foi realizada por 10-fold cross-validation (seed = 0), assegurando reprodutibilidade.

O desempenho foi avaliado por área sob a curva ROC (AUC-ROC), acurácia, precisão (valor preditivo positivo — PPV), valor preditivo negativo (NPV), sensibilidade (recall), especificidade e F1-score. O ponto de corte ótimo foi definido pelo índice de Youden. Além disso, foram analisados cenários com pontos de corte fixos em 0,25 e 0,50 para avaliar implicações operacionais na prática em saúde pública (trade-offs entre sensibilidade e especificidade). As curvas ROC dos principais modelos estão representadas na **Figura 2**, e os resultados detalhados encontram-se na **Tabela 2**.

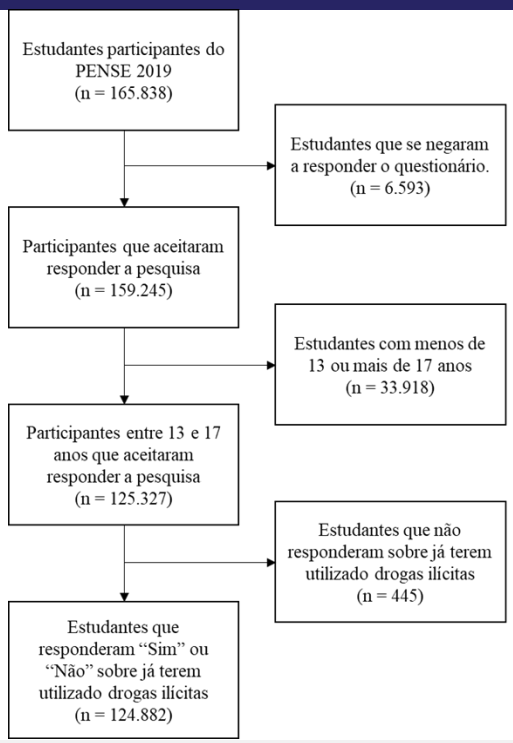
Por se tratar de base de dados secundária, de acesso público e anonimizada, não houve necessidade de novo consentimento. Para garantir reprodutibilidade, todas as análises foram realizadas com seed fixada e versões de bibliotecas devidamente documentadas.

RESULTADOS

Dos 165.838 estudantes inicialmente incluídos na base da PeNSE, 159.245 aceitaram participar da pesquisa. Após a aplicação dos critérios de inclusão, a amostra final compreendeu 124.965 adolescentes entre 13 e 17 anos (**Figura 1**).

A caracterização sociodemográfica (**Tabela 1**) evidenciou distribuição equilibrada por sexo (49,2% masculino), predomínio da faixa etária de 13 a 15 anos (65,8%) e autodeclaração parda (42,8%). Entre os comportamentos de risco, 62,1% relataram uso prévio de álcool e 25,2% uso de tabaco. Adicionalmente, variáveis relacionadas a contexto familiar e suporte social mostraram diferenças estatisticamente significativas entre os grupos de adolescentes que já haviam experimentado drogas ilícitas e aqueles que não haviam. A análise de colinearidade entre variáveis independentes (**Figura 2**) não demonstrou correlações acima de r = 0,7, sugerindo ausência de multicolinearidade relevante.

FIGURA 1. Fluxograma de inclusão e exclusão de participantes



Referência: Fonte prórpia (2025). Fluxograma de inclusão e exclusão dos participantes da Pesquisa Nacional de Saúde Escolar (PeNSE 2019) para composição da amostra final do estudo (n = 124.965). Fonte: autores (2025), com base na PeNSE 2019.

CC BY

DOI: 10.53843/bms.v11i15.1148

Tabela 1. Número de internações de crianças e adolescentes com transtornos mentais e comportamentais no Brasil e nas regiões brasileiras por ano de atendimento.

	Total		Experimentou drogas ilícitas		Não experimentou drogas ilícitas		Valor-p
	n	%	n	%	n	%	
Características sociodemográficas							
Sexo masculino, n (%)	61,417	49.18%	7,630	50.92%	53,787	48.94%	< 0.001
13 a 15 anos, n (%)	82,111	65.75%	6,241	41.65%	75,870	69.04%	< 0.001
16 a 17 anos, n (%)	42,402	33.95%	8,668	57.84%	33,734	30.70%	< 0.001
Branco, n (%)	47,512	38.05%	5,727	38.22%	41,785	38.02%	< 0.001
Preto, n (%)	13,532	10.84%	1,999	13.34%	11,533	10.49%	< 0.001
Amarelo, n (%)	4,390	3.52%	531	3.54%	3,859	3.51%	< 0.001
Pardo, n (%)	53,396	42.76%	6,003	40.06%	47,393	43.12%	< 0.001
Indígena, n (%)	3,505	2.81%	393	2.62%	3,112	2.83%	< 0.001
Uso de tabaco, n (%)	25,228	20.20%	11,974	79.91%	13,254	12.06%	< 0.001
Uso de álcool, n (%)	77,587	62.13%	14,605	97.46%	62,982	57.31%	< 0.001
Planos para o fim do ensino fundamental							
Somente continuar estudando	20,459	16.38%	743	4.96%	19,716	17.94%	< 0.001

CC BY

DOI: 10.53843/bms.v11i15.1148

Somente trabalhar	1,956	1.57%	287	1.92%	1,669	1.52%	< 0.001
Continuar estudando e trabalhar	33,090	26.50%	2,792	18.63%	30,298	27.57%	< 0.001
Planos para o fim do ensino médio							
Somente continuar estudando	18,304	14.66%	1,582	10.56%	16,722	15.22%	< 0.001
Somente trabalha	7,422	5.94%	1,116	7.45%	6,306	5.74%	< 0.001
Continuar estudando e trabalhar	83,125	66.56%	10,027	66.91%	73,098	66.52%	< 0.001
Condições de moradia							
Mora com a mãe, n (%)	110,947	88.84%	12,515	83.52%	98,432	89.57%	< 0.001
Mora com o pai, n (%)	78,614	62.95%	7,773	51.87%	70,841	64.46%	< 0.001
Possui celular, n (%)	109,451	87.64%	13,190	88.02%	96,261	87.59%	0.077
Possui Computador/ Notebook, n (%)	84,558	67.71%	10,020	66.87%	74,538	67.83%	0.013
Possui Internet, n (%)	113,672	91.02%	13,783	91.98%	99,889	90.89%	< 0.001
Possui Carro, n (%)	82,718	66.24%	9,752	65.08%	72,966	66.39%	< 0.001
Possui moto, n (%)	46,701	37.40%	5,396	36.01%	41,305	37.59%	< 0.001

CC BY

DOI: 10.53843/bms.v11i15.1148

Empregado doméstico, n (%)	21,509	17.22%	2,800	18.69%	18,709	17.02%	< 0.001
Número de banheiros em casa							
Nenhum, n (%)	21,924	1.54%	187	1.25%	1,737	1.58%	0.008
Um, n (%)	56,684	45.39%	6,903	46.07%	49,781	45.30%	0.008
Dois, n (%)	41,120	32.93%	4,873	32.52%	36,247	32.98%	0.008
Três, n (%)	15,752	12.61%	1,859	12.41%	13,893	12.64%	0.008
Quatro ou mais, n (%)	9,343	7.48%	1,159	7.73%	8,184	7.45%	0.008
Escolaridade da mãe							
Nenhuma, n (%)	3,293	2.64%	445	2.97%	2,848	2.59%	< 0.001
Ensino fundamental incompleto, n (%)	14,837	11.88%	1,956	13.05%	12,881	11.72%	< 0.001
Ensino fundamental completo, n (%)	5,834	4.67%	703	4.69%	5,131	4.67%	< 0.001
Ensino médio incompleto, n (%)	7,668	6.14%	1,043	6.96%	6,625	6.03%	< 0.001
Ensino médio completo, n (%)	24,473	19.60%	3,161	21.09%	21,312	19.39%	< 0.001
Ensino superior incompleto, n (%)	9,034	7.23%	1,205	8.04%	7,829	7.12%	< 0.001
Ensino superior completo, n (%)	39,973	32.01%	4,634	30.92%	35,339	32.16%	< 0.001

CC BY

DOI: 10.53843/bms.v11i15.1148

Café da manhã							
Todos os dias	68,437	54.80%	6,346	42.35%	62,091	56.50%	< 0.001
5-6 dias/semana	5,607	4.49%	761	5.08%	4,846	4.41%	< 0.001
3-4 dias/semana	5,657	4.53%	880	5.87%	4,777	4.35%	< 0.001
1-2 dias/semana	4,557	3.65%	614	4.10%	3,943	3.59%	< 0.001
Raramente	28,183	22.57%	4,197	28.01%	23,986	21.83%	< 0.001
Almoço com os pais							
Todos os dias	74,829	59.92%	6,648	44.36%	68,181	62.04%	< 0.001
5-6 dias/semana	7,220	5.78%	957	6.39%	6,263	5.70%	< 0.001
3-4 dias/semana	7,959	6.37%	1,267	8.46%	6,692	6.09%	< 0.001
1-2 dias/semana	7,125	5.71%	1,020	6.81%	6,105	5.56%	< 0.001
Raramente	18,936	15.16%	3,228	21.54%	15,708	14.29%	< 0.001
Pais fumantes							
Nenhum, n (%)	98,824	79.13%	10,300	68.74%	88,524	80.55%	< 0.001
Apenas responsável masculino, n (%)	13,110	10.50%	2,099	14.01%	11,011	10.02%	< 0.001

CC BY

DOI: 10.53843/bms.v11i15.1148

Apenas responsável feminino, n (%)	7,079	5.67%	1,366	9.12%	5,713	5.20%	< 0.001
Ambos responsáveis, n (%)	3,691	2.96%	859	5.73%	2,832	2.58%	< 0.001
Número de amigos próximos							
Nenhum, n (%)	4,195	3.36%	684	4.56%	3,511	3.19%	< 0.001
Um, n (%)	6,910	5.53%	982	6.55%	5,928	5.39%	< 0.001
Dois, n (%)	16,926	13.55%	2,435	16.25%	14,491	13.19%	< 0.001
Três ou mais, n (%)	96,610	77.36%	10,838	72.33%	85,772	78.05%	< 0.001
NOS ÚLTIMOS 30 DIAS, com que frequência os colegas de sua escola trataram você bem e/ou foram prestativos com você?							
Nunca, n (%)	6,718	5.38%	1,016	6.78%	5,702	5.19%	< 0.001
Raramente, n (%)	11,529	9.23%	1,721	11.48%	9,808	8.92%	< 0.001
Às vezes, n (%)	24,338	19.49%	3,051	20.36%	21,287	19.37%	< 0.001
Na maioria das vezes, n (%)	46,678	37.38%	5,399	36.03%	41,279	37.56%	< 0.001
Sempre, n (%)	35,233	28.21%	3,735	24.92%	31,498	28.66%	< 0.001
NOS ÚLTIMOS 30 DIAS, com que frequência sua mãe, pai ouresponsável entendeu seus problemas e preocupações?							
Nunca, n (%)	19,512	15.62%	3,542	23.64%	15,970	14.53%	< 0.001
Raramente, n (%)	21,654	17.34%	3,268	21.81%	18,386	16.73%	< 0.001
Às vezes, n (%)	27,455	21.98%	3,353	22.38%	24,102	21.93%	< 0.001

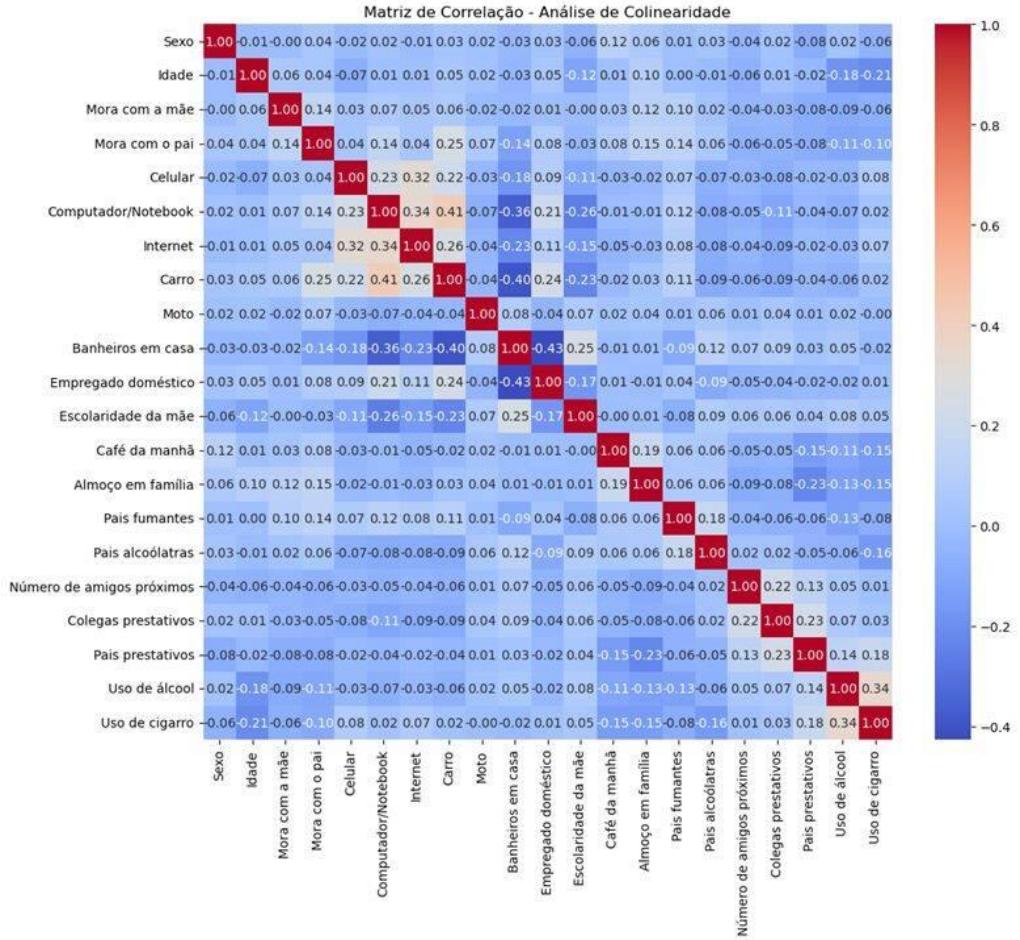
CC BY

DOI: 10.53843/bms.v11i15.1148

Na maioria das vezes, n (%)	29,177	23.36%	2,752	18.37%	26,425	24.05%	< 0.001
Sempre, n (%)	26,572	21.28%	2 ,009	13.41%	24,563	22.35%	< 0.001

Legenda: Distribuição sociodemográfica e comportamental dos adolescentes incluídos na amostra final (n = 124.965). Dados expressos em frequências relativas (%). Fonte: autores (2025), com base na PeNSE 2019.

FIGURA 2. Análise de colinearidade



Referência: Matriz de correlação de Pearson entre variáveis preditoras. Nenhuma variável apresentou correlação >0,70, indicando ausência de colinearidade relevante. ROC = Receiver Operating Characteristic; AUC = Area Under the Curve. Fonte: autores (2025), com base na PeNSE 2019.

No desempenho dos modelos, a Regressão Logística apresentou a maior área sob a curva ROC (AUC = 0,90), seguida por Random Forest, Extra Trees e Naive Bayes (todos com AUC = 0,89), enquanto o K-Nearest Neighbors obteve AUC inferior (0,83). As curvas ROC comparativas estão apresentadas na **Figura 3**.

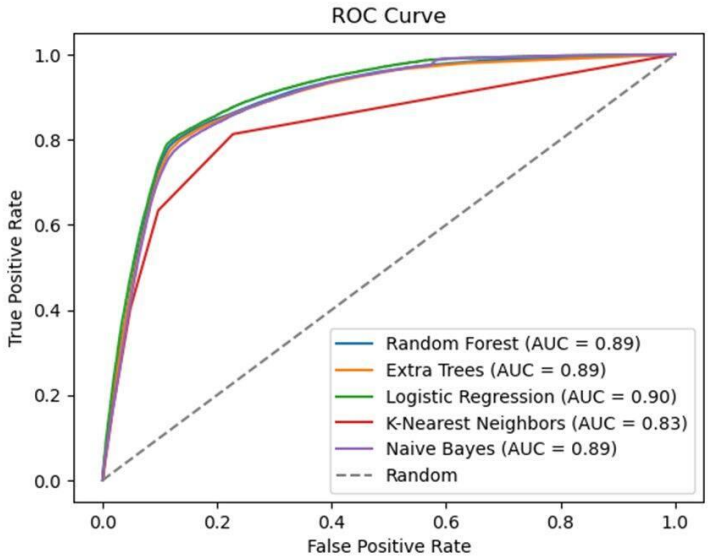
As métricas detalhadas em diferentes pontos de corte encontram-se na **Tabela 2**. No limiar de 0,25, a Regressão Logística apresentou a maior acurácia (0,88), o Naive Bayes destacou-se em sensibilidade (0,82) e o KNN em especificidade (0,90). No ponto de corte de 0,50, Regressão Logística, Random Forest e Extra Trees alcançaram acurácia de 0,89 e especificidade de 0,96, enquanto o Naive Bayes manteve maior sensibilidade (0,79). Esses resultados evidenciam diferentes trade-offs, úteis em cenários de triagem (alta sensibilidade) ou confirmação diagnóstica (alta especificidade).

Para a análise de importância das variáveis, utilizou-se o Random Forest, em razão de sua robustez interpretativa. As cinco variáveis de maior impacto na predição foram: uso prévio de álcool, escolaridade materna, presença de colegas prestativos, suporte parental e consumo de álcool pelos pais (**Figura 4**)

CC BY

DOI: 10.53843/bms.v11i15.1148

FIGURA 3. Curvas ROC comparando modelos de Machine Learning



Referência: Curvas ROC (Receiver Operating Characteristic) para os modelos Regressão Logística, Random Forest, Extra Trees, Naive Bayes e K-Nearest Neighbors na predição do consumo de drogas ilícitas entre adolescentes (PeNSE 2019). AUC = Área sob a curva. Fonte: autores (2025), com base na PeNSE 2019.

Tabela 2. Desempenho dos modelos de Machine Learning.

Modelo	AUC	Threshold	Positive Predictive Value (Precisão)	Negative Predictive Value	Sensitivity (Recall)	Specificity	Accuracy	F1 Score
Random Forest	0.89	0.19	0.46	0.97	0.8	0.87	0.86	0.59
		0.25	0.48	0.97	0.78	0.89	0.87	0.6
		0.5	0.57	0.92	0.5	0.96	0.89	0.47
Extra Trees	0.89	0.17	0.44	0.97	0.8	0.86	0.85	0.57
		0.25	0.48	0.97	0.77	0.89	0.87	0.59

CC BY

DOI: 10.53843/bms.v11i15.1148

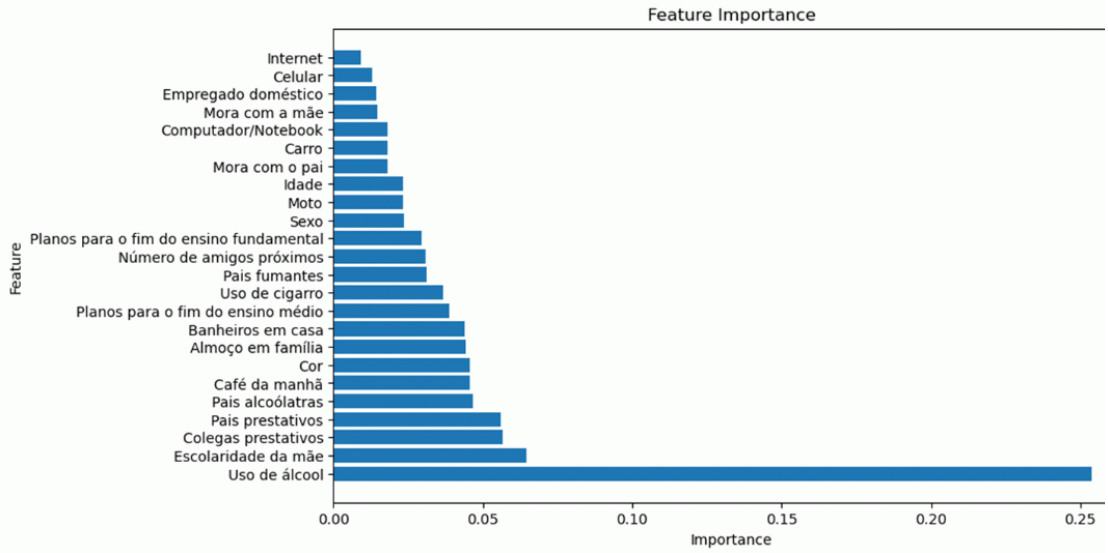
		0.5	0.56	0.92	0.41	0.96	0.89	0.47
Logistic Regression	**0.9**	0.11	0.47	0.97	0.8	0.88	0.87	0.6
		0.25	0.49	0.97	0.78	0.89	0.88	0.6
		0.5	0.58	0.93	0.45	0.96	0.89	0.51
K-Nearest Neighbors	**0.83**	0.2	0.33	0.97	0.81	0.77	0.78	0.47
		0.25	0.47	0.95	0.63	0.9	0.87	0.54
		0.5	0.53	0.92	0.4	0.95	0.89	0.46
Naive Bayes	**0.89**	0.35	0.42	0.97	0.8	0.85	0.84	0.55
		0.25	0.39	0.97	0.82	0.83	0.83	0.53
		0.5	0.44	0.97	0.79	0.87	0.86	0.57

Legenda: Desempenho dos modelos de Machine Learning (Regressão Logística, Random Forest, Extra Trees, Naive Bayes e KNN) na predição do consumo de drogas ilícitas entre adolescentes brasileiros. Resultados apresentados por área sob a curva ROC (AUC), acurácia, sensibilidade e especificidade em diferentes pontos de corte (0,25 e 0,50). ROC = Receiver Operating Characteristic; AUC = Área sob a curva; KNN = K-nearest neighbors. Fonte: autores (2025), com base na PeNSE 2019.

FIGURA 4. Importância das variáveis pelo Random Forest

CC BY

DOI: 10.53843/bms.v11i15.1148



Referência: Ranking das variáveis preditoras de maior impacto no modelo Random Forest: uso prévio de álcool, escolaridade materna, presença de colegas prestativos, suporte parental e consumo de álcool pelos pais. Fonte: autores (2025), com base na PeNSE 2019.

DISCUSSÃO

O presente estudo utilizou dados da Pesquisa Nacional de Saúde Escolar (PeNSE 2019) para desenvolver modelos de Aprendizado de Máquina (Machine Learning, ML) voltados à predição do consumo de drogas ilícitas entre adolescentes brasileiros. A Regressão Logística apresentou o melhor desempenho global (AUC-ROC = 0,90), enquanto o Random Forest foi empregado para análise de importância das variáveis, permitindo interpretação mais clara dos fatores associados. Entre os principais preditores identificados destacaram-se o uso prévio de álcool, a escolaridade materna, o suporte de colegas e pais e o consumo de álcool pelos responsáveis.

A análise das curvas ROC (**Figura 3**) demonstrou desempenho superior da Regressão Logística em relação à linha de referência, com valores próximos para Random Forest, Extra Trees e Naive Bayes, e desempenho inferior do KNN (AUC = 0,83). A interpretação dos pontos de corte (**Tabela 2**) é particularmente relevante em termos de aplicabilidade prática: no limiar de 0,25, a maior sensibilidade dos modelos — sobretudo do Naive Bayes — sugere utilidade em estratégias de triagem em saúde pública, em que a prioridade é identificar o maior número possível de adolescentes em risco, mesmo com custo de maior número de falsos positivos. Já no limiar de 0,50, o ganho em especificidade (até 0,96 para Regressão Logística, Random Forest e Extra Trees) pode ser mais adequado em contextos de alocação restrita de recursos, priorizando intervenções em casos de maior probabilidade real de uso. Essa flexibilidade reforça o potencial dos modelos de ML como ferramentas adaptáveis às necessidades de diferentes políticas públicas.

Esses achados estão em consonância com estudos internacionais que empregaram ML na predição de comportamentos de risco. Afzali et al. demonstraram que modelos baseados em regularização, como a rede elástica, foram eficazes na previsão do uso de álcool entre jovens, destacando a relevância de fatores individuais e contextuais [7]. Revisões sistemáticas, como a de Nawi et al., reforçam o papel central de fatores familiares e comunitários no uso de substâncias [2]. Do mesmo modo, Tinner et al. destacaram a limitada efetividade de intervenções preventivas tradicionais em ambientes escolar e familiar, o que sustenta o papel complementar do ML em oferecer abordagens mais direcionadas [1].

Outro aspecto importante foi a análise da importância das variáveis, realizada pelo Random Forest (**Figura 4**). O uso prévio de álcool foi o preditor mais forte, em linha com evidências de que a experimentação precoce dessa substância aumenta a probabilidade de envolvimento posterior com drogas ilícitas [2,7]. Fatores familiares, como escolaridade materna e suporte parental, reforçam o papel do ambiente doméstico na proteção ou no risco para o uso de drogas, enquanto a presença de colegas

CC BY

DOI: 10.53843/bms.v11i15.1148

prestativos indica a relevância das redes sociais na adolescência. Esses achados ressaltam a necessidade de políticas intersetoriais que integrem ações escolares, familiares e comunitárias.

Apesar de suas contribuições, o estudo apresenta limitações que devem ser consideradas. O delineamento transversal não permite inferir relações causais entre os fatores preditivos e o consumo de drogas ilícitas. A utilização de dados autorrelatados pode introduzir viés de aferição, com possibilidade de subnotificação ou superestimação. Além disso, a ausência de variáveis psicossociais complexas, como diagnósticos psiquiátricos formais, traços de personalidade ou histórico detalhado de adversidades, pode ter reduzido a capacidade preditiva dos modelos. Outro ponto refere-se à imputação pela média utilizada como estratégia primária para lidar com valores ausentes, reconhecida como abordagem simples e transparente, mas que pode atenuar variabilidade e associações. Estudos futuros poderiam empregar métodos mais robustos, como imputação múltipla. Por fim, a validade externa do modelo é limitada, uma vez que treinamento e teste foram realizados na mesma base (PeNSE 2019); ainda que a validação cruzada reduza esse problema, será essencial validar os modelos em bases independentes e longitudinais.

Em síntese, nossos resultados reforçam a utilidade do ML como ferramenta complementar na saúde pública, oferecendo não apenas maior precisão na identificação de adolescentes em risco, mas também flexibilidade para diferentes cenários de aplicação. Ao fornecer subsídios para triagem precoce em escolas e orientação de estratégias familiares, esses modelos podem contribuir para o uso mais eficiente de recursos e para o desenho de políticas públicas mais custo-efetivas.

CONCLUSÃO

A aplicação de modelos de Machine Learning mostrou-se viável para predizer o risco de consumo de drogas ilícitas entre adolescentes brasileiros. A Regressão Logística apresentou o melhor desempenho global, enquanto o Random Forest contribuiu para interpretação dos fatores de risco.

Os resultados reforçam a importância de fatores familiares e sociais — em especial uso prévio de álcool, suporte parental e escolaridade materna — como potenciais alvos de intervenção.

Tais evidências podem subsidiar estratégias preventivas em larga escala, otimizar a alocação de recursos e orientar políticas públicas voltadas a adolescentes em maior vulnerabilidade. Estudos longitudinais e a validação externa em diferentes coortes são recomendados para fortalecer a generalização dos achados.

CONFLITOS DE INTERESSE

Os autores declaram não haver conflito de interesse no desenvolvimento da pesquisa.

FINANCIAMENTO

Os autores declaram não haver financiamento no desenvolvimento da pesquisa.

REFERÊNCIAS

1. Tinner L, Palmer JC, Lloyd EC, Caldwell DM, MacArthur GJ, Dias K, et al. Individual-, family- and school-based interventions to prevent multiple risk behaviours relating to alcohol, tobacco and drug use in young people aged 8–25 years: a systematic review and meta-analysis. BMC Public Health. 2022;22(1):1111. doi:10.1186/s12889-022-13072-5.

2. Nawi AM, Ismail R, Ibrahim F, Hassan MR, Manaf MRA, Amit N, et al. Risk and protective factors of drug abuse among adolescents: a systematic review. BMC Public Health. 2021;21(1):2088. doi:10.1186/s12889-021-11906-2.

3. Instituto Brasileiro de Geografia e Estatística (IBGE). Pesquisa Nacional de Saúde do Escolar: 2019. Rio de Janeiro: IBGE; 2021. Disponível em: https://biblioteca.ibge.gov.br/visualizacao/livros/liv101852.pdf. Acesso em: 20 ago 2025.

4. Artero AS. Aprendizado de máquina: conceitos e algoritmos. In: XXII Conference on Graphics, Patterns and Images – SIBGRAPI. IEEE; 2009. p. 215–28.

CC BY

DOI: 10.53843/bms.v11i15.1148

5. Haykin S. Redes neurais: princípios e prática. 2a ed. Porto Alegre: Bookman; 2001.

6. Silva IN, Spatti DH, Flauzino RA. Redes neurais artificiais para engenharia e ciências aplicadas: curso prático. São Paulo: Artliber; 2010.

7. Afzali MH, Sunderland M, Stewart S, Massé B, Séguin JR, Newton N, et al. Machine-learning prediction of adolescent alcohol use: a cross-study, cross-cultural validation. Addiction. 2019;114(4):662–71. doi:10.1111/add.14504.

8. Vellido A. The importance of interpretability and visualization in machine learning for applications in medicine and health care. Neural Comput Appl. 2020;32:18069–83. doi:10.1007/s00521-019-04051-w.

9. Rajkomar A, Dean J, Kohane I. Machine learning in medicine. N Engl J Med. 2019;380(14):1347–58. doi:10.1056/NEJMra1814259.

10. Beam AL, Kohane IS. Big data and machine learning in health care. JAMA. 2018;319(13):1317–8. doi:10.1001/jama.2017.18391.

11. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56. doi:10.1038/s41591-018-0300-7.

12. Patil RC, Tavaragi MS, Sushma C. Inhalant abuse in adolescents in North Karnataka: a case series. J Psychiatry Spectrum. 2022;1(2):133–5. doi:10.4103/jopsys.jopsys_1_22.

13. Karlovšek MZ, Alibegović A, Balažic J. Our experiences with fatal ecstasy abuse (two case reports). Forensic Sci Int. 2005;147 Suppl:S77–80. doi:10.1016/j.forsciint.2004.09.084.

14. Fineschi V, Masti A. Fatal poisoning by MDMA (ecstasy) and MDEA: a case report. Int J Legal Med. 1996;108(5):272–5. doi:10.1007/BF01369826.

15. Salehi F, Hassanzadeh Taheri MM, Riasi H, Mehrpour O. Recurrent syncope following substance abuse: a case report. Emergency. 2017;5(1):e47. PMID:28286854; PMCID:PMC5325918.

16. Hoorn EJ; Paediatric Death Review Committee, Office of the Chief Coroner of Ontario. A fatal case of ecstasy poisoning. Paediatr Child Health. 2001;6(7):491. doi:10.1093/pch/6.7.491.

CC BY