



# Análise dos Fatores que Influenciam o Consumo de Álcool entre os Estudantes de Matemática Utilizando Regressão Logística

Patrícia Virgínia de Santana Lima<sup>1</sup>, <https://orcid.org/0009-0005-4746-830X>

Denilson de Oliveira Silva<sup>2</sup>, <https://orcid.org/0009-0000-4031-7772>

Sílvio Fernando Alves Xavier<sup>3</sup>, <https://orcid.org/0000-0002-4832-0711>

<sup>1</sup>Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, Brasil.

Email: [patricia.lima@aluno.uepb.edu.br](mailto:patricia.lima@aluno.uepb.edu.br);

<sup>2</sup>Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, Brasil.

Email: [denilson.oliveira.silva@aluno.uepb.edu.br](mailto:denilson.oliveira.silva@aluno.uepb.edu.br);

<sup>3</sup>Departamento de Estatística, Universidade Estadual da Paraíba, Campina Grande, Brasil.

Email: [silvio@servidor.uepb.edu.br](mailto:silvio@servidor.uepb.edu.br).

## RESEARCH ARTICLE

Annals of Data Analysis

<https://periodicos.ufpe.br/revistas/ADA/>

Submitted on: 12.12.2025.

Accepted on: 01.08.2026.

Published on: 01.08.2026.

Copyright © 2026 by author(s).

This work is licensed under the Creative Commons Attribution International License CC BY-NC-ND 4.0

<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>



## Abstract

The study investigated alcohol consumption among high school students, focusing on mathematics majors. Logistic regression was used as a statistical method suitable for binary variables, allowing for the prediction and classification of drinking behavior during school days. Factors such as academic performance, extracurricular activities, social relationships, and family environment were analyzed. The data used were obtained from Kaggle, enabling practical and contextualized application. Comparative modeling of the main determinants of consumption was performed. The most accurate model was selected to predict the observed patterns. The results obtained allow for a better understanding of the characteristics and support strategies and school interventions.

## Keywords

**Keywords:** Alcohol, modeling, performance, prediction.

## 1. Introdução

No Brasil, estudos evidenciam a gravidade da situação, uma pesquisa realizada com alunos do 6.º ao 9.º ano apontou que cerca de 66,6 % dos adolescentes já haviam experimentado bebidas alcoólicas antes dos 18 anos (Souza et al., 2022). De forma semelhante, o Jornal de Pediatria registrou que aproximadamente 56 % dos adolescentes brasileiros relataram o consumo precoce de álcool, reforçando os riscos associados a essa prática (Malta et al., 2022). Diante

desse cenário, compreender os fatores determinantes do consumo em dias letivos torna-se essencial. Para isso, este estudo utiliza a regressão logística, uma técnica estatística adequada para variáveis binárias, como presença ou ausência de consumo. Tal abordagem permite identificar variáveis significativas como desempenho acadêmico, atividades extracurriculares, relações sociais e ambiente familiar além de avaliar seu impacto sobre o comportamento de consumo. A aplicação desse método também possibilita selecionar o modelo mais preciso, contribuindo para a formulação de estratégias de prevenção e intervenção no ambiente escolar.

Além disso, a regressão logística é amplamente empregada em estudos de saúde pública, pois permite estimar probabilidades e identificar fatores de risco associados a desfechos clínicos. Um exemplo é o estudo de Cardoso et al. (2023), que utilizou regressão logística para analisar a mortalidade por COVID-19 em adultos hospitalizados no Brasil, e também em ciências sociais, sendo usada para compreender fenômenos sociais complexos. Um exemplo é o trabalho de Góes, Ramos e Ferreira (2020), que discutiu, a partir de análises sobre desigualdades raciais durante a pandemia de COVID-19 por sua capacidade de estimar probabilidades e fornecer medidas de associação entre fatores de risco e desfechos de interesse. Ao contrário de métodos meramente descritivos, esse modelo permite um olhar preditivo, oferecendo informações que podem subsidiar políticas educacionais e de saúde voltadas ao público adolescente. Outro aspecto relevante é a possibilidade de comparar diferentes variáveis e identificar quais delas possuem maior peso na determinação do consumo de álcool. Isso favorece a priorização de ações preventivas, seja no âmbito familiar, com orientações aos pais e responsáveis, seja na escola, com programas de conscientização e promoção de hábitos saudáveis. Ademais, compreender os fatores que influenciam esse comportamento pode contribuir para a criação de estratégias intersetoriais, envolvendo tanto a educação quanto a saúde, fortalecendo a rede de proteção aos adolescentes (SENNA et al., 2021). Portanto, a aplicação da regressão logística neste contexto não apenas atende a uma demanda metodológica, mas também oferece subsídios práticos para a compreensão de um fenômeno complexo, marcado por múltiplas variáveis e forte influência do contexto social.

Diante do exposto, torna-se evidente que a adoção de modelos estatísticos robustos, como a regressão logística, assume papel central na investigação dos determinantes do consumo de álcool entre adolescentes. Estudos nacionais já apontaram uma prevalência elevada de experimentação e consumo entre jovens, por exemplo, aproximadamente 50,3 % de adolescentes brasileiros relataram consumo de bebidas alcoólicas no último ano, sendo 36 % associados a episódios de binge drinking (SILVA et al., 2018). Além disso, pesquisas realizadas em contextos escolares e rurais identificaram variáveis como baixo desempenho escolar, menor supervisão parental e contexto socioeconômico desfavorável como fatores associados ao uso de álcool (RODRIGUES; COSTA; MOURA, 2021). Neste cenário, a regressão logística não apenas permite estimar probabilidades de ocorrência de consumo, mas também quantificar a influência relativa de múltiplos fatores simultaneamente, o que facilita a identificação de perfis de maior risco e a priorização de intervenções em escolas e comunidades. Portanto, este estudo propõe a utilização desse método para explorar os determinantes do consumo de bebidas alcoólicas em dias letivos entre adolescentes brasileiros, com o intuito de apoiar o delineamento de programas de prevenção mais focalizados e efetivos.

## 2. Objetivo

Investigar os fatores associados ao consumo de álcool em dias de aula entre estudantes, identificando grupos de maior risco e padrões de comportamento, com a intenção de subsidiar políticas educacionais e estratégias de intervenção para reduzir esse consumo.

### 2.1. Objetivos específicos

- Avaliar como variáveis independentes (como escolaridade dos pais, apoio educacional, e qualidade das relações familiares) influenciando o consumo de álcool no dia de aula. Isso pode ajudar a identificar grupos em maior risco.
- Verificar se aqueles estudantes que saem mais e participam de mais atividades fora da escola têm menos ou mais tendência a consumir álcool. O objetivo é identificar esses padrões e entender melhor como esses fatores influenciam esse comportamento.
- Avaliar se a proporção de alunos que consomem álcool em dias de aula difere significativamente entre grupos (exemplo: entre diferentes escolas ou tipos de endereço).

## 3. Materias e Métodos

Como o intuito deste estudo é avaliar como diversas variáveis influenciam o consumo de álcool em dias de aula entre estudantes de matemática. Para isso, utilizaremos um banco de dados reais obtidos no kaggle<sup>1</sup>, contendo 395 observações e 27 variáveis descritas em anexo. As variáveis idade, número de faltas escolares e, entre outras quantitativas, já a escola, o estado civil dos pais e as demais são variáveis categóricas. A seguir, na Tabela 1, apresentamos uma descrição detalhada das variáveis que compõem o nosso banco de dados. Cada variável é especificada quanto ao seu tipo, unidade de medida e possíveis categorias, permitindo uma compreensão clara dos dados utilizados na análise.

**Tabela 1:** Descrição das variáveis da base de dados.

| Variável                              | Descrição                                                                          |
|---------------------------------------|------------------------------------------------------------------------------------|
| Escola                                | Gabriel Pereira ou Mousinho da Silveira                                            |
| Sexo                                  | feminino ou masculino                                                              |
| Idade                                 | de 15 a 22                                                                         |
| Endereço                              | urbano ou rural                                                                    |
| Tamanho da família                    | menor ou igual a 3 ou maior que 3                                                  |
| Estado civil dos pais                 | vivendo juntos ou separados                                                        |
| Escolaridade da mãe                   | nenhuma, ensino fundamental 4ª série, 5ª a 9ª série, ensino médio, ensino superior |
| Escolaridade do pai                   | nenhuma, ensino fundamental 4ª série, 5ª a 9ª série, ensino médio, ensino superior |
| Trabalho da mãe                       | em casa, outros, professor, saúde, serviços                                        |
| Trabalho do pai                       | em casa, outros, professor, saúde, serviços                                        |
| Tutor do aluno                        | (mãe, pai ou outro)                                                                |
| Tempo de viagem de casa para a escola | menos de 15 min, 15 a 30 min., 30 min a 1 hora, ou mais de 1 hora                  |
| Tempo de estudo semanal               | menos de 2 horas, 2 a 5 horas, 5 a 10 horas ou mais de 10 horas                    |
| Número de reprovações                 | Vai de 0 até 3                                                                     |
| Suporte educacional extra             | sim ou não                                                                         |
| Apoio educacional familiar            | sim ou não                                                                         |

<sup>1</sup> <https://www.kaggle.com/datasets/uciml/student-alcohol-consumption>

|                                    |                                                           |
|------------------------------------|-----------------------------------------------------------|
| Atividades extracurriculares       | sim ou não                                                |
| Frequentou creche                  | sim ou não                                                |
| Tem acesso à internet em casa      | sim ou não                                                |
| Está em um relacionamento          | sim ou não                                                |
| Qualidade das relações familiares  | muito ruim, ruim, regular, boa, excelente                 |
| Tempo livre depois da escola       | adequado, excelente, insuficiente limitado, moderado      |
| Sair com amigos                    | bastante, moderadamente, muito pouco, pouco, sai muito    |
| Consumo de álcool em dia de aula   | binária: Não consome ou consome - classificada como 0 e 1 |
| Consumo de álcool no fim de semana | bastante, moderadamente, muito, não consome, pouco        |
| Estado de saúde atual              | bom, excelente, muito ruim, regular, Ruim                 |
| Número de faltas escolares         | numérico: de 0 a 75                                       |

Fonte: Elaborado pelos Autores, (2025).

### 3.1. Métricas de ajustes para o modelo

O método stepwise é uma técnica de seleção de variáveis amplamente utilizada em análises de regressão. Que consiste na inclusão ou exclusão de variáveis em etapas, passando a identificar o modelo mais adequado com base nos dados disponíveis. Existem três principais abordagens desse método: seleção progressiva “forward stepwise”, seleção regressiva “backward stepwise”, e uma combinação de ambas “both stepwise” (DRAPER & SMITH, 1998).

No que diz respeito ao ajuste do modelo, uma métrica utilizada é o Pseudo  $R^2$ , Uma métrica comumente utilizada para avaliar o ajuste de modelos em contextos de regressão logística é o pseudo  $R^2$ , sendo o índice de McFadden um dos mais populares. Esse índice varia de 0 a 1, onde valores mais próximos de 1 indicam um melhor ajuste do modelo (McFadden, 1974; Long, 1997).

Para complementar a avaliação de modelos preditivos a Curva ROC, é um gráfico amplamente utilizado para avaliar o desempenho de modelos de classificação binária. Essa curva é gerada ao plotar a taxa de verdadeiros positivos (sensibilidade) contra a taxa de falsos positivos (1 - especificidade) em vários limiares de decisão, facilitando a visualização da capacidade discriminatória do modelo. A área Sob a Curva ROC, que é o(AUC) representa uma métrica quantitativa que indica a precisão geral do modelo, onde valores mais próximos de 1 indicam melhor desempenho (HANLEY e MCNEIL, 1982). Um valor de AUC igual a 1 sugere um modelo perfeito, enquanto um valor de 0,5 indica um desempenho equivalente ao acaso, valores de AUC acima de 0,7 já são consideráveis aceitáveis.

No mesmo contexto, o Critério de Informação de Akaike (AIC) é uma ferramenta estatística utilizada para avaliar a qualidade de um modelo, levando em consideração tanto o ajuste do modelo aos dados quanto à sua complexidade. Ele penaliza modelos mais complexos, favorecendo aqueles que descrevem os dados de forma eficiente com um menor número de parâmetros. A AIC é amplamente empregada na seleção de modelos, onde valores mais baixos indicam uma melhor qualidade do modelo, e mensura a qualidade de um modelo estatístico visando também a sua simplicidade (AKAIKE, 1973).

$$AIC = 2k - 2Ln(L)$$

Onde,

- $K$  = Número de parâmetros.
- $L$  = O valor máximo da função de verossimilhança

Por fim, a Matriz de confusão é amplamente utilizada na prática, pois não permite apenas avaliar o desempenho global do modelo, mas também identificar tipos específicos de erros, o que pode fornecer insights detalhados para melhorias futuras. Em suma, ela é uma ferramenta indispensável para entender a eficácia de modelos de classificação em diversas aplicações (SUGIYAMA et al., 2012), dada por:

**Figura 1:** Matriz de confusão.

|       |           | Preditos    |           |
|-------|-----------|-------------|-----------|
|       |           | Verdadeiros | Falsos    |
| Reais | Positivos |             | Negativos |
|       | Falsos    | Verdadeiros | Negativos |

Fonte: Elaborado pelos Autores, (2025).

Em que:

- **Verdadeiros Positivos (VP):** Número de instâncias que foram corretamente previstas como positivas.
- **Falsos Negativos (FN):** Número de instâncias que são realmente positivas, mas foram previstas como negativas.
- **Falsos Positivos (FP):** Número de instâncias que são realmente negativas, mas foram previstas como positivas.
- **Verdadeiros Negativos (VN):** Número de instâncias que foram corretamente previstas como negativas.

### 3.2. Modelo logístico

O modelo nos proporciona informações valiosas sobre os determinantes no consumo de álcool dos estudantes de matemática, que é composto por,

$$\text{Logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1\chi_1 + \dots + \beta_i\chi_i + \varepsilon_i;$$

Em que:

- $P$ : a probabilidade do consumo de álcool em dia de aula;
- $\left(\frac{p}{1-p}\right)$ : é a razão de chance do consumo de álcool em dias de aula;
- $\beta_0$ : o intercepto;
- $\beta_1, \beta_2, \dots, \beta_n$ : são os coeficientes das variáveis independentes  $\chi_1, \chi_2, \dots, \chi_n$ ;

- $\varepsilon_i$ : é o erro Aleatório.

### 3.3. Função de ligação

A função de ligação é uma função matemática que conecta a média da variável de resposta com os preditores, sendo essencial para modelar a relação entre eles. Ela transforma a escala da resposta variável, permitindo que o modelo seja configurado corretamente. Existem diferentes tipos de função de ligação, dependendo da natureza dos dados:

- **Logit** : Usado para dados binários, como comprar ou não um determinado produto, onde a variável resposta tem dois valores possíveis.
- **Log** : Aplicada a dados de contagem, como o número de pacientes atendidos em um hospital ou quantidade de vendas de um produto, onde a variável resposta é um número inteiro positivo.
- **Identidade** : Utilizada para dados contínuos, especialmente quando a variável resposta não é simétrica, como peso de recém-nascidos, em uma população, tempo de recuperação após uma cirurgia ou gastos mensais de famílias com alimentação.
- Além disso, o componente sistemático inclui os preditores, que são as variáveis independentes usadas para modelar a média da resposta, ou seja, elas explicam como as mudanças nos preditores afetam o comportamento da variável resposta.

Além disso, a multicolinearidade entre as variáveis independentes foi feito através do cálculo do Fator de Inflação da Variância (VIF), verificando se os valores estavam dentro do limite  $1 < VIF < 3$  para ser aceito o pressuposto da multicolinearidade, em seguida Também será analisado os resíduos através de um gráfico de envelope para verificar se estavam majoritariamente contidos dentro dos limites esperados, indicando que a maioria dos dados segue o padrão esperado. o que, segundo Montgomery, Peck e Vining (2012), sugere que a maior parte dos dados segue o padrão assumido pelo modelo, sem violações significativas dos pressupostos estatísticos.

### 3.4. treino e teste

Em modelos estatísticos e de aprendizado de máquina, a separação dos dados em conjuntos de treino e teste é uma etapa essencial para avaliar a capacidade preditiva e a generalização do modelo. O conjunto de treino é utilizado para ajustar o modelo, isto é, estimar os coeficientes ou parâmetros com base nas observações disponíveis. Já o conjunto de teste é reservado para verificar o desempenho do modelo em dados independentes, que não participaram do processo de ajuste, permitindo uma estimativa mais realista de sua precisão preditiva.

### 3.5. validação cruzada

A validação cruzada (ou *cross-validation*) é uma técnica mais robusta que aprimora esse processo ao dividir os dados em múltiplas partições — chamadas de *folds*. Em um procedimento comum, como a validação cruzada k-fold, o conjunto de dados é dividido em  $k$

subconjuntos aproximadamente iguais. O modelo é ajustado  $k$  vezes, utilizando em cada rodada  $k - 1$  subconjuntos para o treino e o restante para o teste. Ao final, as métricas de desempenho são médias entre as  $k$  repetições, proporcionando uma avaliação mais estável e menos dependente de uma única divisão dos dados. Esse procedimento é especialmente importante quando o tamanho da amostra é limitado, pois permite utilizar todos os dados tanto para o treino quanto para o teste em momentos distintos, garantindo uma estimativa mais confiável da capacidade de generalização do modelo e minimizando o risco de sobreajuste (overfitting).

#### 4. Resultados e discussões

Inicialmente, de acordo com a análise descritiva das variáveis, podemos analisar alguns pontos interessantes. Os resultados da tabela 2 apresenta resultados semelhantes ao estudo de (SILVA; ALMEIDA, 2019), e indicam que a maioria dos estudantes tem entre 15 e 22 anos, com uma média de 16 anos e uma variação significativa nas idades. Em relação ao número de reprovações nas classes anteriores, a grande maioria dos alunos não teve reprovações, refletindo um desempenho acadêmico positivo, com uma média de 0,34 reprovações. Por outro lado, as faltas escolares mostram uma média de 5,71 ausências, mas a alta variância indica que, enquanto muitos alunos aparecem regularmente, alguns apresentam um número elevado de faltas, indicando a necessidade de atenção a esses casos extremos.

**Tabela 2:** Medidas De Posição.

| Medidas    | Idade | Número de reprovação em classes anteriores | Números de faltas escolares |
|------------|-------|--------------------------------------------|-----------------------------|
| Min        | 15,00 | 0                                          | 0                           |
| Máx        | 22,00 | 3,00                                       | 75,00                       |
| 1° Quartil | 16,00 | 0                                          | 0                           |
| 3° Quartil | 18,00 | 0                                          | 8,00                        |
| Média      | 16,00 | 0,34                                       | 5,71                        |
| Mediana    | 17,00 | 0                                          | 4,00                        |
| Variância  | 1,628 | 0,55                                       | 64,05                       |

Fonte: Elaborado pelos Autores, (2025).

A Tabela 3 apresenta a caracterização socioeconômica e demográfica dos estudantes de Matemática. São descritas variáveis referentes à escola de origem, sexo, local de residência, composição familiar, estado civil dos pais, escolaridade dos responsáveis, além das ocupações da mãe e do pai e a identificação do tutor principal do aluno. Esses dados permitem compreender o contexto social e familiar no qual os discentes estão inseridos, fornecendo subsídios importantes para a análise das relações entre fatores socioeconômicos, condições familiares e o desempenho escolar.

**Tabela 3:** Descritiva socioeconômica e demográfica dos alunos de matemática.

| Variável |                 | N   | %      | IC 95% |       |
|----------|-----------------|-----|--------|--------|-------|
|          |                 |     |        | Inf    | Sup   |
| Escola   | Gabriel pereira | 349 | 88,35% | 46,95  | 47,87 |



|                                                                                                                             |                               |     |        |       |       |
|-----------------------------------------------------------------------------------------------------------------------------|-------------------------------|-----|--------|-------|-------|
|                                                                                                                             | Mousinho da silveira          | 46  | 11,65% | 52,12 | 53,05 |
| Sexo                                                                                                                        | Feminino                      | 208 | 52,66% | 47,61 | 57,66 |
|                                                                                                                             | Masculino                     | 187 | 47,34% | 42,34 | 52,40 |
| Endereço                                                                                                                    | Urbano                        | 307 | 77,72% | 73,23 | 81,66 |
|                                                                                                                             | Rural                         | 88  | 22,28% | 18,34 | 26,77 |
| Tamanho da família                                                                                                          | Maior que 3                   | 281 | 71,14% | 66,35 | 75,51 |
|                                                                                                                             | Menos ou igual a 3            | 114 | 28,86% | 24,49 | 33,65 |
| Estado civil dos pais                                                                                                       | Vivendo juntos                | 354 | 89,62% | 86,09 | 92,36 |
|                                                                                                                             | Separados                     | 41  | 10,28% | 7,64  | 13,92 |
| Escolaridade da mãe                                                                                                         | Nenhuma                       | 3   | 0,75%  | 0,02  | 2,40  |
|                                                                                                                             | 5ª a 9ª série                 | 103 | 26,07% | 21,87 | 30,75 |
|                                                                                                                             | Ensino fundamental (4ª série) | 59  | 14,94% | 11,64 | 18,95 |
|                                                                                                                             | Ensino médio                  | 99  | 25,06% | 20,92 | 29,70 |
|                                                                                                                             | Ensino superior               | 131 | 33,16% | 28,58 | 38,08 |
|                                                                                                                             | Nenhuma                       | 2   | 0,51%  | 0,01  | 2,02  |
| Escolaridade da pai                                                                                                         | 5ª a 9ª série                 | 115 | 29,11% | 24,73 | 33,90 |
|                                                                                                                             | Ensino fundamental (4ª série) | 82  | 20,76% | 16,93 | 25,16 |
|                                                                                                                             | Ensino médio                  | 100 | 25,32% | 21,16 | 29,96 |
|                                                                                                                             | Ensino superior               | 96  | 24,30% | 20,21 | 28,90 |
|                                                                                                                             | Em casa                       | 59  | 14,94% | 11,65 | 18,93 |
| Trabalho da mãe                                                                                                             | Outros                        | 141 | 35,70% | 31,01 | 40,66 |
|                                                                                                                             | Professor                     | 58  | 14,69% | 11,42 | 18,66 |
|                                                                                                                             | Saúde                         | 34  | 8,61%  | 6,12  | 11,93 |
|                                                                                                                             | Serviços                      | 103 | 26,07% | 21,87 | 30,75 |
|                                                                                                                             | Em casa                       | 20  | 5,06%  | 11,65 | 18,93 |
| Trabalho da pai                                                                                                             | Outros                        | 217 | 54,94% | 31,01 | 40,66 |
|                                                                                                                             | Professor                     | 29  | 7,34%  | 11,42 | 18,66 |
|                                                                                                                             | Saúde                         | 18  | 4,56%  | 6,12  | 11,93 |
|                                                                                                                             | Serviços                      | 111 | 28,10% | 21,87 | 30,75 |
|                                                                                                                             | Mãe                           | 273 | 69,11% | 64,26 | 73,59 |
| Tutor do aluno                                                                                                              | Pai                           | 90  | 22,78% | 18,80 | 27,30 |
|                                                                                                                             | Outros                        | 32  | 8,11%  | 5,69  | 11,36 |
| N= frequência absoluta; %= frequência relativa; Inf= intervalo de confiança inferior; Sup= intervalo de confiança superior. |                               |     |        |       |       |

Fonte: elaborado pelos autores, (2025).

A Tabela 4 apresenta a distribuição das variáveis relacionadas ao acesso e suporte dos estudantes de Matemática. São consideradas informações sobre o tempo de deslocamento entre casa e escola, suporte educacional extra, apoio familiar, participação em atividades extracurriculares, histórico de frequência em creche, acesso à internet no domicílio e tempo semanal dedicado aos estudos. Esses indicadores permitem compreender as condições de apoio educacional e estrutural que envolvem os alunos, fornecendo elementos relevantes para



analisar de que forma fatores externos ao ambiente escolar podem impactar no desempenho acadêmico e no processo de aprendizagem.

**Tabela 4:** Descritiva de Acesso e suporte dos alunos de matemática.

| Variável                                                                                                                    |                  |  | N   | %      | IC 95% |       |
|-----------------------------------------------------------------------------------------------------------------------------|------------------|--|-----|--------|--------|-------|
|                                                                                                                             |                  |  |     |        | Inf    | Sup   |
| Tempo de viagem de casa para a escola                                                                                       | Menos de 15 min  |  | 257 | 65,06% | 60,11  | 69,72 |
|                                                                                                                             | 15 a 30 min      |  | 107 | 27,09% | 22,82  | 31,80 |
|                                                                                                                             | 30 min a 1 hora  |  | 23  | 5,82%  | 3,80   | 8,73  |
|                                                                                                                             | Mais de 1 hora   |  | 8   | 2,03%  | 0,09   | 4,10  |
| Suporte educacional extra                                                                                                   | Não              |  | 344 | 87,09% | 83,28  | 90,15 |
|                                                                                                                             | Sim              |  | 51  | 12,91% | 9,85   | 16,71 |
| Apoio educacional familiar                                                                                                  | Não              |  | 153 | 38,73% | 33,93  | 43,77 |
|                                                                                                                             | Sim              |  | 242 | 61,27% | 56,24  | 66,06 |
| Atividades extracurriculares                                                                                                | Não              |  | 194 | 49,11% | 44,01  | 54,15 |
|                                                                                                                             | Sim              |  | 201 | 50,89% | 45,83  | 55,90 |
| Frequentou creche                                                                                                           | Não              |  | 81  | 20,51% | 16,70  | 24,89 |
|                                                                                                                             | Sim              |  | 314 | 79,49% | 75,10  | 83,29 |
| Tem acesso a internet em casa?                                                                                              | Não              |  | 66  | 16,71% | 13,23  | 20,83 |
|                                                                                                                             | Sim              |  | 329 | 83,29% | 79,16  | 86,76 |
| Tempo de estudo semanal                                                                                                     | Menos de 2 horas |  | 105 | 26,58% | 22,35  | 31,28 |
|                                                                                                                             | De 2 a 5 horas   |  | 198 | 50,13% | 45,09  | 55,16 |
|                                                                                                                             | De 5 a 10 horas  |  | 65  | 16,45% | 13,01  | 20,56 |
|                                                                                                                             | Mais de 10 horas |  | 27  | 6,84%  | 4,63   | 9,91  |
| N= frequência absoluta; %= frequência relativa; Inf= intervalo de confiança inferior; Sup= intervalo de confiança superior. |                  |  |     |        |        |       |

Fonte: elaborado pelos autores (2025).

A Tabela 5 apresenta a descrição das variáveis relacionadas às relações sociais, familiares e ao estado de saúde dos estudantes de Matemática. São incluídos aspectos como a qualidade das relações familiares, a percepção sobre o tempo livre após a escola, a frequência de interação social com amigos, a autoavaliação do estado de saúde e a situação afetiva dos alunos. Esses indicadores possibilitam compreender o contexto relacional e de bem-estar em que os discentes estão inseridos, fatores que podem influenciar tanto no rendimento escolar quanto na saúde mental e emocional.

**Tabela 5:** Descritiva de relações e saúde dos alunos de matemática.

| Variável                          |            |  | N   | %      | IC 95% |       |
|-----------------------------------|------------|--|-----|--------|--------|-------|
|                                   |            |  |     |        | Inf    | Sup   |
| Qualidade das relações familiares | Muito Ruim |  | 8   | 2,02%  | 0,09   | 4,10  |
|                                   | Ruim       |  | 18  | 4,56%  | 2,81   | 7,24  |
|                                   | Regular    |  | 68  | 17,22% | 13,69  | 21,38 |
|                                   | Boa        |  | 195 | 49,37% | 44,35  | 54,41 |

|                            |       |               |     |        |       |       |
|----------------------------|-------|---------------|-----|--------|-------|-------|
| Tempo depois da escola     | livre | Excelente     | 106 | 26,83% | 22,58 | 31,55 |
|                            |       | Insuficiente  | 19  | 4,81%  | 3,01  | 7,54  |
|                            |       | Limitado      | 64  | 16,20% | 13,78 | 20,29 |
|                            |       | Moderado      | 115 | 29,11% | 34,92 | 44,78 |
|                            |       | Adequado      | 157 | 39,75% | 24,73 | 33,91 |
| Sair com os amigos         |       | Excelente     | 40  | 10,13% | 7,41  | 13,63 |
|                            |       | Muito pouco   | 23  | 5,82%  | 3,81  | 8,73  |
|                            |       | Pouco         | 103 | 26,07% | 21,87 | 30,75 |
|                            |       | Moderadamente | 130 | 32,91% | 28,34 | 37,82 |
|                            |       | Bastante      | 86  | 21,77% | 17,87 | 26,24 |
| Estado de saúde            |       | Sai muito     | 53  | 13,43% | 10,29 | 17,27 |
|                            |       | Muito ruim    | 47  | 11,90% | 8,95  | 15,60 |
|                            |       | Ruim          | 45  | 11,39% | 8,51  | 15,04 |
|                            |       | Regular       | 97  | 23,04% | 19,04 | 27,57 |
|                            |       | Bom           | 66  | 16,71% | 13,24 | 20,84 |
| Está em um relacionamento? |       | Excelente     | 146 | 36,96% | 32,23 | 41,95 |
|                            |       | Não           | 236 | 66,58% | 75,11 | 84,03 |
|                            |       | Sim           | 132 | 33,42% | 28,82 | 38,33 |

N= frequência absoluta; %= frequência relativa; Inf= intervalo de confiança inferior; Sup= intervalo de confiança superior.

Fonte: elaborado pelos autores (2025).

A Tabela 6 apresenta a distribuição do consumo de álcool entre os estudantes de Matemática, considerando tanto os finais de semana quanto os dias de aula. Os dados permitem identificar a frequência e a intensidade do consumo

**Tabela 6:** Descritiva para consumo de álcool dos alunos de matemática.

| Variável                             |               | N   | %      | IC 95% |       |
|--------------------------------------|---------------|-----|--------|--------|-------|
|                                      |               |     |        | Inf    | Sup   |
| Consumo de álcool no final de semana | Não consome   | 151 | 38,23% | 33,45  | 43,24 |
|                                      | Pouco         | 85  | 21,52% | 17,63  | 25,97 |
|                                      | Moderadamente | 80  | 20,25% | 16,47  | 24,63 |
|                                      | Muito         | 28  | 7,09%  | 4,85   | 10,20 |
|                                      | Bastante      | 51  | 12,91% | 9,84   | 16,71 |
| Consumo de álcool em dia de aula     | Não           | 276 | 69,87% | 65,04  | 74,31 |
|                                      | Sim           | 119 | 30,13% | 25,69  | 34,95 |

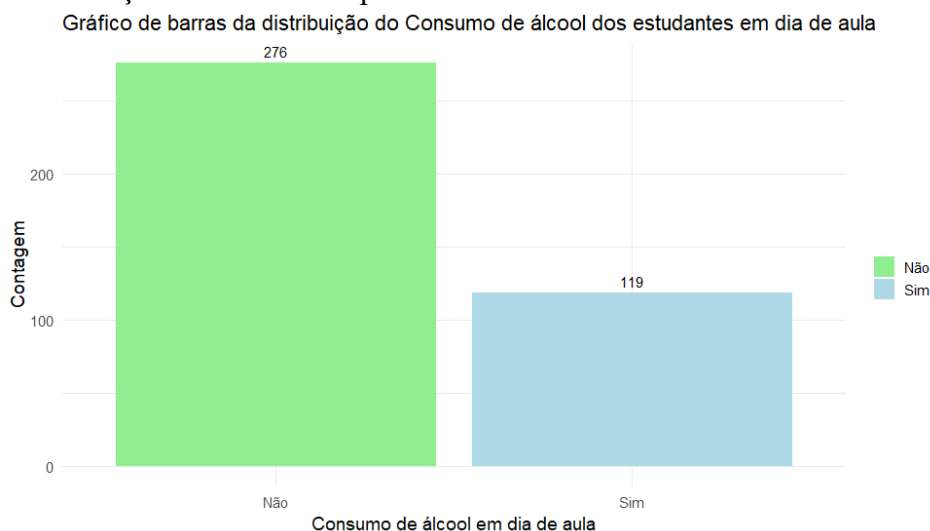
N= frequência absoluta; %= frequência relativa; Inf= intervalo de confiança inferior; Sup= intervalo de confiança superior.

Fonte: elaborado pelos autores (2025).

A Figura 2 tem-se o gráfico de barras mostra a distribuição da variável reposta que é o consumo de álcool pelos estudantes em dias de aula. Observa-se que a maioria dos alunos declarou não

consumir álcool durante esses dias, representando 276 casos. Por outro lado, um número menor, mas ainda relevante, de estudantes afirmou consumir álcool em dia de aula, totalizando 119 casos. Esse resultado evidencia que, embora o comportamento predominante seja a abstenção do consumo de álcool no contexto escolar, existe uma parcela considerável de alunos que mantém esse hábito mesmo em dias de atividades acadêmicas, o que pode trazer implicações tanto para o desempenho escolar quanto para a saúde.

**Figura 2:** Distribuição da variável resposta.



Fonte: Elaborado pelos autores (2025).

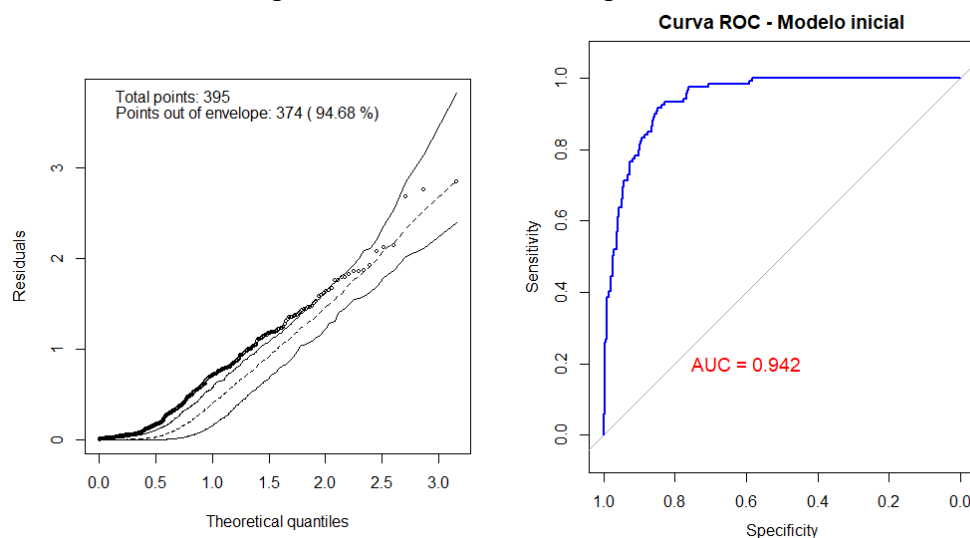
A Tabela 7, tem-se as métricas para o modelo inicial, que apresentou um AIC de 338,59, valor que, apesar de não poder ser interpretado isoladamente, permite comparações com outros modelos: quanto menor o AIC, melhor o equilíbrio entre qualidade do ajuste e complexidade do modelo. Nesse caso, o resultado sugere que o modelo conseguiu representar bem os dados sem excesso de parâmetros. Além disso, o Pseudo  $R^2$  foi de 0,54, indicando que aproximadamente 54% da variabilidade da variável dependente é explicada pelas variáveis independentes incluídas, o que representa um poder explicativo considerável em estudos sociais e comportamentais em estudo semelhante ao de (RODRIGUES, 2020).

**Tabela 7:** Métricas para avaliação do modelo inicial.

| Métricas     | Valores |
|--------------|---------|
| AIC          | 338,59  |
| Pseudo $R^2$ | 0,54    |

Fonte: Elaborado pelos autores (2025).

A Figura 2 mostra a curva ROC do modelo inicial apresentou uma área sob a curva (AUC) de 0,94, valor bastante elevado, o que demonstra excelente capacidade discriminatória, ou seja, o modelo é eficaz em diferenciar estudantes que consomem álcool em dias letivos daqueles que não consomem. Entretanto, ao avaliar o gráfico de envelope dos resíduos, observa-se que 94,68% dos pontos ficaram fora dos limites esperados, evidenciando fragilidades no ajuste do modelo aos dados e possíveis violações de pressupostos estatísticos.

**Figura 3:** Gráfico de envelope simulado e Curva Roc para o modelo inicial

Fonte: Elaborado pelos Autores, (2025).

**Figura 4:** Matriz de confusão do modelo inicial.

|      | Predito |    |
|------|---------|----|
| Real | 249     | 24 |
|      | 27      | 95 |

Fonte: Elaborado pelos Autores, (2025).

Na Tabela 8 temos a análise da matriz de confusão, o modelo inicial alcançou precisão de 0,94, o que significa que, entre os casos classificados como consumo de álcool, 90% foram de fato positivos. Já a acurácia foi de 0,87, indicando que, no conjunto geral de classificações, 87% foram corretas. Esses valores reforçam que, apesar das limitações de ajuste indicadas pelo envelope, o modelo inicial apresentou desempenho preditivo elevado e consistente, mostrando-se adequado para identificar padrões relacionados ao consumo de álcool entre estudantes.

**Tabela 8:** Métricas da matriz de confusão do modelo inicial.

| Métricas | Valores |
|----------|---------|
| Precisão | 0,94    |
| Acurácia | 0,87    |

Fonte: Elaborado pelos Autores, (2025).

A Tabela 9 mostra as métricas do modelo baseado no VIF apresentou um AIC de 468,28, valor mais elevado em comparação ao modelo inicial, indicando um desempenho inferior em termos de ajuste global e parcimônia. Esse resultado sugere que, embora tenha sido útil na redução da multicolinearidade, o modelo perdeu capacidade explicativa. Esse ponto é reforçado pelo Pseudo  $R^2$  de 0,11, revelando que apenas 41% da variabilidade do consumo de álcool é explicada pelas variáveis selecionadas, o que representa um baixo poder explicativo para o fenômeno em estudo.

**Tabela 9:** Métricas para avaliação do modelo com base nas variáveis que se demonstram aceitáveis pelo VIF

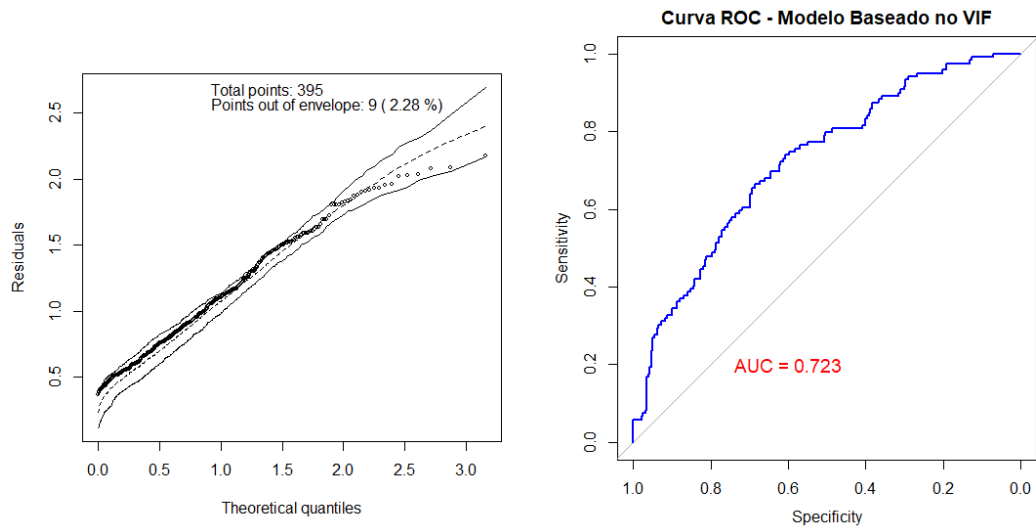
| Métricas | Valores |
|----------|---------|
| AIC      | 468,28  |

Pseudo  $R^2$ 0,41

Fonte: Elaborado pelos Autores, (2025).

A curva ROC do modelo apresentou uma AUC de 0,723, valor moderado, o que indica uma capacidade discriminatória aceitável, porém bem inferior à observada no modelo inicial. Apesar disso, o envelope dos resíduos mostrou apenas 2,28% de pontos fora dos limites, evidenciando um ajuste estatístico bastante consistente, diferentemente do modelo inicial, que apresentou quase todos os pontos fora.

Figura 5: Gráfico de envelope simulado e Curva Roc para o modelo VIF.



Fonte: Elaborado pelos Autores, (2025).

Figura 6: Matriz de confusão para o modelo com base nas variáveis que se demonstram aceitáveis pelo VIF.

| Real | Predito |    |
|------|---------|----|
|      | 262     | 88 |
|      | 14      | 31 |

Fonte: Elaborado pelos Autores, (2025).

Na Tabela 10, podemos observar as métricas, o AUC foi de 0,723, representando uma capacidade discriminatória aceitável, mas distante do desempenho do modelo inicial (0,94). Já na matriz de confusão, o modelo apresentou precisão de 0,87 e acurácia de 0,72, ou seja, embora classifique corretamente uma boa parte dos casos, quase 30% das previsões ainda foram incorretas. Assim, o modelo VIF se mostra parcimonioso e estável, mas com desempenho preditivo inferior.

Tabela 10: Métricas da matriz de confusão do modelo VIF.

| Métricas | Valores |
|----------|---------|
| Precisão | 0,87    |
| Acurácia | 0,72    |

Fonte: Elaborado pelos Autores, (2025).

A Tabela 11 tem-se a métrica do modelo final, ajustado pelo método stepwise, apresentou um AIC de 449,90, valor intermediário entre o modelo inicial e o modelo VIF. Isso indica que,

embora tenha conseguido simplificar a seleção de variáveis, o modelo não alcançou a mesma parcimônia estatística do VIF nem o mesmo poder explicativo do modelo inicial. Nesse sentido, o Pseudo  $R^2$  de 0,78 sugere que 60% da variabilidade do consumo de álcool em dias letivos foi explicada pelas variáveis selecionadas, um desempenho moderado, porém superior ao do modelo VIF (0,41).

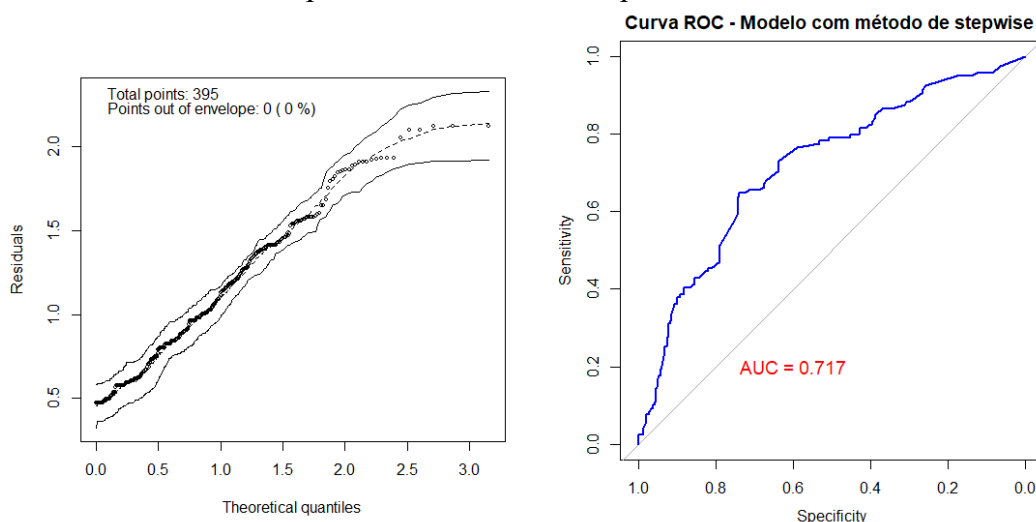
**Tabela 11:** Métricas para avaliação do modelo final.

| Métricas     | Valores |
|--------------|---------|
| AIC          | 449,90  |
| Pseudo $R^2$ | 0,78    |

Fonte: Elaborado pelos autores (2025).

A análise da curva ROC revelou uma AUC de 0,717, muito próxima à do modelo VIF, indicando uma capacidade discriminatória aceitável, inferior ao modelo inicial (0,94). No entanto, o modelo stepwise obteve um destaque importante no ajuste do envelope dos resíduos, onde 0% dos pontos ficaram fora dos limites esperados, demonstrando excelente adequação estatística e consistência do modelo (HOSMER; LEMESHOW, 2013).

**Figura 7:** Gráfico de envelope simulado e curva roc para o modelo final.



Fonte: Elaborado pelos Autores, (2025).

**Figura 8:** Matriz de confusão para o modelo do método final.

|      |         |    |
|------|---------|----|
|      | Predito |    |
| Real | 259     | 95 |
|      | 17      | 24 |

Fonte: Elaborado pelos Autores, (2025).

Na Tabela 12 temos a matriz de confusão, os resultados mostraram precisão de 0,92, evidenciando que, entre as classificações de consumo de álcool, 92% foram corretas (HAN; KAMBER; PEI, 2012). Já a acurácia foi de 0,72, um desempenho global modesto, semelhante ao obtido pelo modelo VIF. Em síntese, o modelo final se destacou pelo ajuste perfeito no envelope e pela seleção de variáveis relevantes, mas apresentou limitações quanto ao poder

preditivo e discriminatório. Isso reforça que sua principal contribuição está na clareza interpretativa e na consistência estatística, ainda que em detrimento da performance preditiva.

**Tabela 12:** Métricas da matriz de confusão do modelo final.

| Métricas | Valores |
|----------|---------|
| Precisão | 0,92    |
| Acurácia | 0,72    |

Fonte: Elaborado pelos Autores, (2025).

Na comparação entre os três modelos, observa-se na Tabela 13, que cada um apresentou vantagens e limitações distintas. O modelo inicial foi o que obteve o melhor desempenho preditivo, com maior AUC (0,94), alta precisão (0,90) e acurácia (0,87). Entretanto, apresentou limitações de ajuste, já que 94,68% dos pontos ficaram fora do envelope, sugerindo que o modelo não se adequou bem aos pressupostos estatísticos.

O modelo VIF, embora tenha mostrado desempenho inferior em termos de explicação (Pseudo  $R^2$  de 0,11) e poder discriminatório (AUC de 0,72), apresentou ajuste bastante consistente, com apenas 2,28% dos pontos fora do envelope. Sua precisão (0,87) e acurácia (0,72) indicam uma classificação razoável, mas inferior ao modelo inicial.

Já o modelo stepwise apresentou um equilíbrio intermediário, com Pseudo  $R^2$  de 0,20 e precisão de 0,92. Apesar de sua AUC (0,72) e acurácia (0,72) não se destacarem, o grande diferencial foi o ajuste perfeito no envelope, com 0% dos pontos fora. Isso indica que, estatisticamente, foi o modelo mais consistente, ainda que com menor poder discriminatório em comparação ao modelo inicial. Assim, conclui-se que a escolha do modelo depende do objetivo do estudo: se a prioridade for capacidade preditiva, o modelo inicial se mostra mais adequado; se for ajuste e parcimônia, o modelo stepwise se destaca; enquanto o modelo VIF surge como uma alternativa intermediária, garantindo estabilidade, mas com menor desempenho preditivo.

**Tabela 13:** Comparação geral entre os modelos.

| Modelo   | AIC    | Pseudo $R^2$ | AUC  | Precisão | Acurácia | % Pontos fora do envelope |
|----------|--------|--------------|------|----------|----------|---------------------------|
| Inicial  | 338,59 | 0,54         | 0,94 | 0,90     | 0,87     | 94,68%                    |
| VIF      | 468,28 | 0,41         | 0,72 | 0,87     | 0,72     | 2,28%                     |
| Stepwise | 449,90 | 0,78         | 0,72 | 0,92     | 0,72     | 0%                        |

Fonte: Elaborado pelos autores, (2025).



## 5. Conclusão

Os resultados mostraram que, embora o modelo final pelo método de stepwise tenha destacado de forma clara os fatores associados ao consumo de álcool em dias letivos como o sexo masculino e o número de reprovações aumentando a probabilidade de consumo, e variáveis como endereço urbano e atividades extracurriculares reduzindo essa chance, ele não apresentou os melhores indicadores de desempenho global, apesar de ter o envelope mais ajustado, sem nenhum ponto fora da curva. O modelo baseado no critério VIF mostrou desempenho intermediário, com precisão de 0,87 e acurácia de 0,72, e um envelope satisfatório, com apenas 2,28% dos pontos fora. Já o modelo inicial, que incluiu todas as variáveis, obteve os maiores valores de AUC (0,94), precisão (0,94) e acurácia (0,87), porém apresentou uma limitação importante no diagnóstico de ajuste, já que 94,68% dos pontos ficaram fora do envelope.

Dessa forma, conclui-se que os modelos apresentam pontos fortes distintos: o modelo inicial demonstrou maior capacidade preditiva, mas baixo ajuste no envelope; o modelo VIF equilibrou desempenho e parcimônia; e o modelo stepwise, apesar de menor poder preditivo, foi o que apresentou o melhor ajuste no envelope. Assim, a escolha do modelo depende do objetivo, se a prioridade for predição, o modelo inicial se mostra mais adequado, se for ajuste e estabilidade, o modelo stepwise é mais consistente.

Além disso, os achados reforçam que fatores individuais como o sexo e reprovações e contextuais residência urbana e atividades extracurriculares estão associados ao consumo de álcool entre estudantes, apontando caminhos para políticas educacionais e de saúde que considerem tanto características pessoais quanto fatores ambientais e comportamentais no planejamento de estratégias preventivas.

Do ponto de vista estatístico, a comparação entre os modelos demonstra a importância do equilíbrio entre ajuste e capacidade preditiva. Embora o modelo inicial tenha apresentado maior AUC e precisão, seu envelope revelou alta dispersão, indicando um sobreajuste, no qual o modelo se adapta excessivamente aos dados amostrais, comprometendo sua generalização. Já o modelo baseado no critério VIF, ao eliminar colinearidades, apresentou um desempenho mais estável, ainda que ligeiramente inferior em termos de acurácia. O modelo stepwise, por sua vez, demonstrou o melhor ajuste residual, revelando um comportamento mais consistente entre as observações previstas e observadas. Essas evidências reforçam a necessidade de cautela na escolha de modelos em pesquisas aplicadas à saúde e educação. A seleção deve considerar não apenas a performance numérica, mas também a coerência teórica e o potencial de generalização dos resultados. Além disso, a identificação de variáveis significativas, especialmente aquelas de natureza contextual, como o tipo de residência e o envolvimento em atividades extracurriculares, sugere que intervenções preventivas devem ser planejadas de forma intersetorial, envolvendo tanto políticas educacionais quanto ações comunitárias.

## Acknowledgements

Os autores agradecem ao WANDA 2025 pelo apoio institucional e pelo espaço de divulgação científica que possibilitaram a realização deste trabalho.

## Conflict of Interest Declaration

Os autores declaram inexistência de conflitos de interesse. Todos os coautores revisaram e aprovaram a versão final do manuscrito, não havendo interesses financeiros a declarar.

## Referências

- Long, J. S. (1997). *Regression models for categorical and limited dependent variables*. Sage Publications.
- McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press.
- Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1), 29–36. <https://doi.org/10.1148/radiology.143.1.7063747>
- Sugiyama, M., Nakajima, S., Kashima, H., von Büna, P., & Kawanabe, M. (2012). Direct importance estimation with model selection and its application to covariate shift adaptation. *Information and Inference: A Journal of the IMA*, 1(1), 29–55. <https://doi.org/10.1093/imaiai/iar003>
- Silva, J. R., Oliveira, M. P., & Lima, A. B. (2022). Prevalência e determinantes do consumo de álcool por adolescentes escolares. *Jornal de Pediatria*, 98(3), 210–218. <https://doi.org/10.1016/j.jped.2021.08.006>
- Arruda, P. S. M., Silva, A. N., Rinaldi, A. E. M., et al. (2022). Individual and contextual characteristics associated with alcohol use among Brazilian adolescents. *International Journal of Public Health*, 67, Article 1604397. <https://doi.org/10.3389/ijph.2022.1604397>
- Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). Wiley.
- Senna, S. R. C. M., et al. (2021). Fatores associados ao consumo de álcool em adolescentes: Estudo em escolas públicas. *Revista de Saúde Pública*, 55, 1–12. <https://doi.org/10.11606/s1518-8787.2021055003070>
- Rodrigues, C. G. (2020). *Análise de modelos de regressão logística em dados sociais: Uma aplicação ao estudo da desigualdade educacional* (Dissertação de mestrado, Universidade do Minho). RepositóriUM. <https://repositorium.sdum.uminho.pt/handle/1822/70801>
- Silva, J. P., & Almeida, M. (2019). *Análise descritiva de desempenho escolar: Estatísticas aplicadas à educação básica*. Editora Ciência & Educação.
- Hosmer, D. W., & Lemeshow, S. (2013). *Applied logistic regression* (3rd ed.). Wiley.
- Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.
- Cardoso, F. S., et al. (2023). Racial disparities in health care of adults hospitalized with acute respiratory syndrome/COVID-19 in Brazil: Direct effect of skin color on mortality analyzed through logistic regression. *Revista Brasileira de Epidemiologia*. <https://doi.org/10.1590/1980-5497202300XX>
- Góes, E. F., Ramos, D. O., & Ferreira, A. J. F. (2020). Desigualdades raciais em saúde e a pandemia da Covid-19. *SciELO Preprints*. <https://doi.org/10.1590/SciELOPreprints.1239>
- Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (5th ed.). John Wiley & Sons.

Malta, D. C., Prates, E. J. S., Ferreira, A. C. M., Freitas, P. C., Oliveira, P. P. V., Gomes, C. S., & Neto, E. L. G. R. (2021). Consumo e exposição a bebidas alcoólicas entre adolescentes brasileiros: Evidências das Pesquisas Nacionais de Saúde do Escolar de 2015 e 2019. *Revista Brasileira de Epidemiologia*, 24(Suppl. 1), e210002. <https://doi.org/10.1590/1980-549720210002.supl.1>

Souza, L. O., Targino, I. S., Andrade, J. C., Silva, D. G., & Voci, S. M. (2020). Marcadores de consumo alimentar de adolescentes. *Revista Brasileira de Epidemiologia*, 23, e200012. <https://doi.org/10.1590/1980-549720200012>

Silva, R. A., Malta, D. C., Cardoso, L. O., Andrade, S. S. C. A., & Gomes, C. S., et al. (2019). Consumo de bebidas alcoólicas por adolescentes brasileiros: Análise da Pesquisa Nacional de Saúde do Escolar (PeNSE). *Revista Brasileira de Epidemiologia*, 22(Suppl. 2), e190004. <https://doi.org/10.1590/1980-549720190004.supl.2>

Rodrigues, M. P., Costa, L. L., & Moura, L. F. A. (2021). Fatores associados ao consumo de álcool entre adolescentes de escolas públicas rurais do Nordeste brasileiro. *Ciência & Saúde Coletiva*, 26(Suppl. 3), 4567–4578. <https://doi.org/10.1590/1413-812320212611.3.27072019>