UNIVERSIDADE FEDERAL DE OURO PRETO

CURSO DE BACHARELADO EM ESTATÍSTICA

SANDRO IVO BIANCHI DA MOTTA

IMPACTO DO TIPO DE ESCOLA NAS NOTAS DO ENEM: ANÁLISE ESTATÍSTICA VIA REGRESSÃO LOGÍSTICA

OURO PRETO

2025

SANDRO IVO BIANCHI DA MOTTA

IMPACTO DO TIPO DE ESCOLA NAS NOTAS DO ENEM: ANÁLISE ESTATÍSTICA VIA REGRESSÃO LOGÍSTICA

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto, como requisito parcial para a obtenção do título de Bacharel em Estatística.

Orientadora: Profa. Dra. Graziela Dutra Rocha Gouvea.

OURO PRETO

2025

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Sandro Ivo Bianchi da Motta

Impacto do tipo de escola nas notas do ENEM: análise estatística via regressão logística

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 31 de março de 2025

Membros da banca

Dra. Graziela Dutra Rocha Gouvêa - Orientadora (Universidade Federal de Ouro Preto)
Dr. Fernando Luiz Pereira de Oliveira (Universidade Federal de Ouro Preto)
Dr. Tiago Martins Pereira (Universidade Federal de Ouro Preto)

Professora Dra. Graziela Dutra Rocha Gouvêa, orientadora do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 31/03/2025

Documento assinado eletronicamente por **Graziela Dutra Rocha Gouvea**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 04/04/2025, às 14:04, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0888831** e o código CRC **B2809E08**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

A todos que, de alguma forma, contribuíram para esta jornada, com um agradecimento especial: A minha esposa Rafaella e meus filhos Heitor e Clara que foram a minha força e inspiração. Aos meus colegas de classe, pelo apoio constante e pelas parcerias construídas ao longo do caminho. E a todos os professores, por compartilharem generosamente seu conhecimento e orientações.

RESUMO

Este trabalho analisa a influência das notas do Exame Nacional do Ensino Médio (ENEM) no tipo de escola frequentada pelos alunos (pública ou privada) utilizando técnicas estatísticas, como análise descritiva e modelos de regressão logística. Foram utilizados os microdados do ENEM 2023 fornecidos pelo INEP, selecionando as notas de Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação. A análise descritiva revelou disparidades significativas no desempenho acadêmico entre os dois tipos de escola, com alunos de escolas privadas apresentando, em média, melhores resultados. O modelo de regressão logística demonstrou que as notas nas diferentes áreas avaliadas têm impacto significativo na probabilidade de um aluno pertencer a uma escola pública ou privada, com exceção da variável Linguagens e Códigos. A validação do modelo indicou uma boa discriminação entre as categorias, utilizando técnicas estatísticas. No entanto, diagnósticos do modelo sugeriram a necessidade de melhorias.

Palavras-chave: ENEM; Regressão Logística; Métodos Estatísticos.

ABSTRACT

This study analyzes the influence of the scores from the National High School Exam (ENEM) on the type of school attended by students (public or private) using statistical techniques such as descriptive analysis and logistic regression models. The 2023 ENEM microdata provided by INEP was used, selecting scores from Natural Sciences, Humanities, Languages and Codes, Mathematics, and Writing. The descriptive analysis revealed significant disparities in academic performance between the two types of schools, with students from private schools generally achieving better results. The logistic regression model demonstrated that scores in the different evaluated areas significantly impact the probability of a student attending a public or private school, except for the Languages and Codes variable. Model validation indicated good discrimination between categories using statistical techniques. However, model diagnostics suggested the need for improvements.

Keywords: ENEM; Logistic Regression; Statistical Methods.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1. INTRODUÇÃO

O Exame Nacional do Ensino Médio (ENEM) foi criado em 1998 com o objetivo de avaliar o desempenho dos estudantes ao fim da educação básica. Inicialmente, sua função principal era ser uma ferramenta de diagnóstico para a melhoria do ensino no Brasil (INEP, 2020).

Com o passar dos anos, o ENEM ganhou novas funções, tornando-se um dos principais meios de acesso ao ensino superior, seja por meio de programas como o Portal Único de Acesso ao Ensino Superior (Sisu), o Programa Universidade Para Todos (Prouni) e o Fundo de Financiamento Estudantil (Fies), ou mesmo para ingresso direto em instituições públicas e privadas. Além disso, o exame passou a ser usado para certificação do ensino médio, ampliando ainda mais sua importância (INEP, 2020).

No entanto, o exame revela também diferenças no sistema educacional do Brasil, especialmente entre as instituições públicas e privadas. O objetivo deste estudo é investigar, usando técnicas estatísticas se o tipo de escola influencia as notas.

A análise baseia-se nos microdados do ENEM 2023 fornecidos pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 2023). A discussão sobre investimentos em educação não é nova, mas ainda é importante, especialmente quando se trata da necessidade de melhorar as condições para os estudantes das escolas públicas. Entender estas diferenças faz parte de esforço contínuo para construir uma sociedade mais justa.

Este trabalho busca contribuir para a discussão sobre a acessibilidade e equidade na educação no Brasil, analisando as notas do ENEM com base no tipo de escola.

2. MATERIAIS E MÉTODOS

Os microdados do Exame Nacional do Ensino Médio (ENEM) de 2023, fornecidos pelo INEP disponível em: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem, acesso em: 30/03/2025, contêm uma ampla variedade de informações sobre os participantes, incluindo dados socioeconômicos, escolares, respostas a questionários e as notas obtidas nas provas. As provas abrangem quatro áreas de conhecimento – Ciências Humanas, Ciências da Natureza, Linguagens e Códigos, Matemática – além de uma redação.

Para o propósito deste estudo, analisar se do tipo de escola afeta o desempenho dos alunos nas provas do ENEM, as seguintes variáveis foram selecionadas a partir dos microdados:

TP_ESCOLA: Indica o tipo de escola frequentada pelo participante durante o ensino médio, 0 para “Privada” e 1 para “Pública”.

NU_NOTA_CN: Nota na prova de Ciências da Natureza.

NU_NOTA_CH: Nota na prova de Ciências Humanas.

NU_NOTA_LC: Nota na prova de Linguagens e Códigos.

NU_NOTA_MT: Nota na prova de Matemática.

NU_NOTA_REDACAO: Nota na prova de Redação.

Estas variáveis foram escolhidas porque representam diretamente o desempenho acadêmico dos estudantes nas áreas avaliadas pelo ENEM, permitindo uma análise da relação entre o tipo de escola frequentada e o desempenho nas provas.

A primeira fase da análise estatística envolve uma exploração descritiva dos dados obtidos do Exame Nacional do Ensino Médio (ENEM) de 2023. Esta etapa é crucial para obter uma compreensão inicial das características dos dados, incluindo a distribuição das notas dos alunos e a proporção de participantes de escolas públicas e privadas. As atividades específicas incluem:

Sumarização dos Dados: Cálculo de medidas de tendência central e de dispersão para as notas em cada área do ENEM (Ciências Humanas, Ciências da Natureza, Linguagens e Códigos, Matemática e Redação) estatísticas de ordem foram utilizadas devido a assimetria dos dados.

Visualização dos Dados: Criação de gráficos, incluindo histogramas e boxplots, para visualizar a distribuição das notas por área de conhecimento e por tipo de escola.

Por meio destas técnicas, é possível identificar padrões, assimetrias na distribuição das notas e possíveis outliers.

Comparação Preliminar: Análise comparativa das notas entre alunos de escolas públicas e privadas, fornecendo uma visão inicial sobre as diferenças de desempenho que podem existir entre estes dois grupos.

Após a análise descritiva, a técnica de regressão logística é aplicada. O objetivo desta etapa é modelar a probabilidade de um aluno pertencer a uma escola pública ou privada com base em suas notas no ENEM. A regressão logística é particularmente adequada para este propósito, pois a variável dependente (tipo de escola) é binária. As etapas incluem:

Preparação dos Dados para Modelagem: Seleção das variáveis de interesse (notas nas áreas do ENEM) e tratamento de dados faltantes ou outliers, se necessário. A variável dependente (tipo de escola) foi codificada como uma variável binária.

Construção do Modelo de Regressão Logística: Ajuste do modelo utilizando a variável tipo de escola como variável dependente e as notas do ENEM como variáveis independentes. A técnica de regressão logística permitirá estimar os *odds ratios*, que indicam a força da associação entre o desempenho nas áreas do ENEM e a probabilidade de estar em uma escola pública versus privada.

Diagnóstico e Validação do Modelo: Avaliação do ajuste do modelo por meio de estatísticas de diagnóstico. Como será dada atenção à multicolinearidade entre as variáveis independentes, utilizando o Fator de Inflação da Variância (VIF) para detectar possíveis problemas.

Interpretação dos Resultados: Análise e interpretação dos coeficientes do modelo, traduzindo-os em termos da relação entre as notas do ENEM e a probabilidade de um aluno estar em uma escola pública ou privada. Discussão sobre a significância estatística e a relevância prática dos resultados.

Finalmente, os resultados sintetizados da análise descritiva e da modelagem de regressão logística, destacando as principais descobertas e sua implicação para a compreensão da diferença das notas em cada tipo de escola no exame.

Modelo de Bernoulli (Logístico)

A famosa frase de George Box (1979), “Todos os modelos estão errados, alguns são úteis”, nos lembra uma verdade fundamental na modelagem estatística e

científica: todos os modelos são simplificações da realidade, construídos sobre suposições que inevitavelmente não conseguem abarcar toda a complexidade do mundo. Essa natureza imperfeita, contudo, não diminui seu valor. Pelo contrário, a utilidade prática de modelos que, apesar de suas limitações, permitem fazer previsões confiáveis e auxiliam na tomada de decisões.

Criar modelos científicos é como construir um quebra-cabeça gigante. A cada nova peça que encaixamos, a imagem fica mais completa. Mas a imagem nunca estará perfeita. O importante é que o quebra-cabeça seja útil e nos ajude a entender o mundo ao nosso redor. Por isso, é preciso sempre revisar e ajustar os modelos, buscando torná-los cada vez mais precisos e aplicáveis.

Neste trabalho, os resultados obtidos através do software R (R Core Team, 2024), e um modelo de regressão logística binomial foi ajustado usando a função glm (*Generalized Linear Models*) disponível no pacote base *stats*. Na regressão logística, a relação entre as variáveis independentes e a variável dependente é modelada usando a função logit, que é a transformação logarítmica da razão das probabilidades de sucesso e fracasso. Essa transformação permite que a variável dependente seja modelada como uma função linear das variáveis independentes, mantendo a previsão dentro do intervalo [0, 1].

O objetivo do uso deste modelo nesta análise, está em prever a probabilidade de um indivíduo pertencer a uma determinada categoria de escola (como pública, ou, privada) com base em suas pontuações nas provas, que servem como variáveis explicativas.

O funcionamento do modelo de regressão logística baseia-se na estimativa da probabilidade de ocorrência de um evento binário, ou seja, um evento com dois possíveis resultados (neste caso, ser aluno de escola pública ou privada). Diferentemente da regressão linear, que estima valores contínuos, a regressão logística busca modelar uma variável dependente categórica, produzindo como saída valores entre 0 e 1, que podem ser interpretados como probabilidades.

Para isso, o modelo utiliza uma função de ligação chamada *logit*, que transforma a probabilidade $\boldsymbol{p}$ do evento de interesse ocorrer em uma escala que vai de $-\infty\ a\ +\infty$. Isso é necessário porque a combinação linear das variáveis explicativas $\boldsymbol{\beta 0 + \beta 1 X1 + \beta 2 X2 + \cdots + \beta n Xn}$ pode assumir qualquer valor real, enquanto a probabilidade deve estar restrita ao intervalo de 0 a 1.

Matematicamente, essa relação é expressa pela seguinte equação:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n$$

Onde:

p é a probabilidade de o evento de interesse ocorrer.

$\frac{1-p}{p}$ é a razão de odds (chance) para a probabilidade

$\beta_1, \beta_2, \ldots, \beta_p$ são os coeficientes do modelo, estimados a partir dos dados. Cada β representa o efeito de uma variável independente X na log-odds do evento.

$X_1, X_2, \ldots, X_n$ são as variáveis independentes. A partir da função logit, podemos usar a função logística (ou sigmóide) para transformar a saída do modelo linear (os valores de log-odds) de volta a uma probabilidade (um valor entre 0 e 1), através da fórmula:

$$p = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n}}$$

Para desenvolver o modelo, um conjunto de variáveis explicativas que representam as pontuações nas áreas do ENEM (Ciências da Natureza, Ciências Humanas, Linguagens e Códigos, Matemática e Redação) foram selecionadas. foi utilizado critérios estatísticos como valor p para selecionar o modelo final, garantindo que incluíssemos variáveis relevantes e significativas para a previsão do tipo de escola.

A função glm foi utilizada para ajustar o modelo, especificando a distribuição dos erros como binomial e a função de ligação como logit, comumente usada em modelos de regressão logística e é a função inversa da função logística. Ela mapeia a variável de resposta (a probabilidade de pertencer a uma categoria de escola) para uma escala contínua, permitindo a modelagem da relação entre as variáveis explicativas (pontuações nas provas) e a variável de resposta (categoria da escola) (HOSMER; LEMESHOW; STURDIVANT, 2013).

Para a adequada interpretação dos coeficientes na regressão logística, é necessário compreender o conceito de chance, definida como a razão entre a probabilidade de ocorrência e a de não ocorrência de um evento. O modelo logístico estima a relação entre as variáveis independentes e a chance do evento, expressa na forma de logaritmo das chances. Assim, a interpretação dos coeficientes ocorre por meio da razão de chances (*Odds Ratio*), que indica a variação proporcional nas chances associada a uma unidade de incremento na variável explicativa.

Chance no contexto de Regressão logística:

$$Chance = \left(\frac{prob(sucesso)}{prob(fracasso)}\right) = \left(\frac{P(Y=1)}{P(Y=0)}\right) = \left(\frac{P(Y=1)}{1-P(Y=1)}\right) = e^{\theta}$$

Exemplo: se a probabilidade de o aluno ser da escola pública(sucesso) for 79% e de escola privada (fracasso) 21%, isto significa se selecionarmos um aluno aleatório teríamos 3.76 vezes mais chance de sucesso de ser de escola pública do que privada.

Na equação logística, a relação entre as variáveis independentes e a chance do evento de interesse é dada por:

$$Chance = \left(\frac{prob(pública = 1)}{prob(privada = 0)}\right) = e^{\beta_0 + \beta 1 X1 + \beta_2 X2 + \ldots + \beta_n Xn}$$

Para a interpretação dos coeficientes é necessário entender o conceito de razão de chances (Odds Ratio):

O odds ratio (OR) é a razão das chances entre dois grupos ou situações diferentes. Especificamente, para um aumento de uma unidade na variável x_1 , a razão de chances é dada por:

$$OddsRatio = \frac{chance\ com\ x_1 + 1}{chance\ com\ x_1} = e^{\beta_1}$$

Pode se pensar que:

Se β_1 for -0.0048, supondo que x_1 seja a variável NU_NOTA_CN, cada aumento unitário na pontuação de Ciências da Natureza resulta em uma diminuição nas log-odds do evento de interesse.

Um OR de exp(-0.0048) = 0.9952115, significa que a cada aumento unitário na pontuação de Ciências da Natureza, as chances do evento de interesse (ser de escola pública) diminuem em aproximadamente 0.48% em relação a categoria de referência (privada).

Suponha que um aluno tenha de pontuação 650 em NU_NOTA_CN, 500 em NU_NOTA_MT, 475 em NU_NOTA_CH, e 730 em NU_NOTA_REDACAO:

Probabilidade = $e^{\beta_0+\beta_{1.650}+\beta_{2.500}+\beta_3,475+\beta_{4.730}}$ = 0.72

3. RESULTADOS

3.1. ESTATÍSTICAS DESCRITIVAS

Dos dados temos, que no ano de 2023 o Enem contou com 3.933.955 inscritos, sendo que, 71.75% deles, participaram do exame no primeiro dia e 68.44% participaram do segundo dia. Por algum motivo 3.31% dos que participaram, não retornaram para a segunda etapa, ou, não obtiveram notas.

Do total de participantes, 1.401.159 (35.6%) responderam em que tipo de escola, pública ou privada, completaram o ensino médio.

Como se trata de uma informação importante no contexto da análise, foi realizada uma filtragem, removendo os 64.4% de indivíduos que não responderam a esta questão.

Por fim, para esta análise, foram considerados somente os que Responderam a Questão Tipo de Escola e que obtiveram notas de alguma competência nos 2 dias de Prova, sendo estes, 1.050.858 indivíduos, 21.06% da escola Privada e 78.94% de Escola Pública.

3.2. DISTRIBUIÇÃO DAS NOTAS POR TIPO DE ESCOLA

Tabela 1 - Relação Notas Privada x Pública Percentis 50%, 25%,75%.

Competência	**Privada, N=221.328**		**Pública, N=829.530**	
	Mediana	**25% 75%**	**Mediana**	**25% 75%**
C. da Natureza	553	[495; 603]	473	[426; 525]
C. Humanas	584	[533; 627]	511	[452; 561]
Linguagens, Códigos	567	[522; 606]	508	[459; 552]
Matemática	647	[549; 719]	493	[416; 586]
Redação	800	[660; 900]	600	[480; 740]

De acordo com a Tabela 1 alunos de escolas privadas apresentam notas mais altas em todas as competências. Entre os alunos das escolas públicas a diferença entre os percentis mostra uma maior dispersão das notas. A tabela com a distribuição dos percentis mostra uma diferença entre o desempenho dos alunos de escolas públicas e privadas, especialmente em áreas como Matemática e Redação.

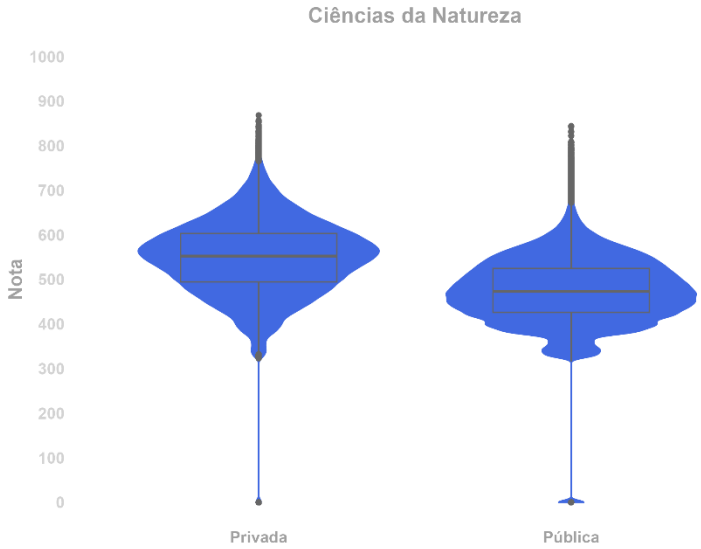

Figura 1 - Gráfico para a distribuição das notas em ciências da natureza.

Nota-se no gráfico da Figura 1 que a mediana das notas de *Ciências da Natureza* para escolas privadas é maior do que para escolas públicas. O desvio padrão é ligeiramente maior para as escolas privadas, ambas as escolas apresentam uma variabilidade significativa.

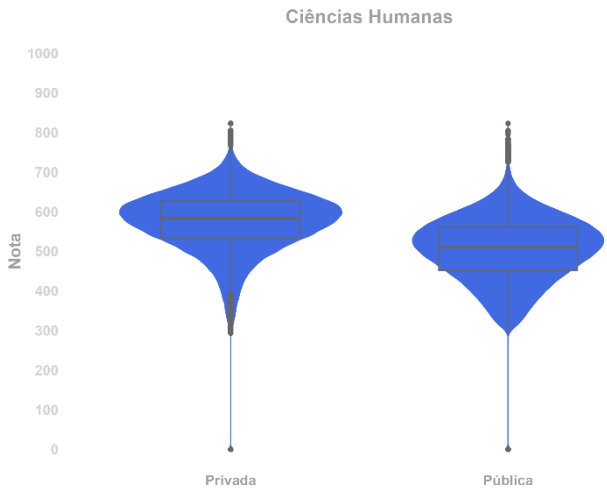

Figura 2 - Gráfico para a distribuição da nota em ciências das humanas.

Em *Ciências Humanas de acordo com a Figura 2* as notas possuem maior variação na escola Pública, ambas possuem uma grande concentração próximas a média e mediana, com um deslocamento a direita.

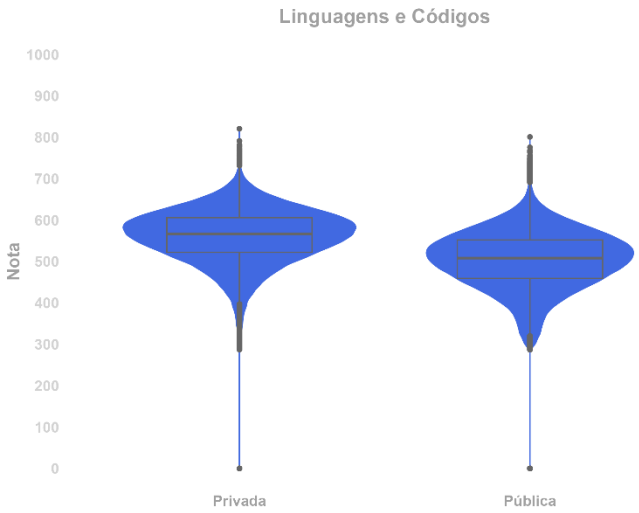

Figura 3 - Gráfico para a distribuição da nota em linguagens e códigos.

No gráfico da Figura 3, referente a Linguagens e Códigos, observa-se uma diferença menor no desempenho entre os dois tipos de escola, que pode não ser estatisticamente significativa para o modelo.

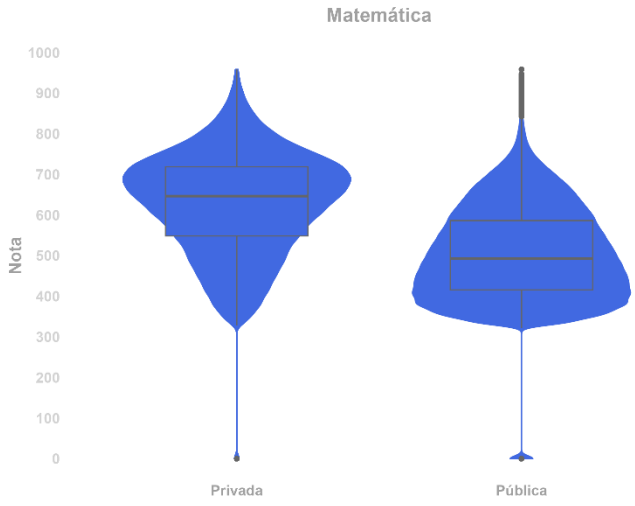

Figura 4 - Gráfico para a distribuição da nota em matemática.

A Figura 4 mostra claramente uma diferença na distribuição da pontuação, temos que mediana das notas de *Matemática* é mais alta para alunos de escolas privadas, o primeiro quartil fica próximo a mediana de estudantes de escola pública. A variabilidade das notas é maior para escolas privadas, indicando uma maior dispersão de notas. Ambos os tipos de escola apresentam outliers, mas escolas

privadas têm maior quantidade de notas altas. A distribuição das notas em escolas públicas parece ser ligeiramente inclinada para notas mais baixas, enquanto a distribuição para escolas privadas é mais simétrica.

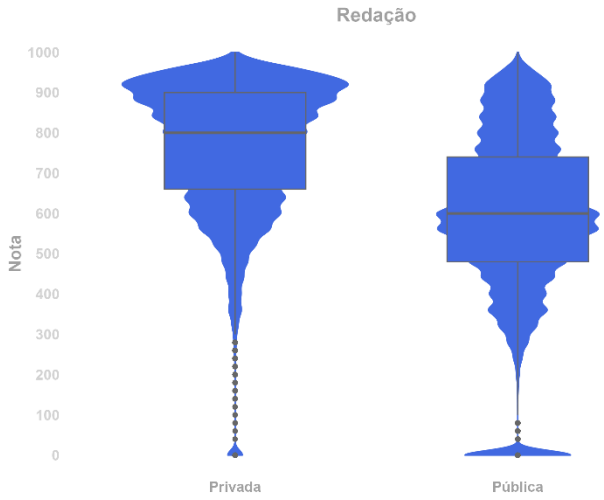

Figura 5 - Gráfico para a distribuição da nota em redação.

No gráfico da Figura 5 temos a nota Redação, onde há maior diferença na pontuação, o gráfico sugere que, em média, há uma boa diferença entre as notas de redação de alunos de escolas privadas e públicas.

3.3. CORRELAÇÃO ENTRE NOTAS

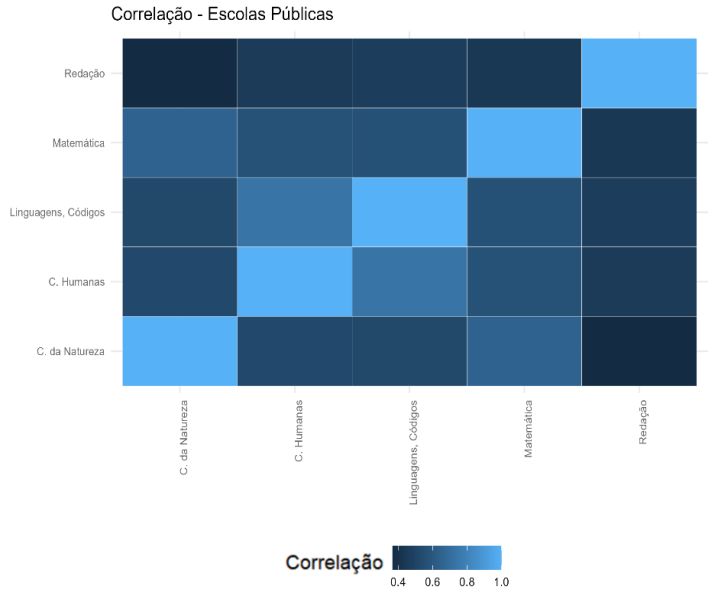

Figura 6 - Gráfico de correlação competência x escola pública.

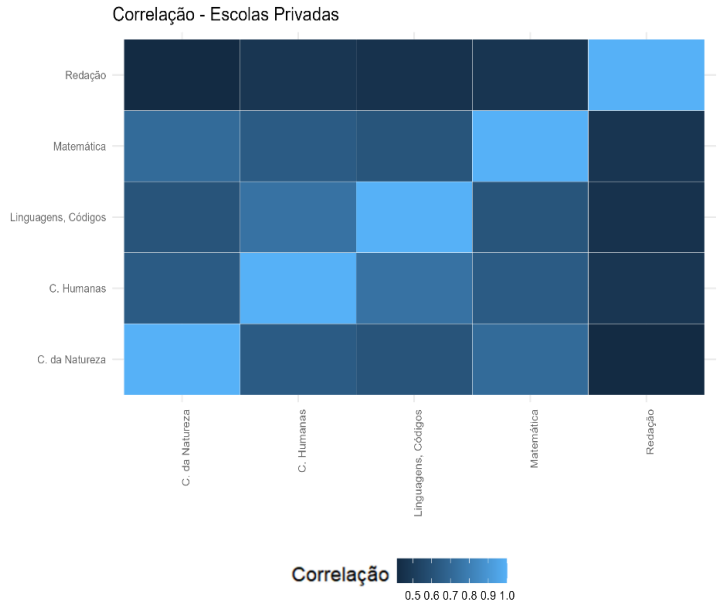

Figura 7 - Gráfico de correlação competência x escola privada.

No Gráfico de calor da Figura 6 e da Figura 7, as notas em cada competência se correlacionam de maneira parecida e positiva, em ambos de tipos de Escola. Ciências e Humanas e Linguagem, Códigos apresentam uma correlação positiva, indicando que os participantes possuem um desempenho parecido nestas competências, assim como Matemática e Ciências da Natureza.

Tabela 2 - Modelo ajustado e Resultados obtidos.

Variável	OR	95% IC	p-valor	VIF
C. da Natureza	0.9952	[0.9951, 0.9953]	<0.001	2.0
C. Humanas	0.9982	[0.9981, 0.9984]	<0.001	2.5
Linguagem, Códigos	1.0000	[0.9999, 1.0001]	0.900	2.4
Matemática	0.9958	[0.9957, 0.9959]	<0.001	2.0
Redação	0.9974	[0.9974, 0.9974]	<0.001	1.3

Os resultados do modelo ajustado com todas as variáveis na Tabela 2 fornecem coeficientes estimados para cada variável explicativa, juntamente com seus intervalos de confiança. Razões de chances (OR) igual a 1 indica que não há associação entre a variável independente e a variável dependente.

Embora todas as variáveis, exceto Linguagens e Códigos, sejam significativas, seus impactos no modelo são pequenos, como evidenciado pelos valores de OR próximos de 1. Por exemplo, para a variável que é a Nota em Nota C. da Natureza, um OR de 0.995 significa que, para cada aumento de uma unidade na pontuação dessa área, a chance de pertencer à categoria de escola Privada diminui em 0.4%, temos também o valor p indicando que todos os coeficientes são significativos com exceção de Linguagens, Códigos com p-valor maior que o nível de significância de 5%, e por fim temos o VIF menor que 10 em todas as variáveis indicando ausência de multicolinearidade, o que é positivo para a estabilidade do modelo.

A Nota de Linguagens e Códigos, não contribui significativamente para diferenciar os tipos de escola, sugerindo que pode ser irrelevante no contexto do modelo e, eventualmente, removida.

Tabela 3 - Um novo modelo é ajustado com a ausência da variável NU_NOTA_LC.

Variável	OR	95% IC	p-valor	VIF
C. da Natureza	0.9952	[0.9951, 0.9953]	<0.001	1.9
C. Humanas	0.9982	[0.9981, 0.9983]	<0.001	1.9
Matemática	0.9958	[0.9957, 0.9959]	<0.001	2.0
Redação	0.9974	[0.9974, 0.9974]	<0.001	1.3

Agora, com a exclusão de NU_NOTA_LC, temos um modelo com todas as variáveis estatisticamente significativas (p-valor < 0,05) na tabela 3.

3.4. DIAGNOSTICO DO MODELO

É importante ressaltar que o modelo de regressão logística assume alguns pressupostos, como a linearidade do logito, a ausência de multicolinearidade entre as variáveis explicativas, a independência dos erros e a adequação do modelo aos dados. Foi realizado uma análise cuidadosa para garantir que esses pressupostos fossem atendidos, garantindo assim a confiabilidade e a validade de nossos resultados.

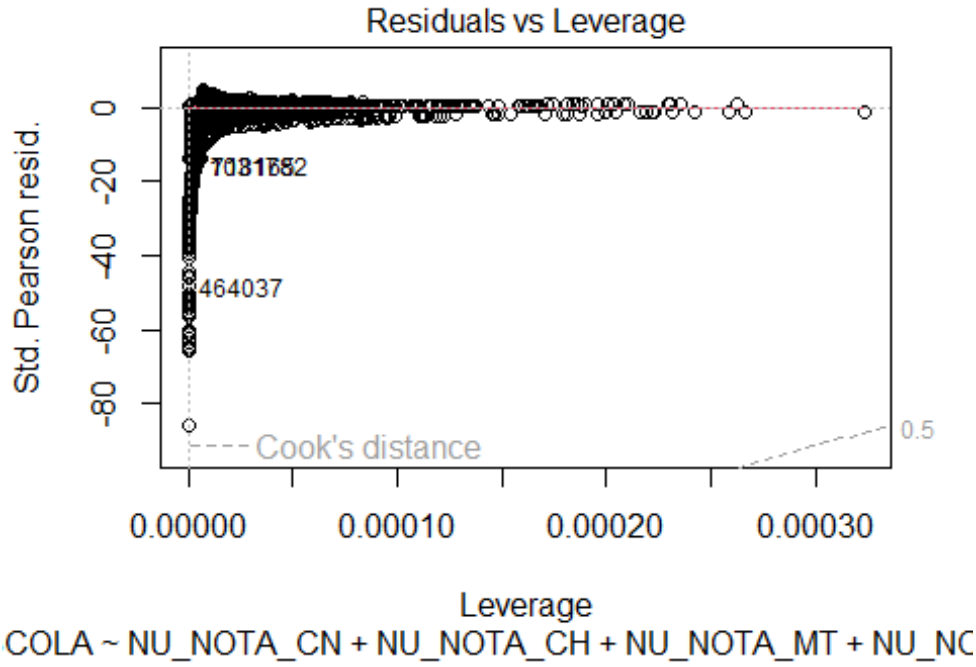

Figura 8 - Gráfico de resíduos x alavancagem

Os resíduos de Pearson versus os valores ajustados pelo modelo da Figura 8 são calculados subtraindo a probabilidade observada da probabilidade ajustada pelo modelo e dividindo pelo erro padrão. Eles são úteis para verificar se a suposição de homoscedasticidade (variância constante dos erros) foi violada. Podemos notar uma alta variação, resíduos padronizados extremamente negativos são evidentes no gráfico. Esses pontos estão muito abaixo da linha zero e indicam que o modelo subestima os valores observados para essas observações.

Espera-se que a dispersão dos pontos em torno da linha horizontal em zero deveria estar uniformemente distribuída, isso sugere que a variância dos erros é constante e que o modelo está adequado. porém é identificado padrão discernível nos pontos e a dispersão muda ao longo dos valores ajustados, isso pode indicar violações na homoscedasticidade e sugerir a necessidade de investigar mais a fundo o modelo.

3.5. RESIDUOS PADRONIZADOS:

A análise dos resíduos padronizados foi realizada para avaliar a adequação do modelo ajustado. Os valores extremos observados nos resíduos padronizados sugerem a presença de outliers ou pontos de alavancagem. Esses pontos podem influenciar desproporcionalmente o ajuste do modelo, resultando em estimativas de coeficientes que não são representativas para a maioria dos dados.

Tabela 4 - Análise de Deviance entre o Modelo Nulo e o Modelo Ajustado.

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1050857	1081905			
1050853	856483	4	225422.2	0

A Tabela 4 apresenta a comparação entre o modelo nulo e o modelo ajustado. Observa-se que a deviance residual reduziu de 1.081.905, no modelo nulo (sem as variáveis independentes), para 856.483, no modelo ajustado. A diferença de deviância foi de 225.422,2, com 4 graus de liberdade, e o valor de p associado é menor que 0, indicando que essa redução é estatisticamente significativa.

Isso significa que o modelo ajustado, que inclui as variáveis explicativas, apresenta um desempenho significativamente melhor do que o modelo nulo, que não considera nenhuma variável. Portanto, as variáveis incluídas contribuem de forma relevante para explicar a variável resposta.

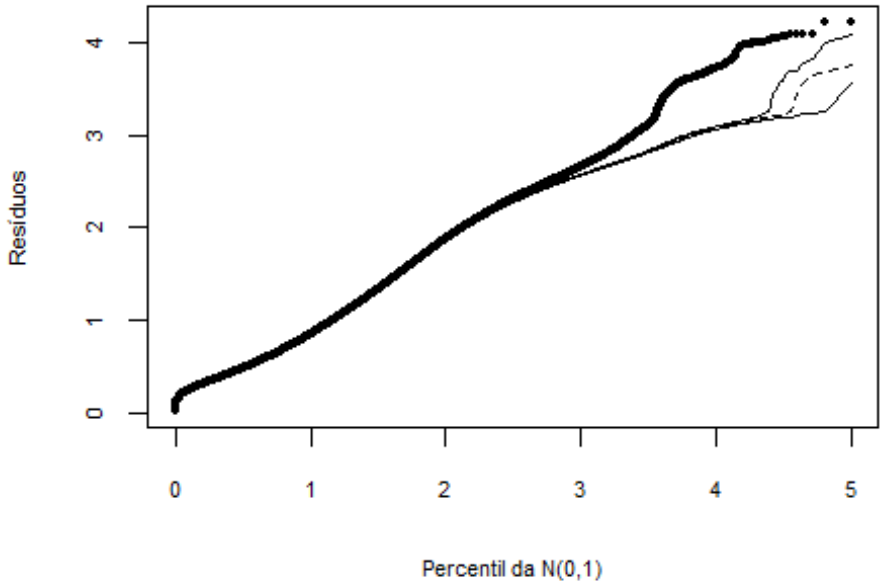

Figura 9 - Gráfico de envelope.

Os pontos na diagonal da Figura 9 representam os resíduos padronizados em relação a uma distribuição normal padrão (N (0,1)). Idealmente, esses pontos dentro do intervalo de confiança proposto pelo modelo, temos um total de 1.050.858 pontos e 399.464 (38,01%) ficaram fora do envelope.

Os envelopes mostram os limites dentro dos quais os resíduos padronizados devem cair se a normalidade for atendida. Se os pontos dos resíduos estiverem dentro desses envelopes, isso sugere que a normalidade é razoavelmente satisfeita, sendo assim como temos pontos fora do envelope, há evidências de que os erros não seguem distribuição normal.

Os envelopes servem para identificar desvios significativos da normalidade nos resíduos do modelo. Os pontos dos resíduos estiverem consistentemente fora dos envelopes, isso pode indicar que os resíduos não estão seguindo uma distribuição normal, o que pode afetar a validade das inferências do modelo.

Tabela 5 - Análise de Variância do Modelo Ajustado.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
MODELO NULO			1050857	1081905.2	
NU_NOTA_CN	1	158354.66	1050856	923550.5	< 0.00001
NU_NOTA_CH	1	20186.35	1050855	903364.2	< 0.00001
NU_NOTA_MT	1	26048.45	1050854	877315.7	< 0.00001
NU_NOTA_REDACAO	1	20832.74	1050853	856483.0	< 0.00001

A análise de variância do modelo na Tabela 5 sugere que todas as variáveis explicativas foram significativas para explicar a variável resposta.

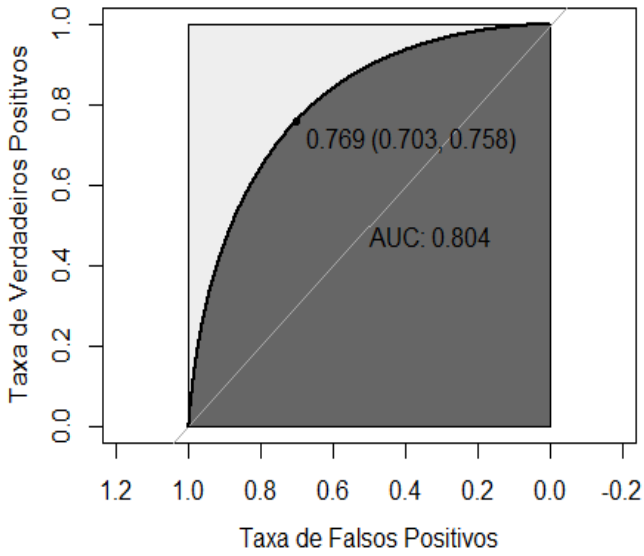

Figura 10 - Gráfico curva ROC.

O valor AUC (Area Under the Curve) da Figura 10 é de 0.804, ele resume o desempenho global do modelo e é utilizada para avaliar o ajuste do modelo de Bernoulli, de maneira semelhante ao coeficiente de determinação R^2.

Com uma AUC de 0.807, o modelo demonstra um bom desempenho na distinção entre as classes (escola pública vs. escola privada). A curva ROC sugere que o modelo mantém uma alta taxa de verdadeiros positivos enquanto minimiza os falsos positivos.

4. CONSIDERAÇÕES FINAIS

Através da aplicação de técnicas estatísticas, como a análise descritiva, a regressão logística e a construção de gráficos de distribuição, pudemos identificar padrões e diferenças significativas entre os dois grupos.

Alunos de escolas privadas apresentaram, em média, notas mais altas em todas as disciplinas do ENEM em comparação com alunos de escolas públicas. A análise dos *violin plots* e *boxplots* revelou que as notas dos alunos de escolas privadas não só são mais altas, mas também apresentam menor variabilidade, indicando um desempenho mais consistente.

A aplicação do modelo de regressão logística mostrou que as notas das diversas disciplinas são preditores significativos do tipo de escola. Os coeficientes da regressão indicaram que, para cada aumento nas notas das disciplinas, a chance de um aluno ser de uma escola privada aumentava, destacando a relação positiva entre melhores desempenhos e a frequência em escolas privadas.

No entanto, uma análise mais detalhada dos diagnósticos do modelo revelou algumas áreas de preocupação:

Os resíduos de Pearson mostraram alta variação entre os valores ajustados e os valores observados, indicando que o modelo pode não estar capturando perfeitamente todas as nuances dos dados. Alguns resíduos padronizados extremos foram identificados, sugerindo a presença de outliers ou pontos de alta alavancagem.

O gráfico de envelope revelou que vários pontos estavam fora das bandas de confiança esperadas, reforçando a necessidade de investigar mais a fundo os outliers e possíveis influências anômalas.

Apesar dos problemas com os resíduos, a AUC do modelo foi 0.804, indicando uma boa capacidade de discriminação global entre os tipos de escola. Isso sugere que, embora o modelo funcione bem para a maioria das observações, há áreas onde ele pode ser melhorado.

Dado os problemas identificados nos resíduos, pode ser benéfico considerar métodos de regressão robusta para minimizar a influência de outliers e melhorar a qualidade do ajuste do modelo. A implementação de uma regressão robusta poderia ajudar a obter coeficientes mais estáveis e um ajuste melhorado para os dados.

É essencial realizar uma investigação detalhada dos pontos de alta alavancagem e outliers para entender se eles representam erros de medição, dados

anômalos ou padrões que não estão sendo capturados pelo modelo atual. Ajustes no modelo ou tratamento desses pontos podem ser necessários para melhorar a precisão e a robustez dos resultados.

Os resultados obtidos na regressão logística reafirmam os resultados observados na análise descritiva. E embora o modelo de regressão logística tenha mostrado uma boa capacidade de discriminação entre alunos de escolas públicas e privadas, como evidenciado pela AUC, a análise diagnóstica sugere que há espaço para melhorias. A consideração de métodos robustos e a investigação de outliers são passos importantes para assegurar que os resultados do modelo sejam precisos e confiáveis.

5. REFERÊNCIAS

BOX, G. E. P. All models are wrong, but some are useful. *In:* **Robustness in Statistics.** Academic Press, 1979.

BURNHAM, K. P.; ANDERSON, D. R. **Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach.** Springer, 2002.

HOSMER, D. W.; LEMESHOW, S.; STURDIVANT, R. X. **Applied Logistic Regression**. 3. ed. New York: Wiley, 2013.

INEP. **Histórico.** Instituto Nacional de Estudos e Pesquisas Educacionais, 2020.

INEP. **Microdados do ENEM 2023.** Instituto Nacional de Estudos e Pesquisas Educacionais, 2023.

GIOLO, S. R. **Introdução à análise de dados categóricos com aplicações**. São Paulo: Edgard Blücher, 2017.

R CORE TEAM. *R: A Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing, 2024. Disponível em: https://www.R-project.org/, 2025.