Antônio Salvador Neto

REGRESSÃO LOGÍSTICA EM MICRODADOS DA EDUCAÇÃO

Vitória

2023

Antônio Salvador Neto

REGRESSÃO LOGÍSTICA EM MICRODADOS DA EDUCAÇÃO

Dissertação de mestrado apresentada ao PROFMAT como parte dos requisitos exigidos para a obtenção do título de Mestre em Matemática

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

MESTRADO PROFISSIONAL EM MATEMÁTICA EM REDE NACIONAL

PROFMAT

Orientador: Prof. Dr. Alancardek Pereira Araujo

Vitória

2023

Ficha catalográfica disponibilizada pelo Sistema Integrado de Bibliotecas - SIBI/UFES e elaborada pelo autor

---

Salvador Neto, Antônio, 1985-
S182r REGRESSÃO LOGÍSTICA EM MICRODADOS DA EDUCAÇÃO / Antônio Salvador Neto. - 2023.
73 f. : il.

Orientador: Alancardek Pereira Araujo.
Dissertação (Mestrado Profissional em Matemática em Rede Nacional) - Universidade Federal do Espírito Santo, Centro de Ciências Exatas.

1. Regressão Logística. 2. Desempenho acadêmico. 3. ENEM. 4. Análise descritiva. I. Pereira Araujo, Alancardek. II. Universidade Federal do Espírito Santo. Centro de Ciências Exatas. III. Título.

CDU: 51

UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO

Centro de Ciências Exatas

Programa de Pós-Graduação em Matemática em Rede Nacional – PROFMAT

“Regressão Logística em Microdados da Educação”

Antônio Salvador Neto

Defesa de Dissertação de Mestrado Profissional submetida ao Programa de Pós-Graduação em Matemática em Rede Nacional da Universidade Federal do Espírito Santo como requisito parcial para a obtenção do título de Mestre em Matemática.

Aprovado em 19/12/2023 por:

_____

Prof.(a) Dr.(a) Alancardek Pereira Araujo
Orientador(a) – UFES

_____

Prof.(a) Dr.(a) Fábio Júlio da Silva Valentim
Membro interno – UFES

_____

Prof. Dr.(a) Giselle Ribeiro de Azeredo Silva Strey
Membro Externo – SEDU-ES

Para Luana que me acompanhou e me deu forças em cada momento e Penha que não pôde estar aqui para ver o fim dessa jornada.

Agradecimentos

Agradeço à Sociedade Brasileira de Matemática e à Universidade Federal do Espírito Santo a pela oportunidade de realização do PROFMAT. Agradeço ainda a todos os professores e colegas que me inspiraram durante o período em que cursei do PROFMAT pela paixão que demonstram ter pela matemática. Agradeço à Juliette Zanetti que me ajudou demais com seu apoio e revisões.

"Saber muito não significa ser inteligente; a inteligência não é só informação, mas também julgamento, a maneira pela qual uma informação é coordenada e utilizada."
(Carl Sagan Cosmos (1980))

Resumo

O desempenho acadêmico é uma preocupação constante no campo da educação. Técnicas estatísticas de regressão aplicadas aos dados obtidos no exame nacional do ensino médio (ENEM) podem auxiliar na compreensão dos fatores que interferem o desempenho acadêmico de estudantes. O trabalho apresenta fundamentação teórica sobre regressão linear e logística bem como técnicas de discretização, manipulação de dados e medidas de qualidade de modelo. Modelos de regressão logística foram ajustados com dados do ENEM dos anos de 2020, 2021 e 2022 de forma independente e para cada disciplina da avaliação. Os modelos se mostraram robustos o suficiente para prever desempenho de alunos baseando-se em dados socioeconômicos. A análise descritiva e dos coeficientes do modelo apontam para uma forte correlação negativa entre o número de pessoas que vivem na mesma residência do aluno e o seu desempenho. Além disso a categoria de renda da família do aluno, ocupação do pai e o tipo de escola que o aluno frequentou tem grande peso sobre o desempenho na avaliação. Outro fator importante foi a idade, idades mais elevadas tendem a a pertencer a categorias de desempenho mais elevadas.

Palavras-chave: Desempenho acadêmico, Educação, Regressão logística, ENEM, Análise descritiva.

Abstract

Academic performance is a constant concern in the field of education. Statistical regression techniques applied to data obtained from the National High School Exam (ENEM) can assist in understanding the factors that influence students' academic performance.

The work provides theoretical foundations on linear and logistic regression, as well as discretization techniques, data manipulation, and model quality measures. Logistic regression models were fitted with ENEM data from the years 2020, 2021, and 2022 independently and for each assessment subject. The models proved robust enough to predict students' performance based on socioeconomic data.

Descriptive analysis and model coefficient examination point to a strong negative correlation between the number of people living in the same residence as the student and their performance. Additionally, the family income category, father's occupation, and the type of school the student attended have a significant impact on performance in the assessment. Another important factor was age; higher ages tend to belong to higher performance categories.

Keywords: Academic performance, Education, Logistic regression, ENEM (National High School Exam), Descriptive analysis.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

O desempenho acadêmico dos estudantes tem sido uma preocupação constante no campo da educação. Uma maneira popular de avaliar o desempenho é através de exames nacionais como o Exame Nacional do Ensino Médio (ENEM) no Brasil. Diversos estudos têm analisado os resultados do ENEM para identificar padrões e preditores do desempenho dos estudantes (SILVEIRA; BARBOSA; SILVA, 2015).

A Teoria de Resposta ao Item (TRI), cuja aplicação no ENEM foi analisada por Ferreira (2018), Tadeu e Costa (2017), Espiritu e Filho (2020), entre outros é uma abordagem sofisticada para a medição do desempenho em testes padronizados. Diferentemente dos métodos tradicionais que consideram apenas o número de respostas corretas para determinar a pontuação de um indivíduo, a TRI leva em conta tanto a dificuldade de cada questão quanto o padrão de respostas do indivíduo ao longo do teste.

Especificamente, a TRI permite uma estimativa mais precisa da habilidade de um estudante, pois reconhece que uma questão difícil corretamente respondida indica uma habilidade maior do que uma questão fácil corretamente respondida. Além disso, a TRI ajuda a criar testes equitativos, permitindo a comparação de pontuações de diferentes versões de um teste, cada uma com um conjunto único de itens. Essas características fazem da TRI um método ideal para exames de grande escala, como o ENEM, que precisa avaliar de maneira justa e precisa a competência de milhões de estudantes a cada ano. (ARAUJO; ANDRADE; BORTOLOTTI, 2009), (PASQUALI; PRIMI, 2003).

As técnicas de regressão são ferramentas poderosas para interpretar e entender os dados em diversas áreas de estudo, incluindo a educação. Elas permitem a exploração da relação entre variáveis independentes e uma variável dependente, fornecendo uma visão profunda das dinâmicas subjacentes a essas relações. (RODRIGUES; MEDEIROS; GOMES, 2013).

Júnia e Mariz (2021), por exemplo, aplica o modelo de regressão logística aos dados do ENEM para decifrar os fatores que afetam o desempenho dos alunos. A regressão logística é particularmente útil em contextos onde a variável dependente é binária (por exemplo, aprovar ou reprovar, passar ou falhar), como comumente ocorre na educação (HOSSMER; LEMESHOW, 2000). Ao utilizar este método, é possível determinar quais fatores estão estatisticamente associados a esses resultados binários, e assim, fornecer insights valiosos para melhorar as estratégias de ensino e aprendizado(FARIA et al., 2021).

De maneira semelhante, De, Moreira e Das (2021) discute o uso da regressão linear e logística como uma ferramenta de análise de dados que podem ser, inclusive, aplicados em uma proposta pedagógica para incentivar os alunos a lidar com esse tipo de ferramenta

estatística cada vez mais necessária para lidar com o grande volume de dados que nos cercam (ALPAYDIN, 2020).

Além disso, Américo e Lacruz (2017) utiliza a regressão linear múltipla para analisar o desempenho escolar nas escolas do Espírito Santo, revelando a importância do ambiente escolar para o desempenho dos estudantes. Sua pesquisa reforça o argumento de que as técnicas de regressão são fundamentais para a compreensão dos fatores que influenciam o desempenho escolar, permitindo a identificação de oportunidades de intervenção para melhorar a qualidade da educação. Viggiano e Mattos (2009) identificaram diferenças regionais no desempenho dos estudantes, uma questão importante na educação brasileira, dadas as variações econômicas e culturais entre as regiões do país.

A relevância do Exame Nacional do Ensino Médio (ENEM) na sociedade brasileira, bem como as discrepâncias entre diferentes redes de ensino, foram analisadas em profundidade por Bessa (2016). Esta visão está alinhada com o trabalho de Hossmer e Lemeshow (2000), que apontam a importância da análise estatística, como a regressão logística, na avaliação do desempenho educacional.

Bessa (2016) argumenta que o ENEM desempenha um papel crucial como um mecanismo de mobilidade social, pois os resultados obtidos neste exame podem proporcionar acesso a oportunidades de ensino superior e emprego. Isso está em conformidade com os estudos de Viggiano e Mattos (2009), que também enfatizaram a influência do desempenho no ENEM nas oportunidades de educação superior. Além disso, Bessa (2016) ressalta o valor do ENEM como um instrumento para a avaliação e classificação das escolas. Conforme indicado por Faria et al. (2021), as classificações baseadas em desempenho em exames como o ENEM têm implicações significativas para a percepção do público em relação à qualidade das instituições de ensino, entretanto existe uma preocupação com as disparidades entre as redes de ensino. O desempenho escolar de estudantes está fortemente correlacionado com o contexto socioeconômico da família. Esta relação foi observada em estudos conduzidos por várias organizações e pesquisadores.

Pires (2015) fez uma análise específica das relações entre a renda familiar e a escolaridade dos pais, com os resultados do ENEM no estado de São Paulo. O estudo concluiu que existe uma influência positiva e significativa da renda familiar e da escolaridade dos pais nos resultados do ENEM. Essa conclusão está alinhada com a análise realizada no trabalho de Américo e Lacruz (2017), que explorara a relação entre o ambiente escolar e o desempenho escolar, utilizando a Prova Brasil como medida de desempenho. Eles concluíram que, embora o ambiente escolar explique parte do desempenho, as características socioeconômicas da família do estudante, como renda familiar, têm um impacto substancial.

De maneira similar, Assunção, Araújo e Almeida (2019) exploraram a influência do histórico familiar (incluindo renda familiar e a escolaridade dos pais) no acesso à educação técnica e vocacional. Eles descobriram que tanto a renda familiar quanto a escolaridade

do pai têm um efeito positivo no acesso a essas instituições, embora a escolaridade da mãe não tenha mostrado significância estatística .

Além disso, em sua análise abrangente das políticas sociais brasileiras, o IPEA (2019) destacou a importância do contexto socioeconômico, incluindo a renda familiar, para o desempenho acadêmico dos estudantes IPEA (2019).

2 Fundamentação teórica

Análise estatística é um ramo da matemática que lida com a coleta, interpretação, análise e apresentação de dados. Ela ajuda a entender as tendências e padrões nos dados e a tomar decisões informadas com base nessas informações. É uma ferramenta essencial em diversos campos, incluindo negócios, economia, ciências sociais, biologia, medicina, engenharia e muito mais (MORETTIN; BUSSAB, 2017).

A análise estatística pode ser dividida em duas categorias principais: estatística descritiva e estatística inferencial. A *estatística descritiva* lida com a descrição e apresentação dos dados. Ela se concentra em resumir e apresentar os dados de maneira fácil de entender, frequentemente por meio de gráficos e tabelas. Medidas como média, mediana, moda e variância são frequentemente usadas na estatística descritiva. A *estatística inferencial* usa os dados para fazer previsões ou inferências sobre uma população maior com base em uma amostra desses dados. Ela usa técnicas como testes de hipóteses, intervalos de confiança e regressão. Até mesmo a forma com que o ENEM infere nota a um aluno através da TRI (teoria de resposta ao ítem) se utiliza do método estatístico inferencial (KARINO; SOUSA, 2012), (KLEIN,) (HAIR; BLACK; SANT'ANNA, 2000) (HAIR, 2009). Uma ferramenta muito útil da estatística inferencial são as regressões. Considere o seguinte exemplo. Ao examinar a correlação entre o número de horas de estudo e o desempenho em avaliações, uma tendência geral é observada. Normalmente, um aumento nas horas de estudo resulta em um aumento proporcional no desempenho na avaliação. É intuitivo perceber que essas duas variáveis estão interligadas de alguma forma. A condução de análises quantitativas frequentemente revela padrões que correlacionam uma variável com outra (ECCLE, 2002). É possível construir um modelo matemático que descreva essa relação e medir a eficácia desse modelo na representação das variáveis correlacionadas (MEYER, 2017). Esse modelo pode ajudar a compreender melhor como uma grandeza (desempenho na prova) está relacionada com outra (tempo de estudo).

Uma vez modelado matematicamente essa relação entre grandezas é possível fazer previsões, como antecipar o desempenho futuro de um estudante com base no número de horas de estudo(DE; MOREIRA; DAS, 2021). Pode-se expandir a aplicação da regressão para modelar fenômenos mais complexos, como a relação entre treinamento em habilidades específicas e o desempenho no trabalho, ou entre a quantidade de prática de um esporte e o desempenho nesse esporte. Existem vários tipos de regressão, cada um adequado a diferentes tipos de relações entre as variáveis (MONTGOMERY, 2012). Outras técnicas de regressão, como a regressão polinomial, exponencial e logística, são derivadas da regressão linear. Além disso, esses modelos podem ser adaptados para incluir mais de uma variável. É importante lembrar que a análise de regressão se preocupa apenas com a dependência

estatística entre as variáveis, não com relações determinísticas ou funcionais. A correlação entre duas variáveis não implica necessariamente que uma causa a outra (HAIR, 2009). As variáveis dependentes e independentes são conceitos fundamentais na análise de dados e estatística. Chama-se variável dependente aquela que está sendo observada e estudada (desempenho na avaliação), enquanto a variável independente é aquela que é manipulada ou controlada para afetar a variável dependente (horas de estudo). Na regressão, a relação entre essas variáveis é bidirecional. Isso significa que para o modelo, não importa se são as horas de estudo que influenciam o desempenho na avaliação ou se é o desempenho na avaliação que influencia as horas de estudo (PEREIRA, 2009).

2.0.1 Lei dos Grandes Números

A Lei dos Grandes Números (**LGN**) é um princípio fundamental na teoria da probabilidade e estatística que descreve o comportamento de médias amostrais à medida que o tamanho da amostra aumenta. Essa lei estabelece que, sob certas condições, a média de um grande número de observações independentes de uma variável aleatória converge para o valor esperado da variável. Em outras palavras, à medida que coletamos mais dados, a média amostral se aproxima da média populacional (MISES, 1957).

Existem duas formas principais da Lei dos Grandes Números: a *Forma Fraca* e a *Forma Forte*.

A Forma Fraca afirma que, para uma sequência de variáveis aleatórias independentes e identicamente distribuídas (i.i.d.), a média amostral converge em probabilidade para a média populacional à medida que o tamanho da amostra aumenta. Matematicamente, se $X_1, X_2, \ldots, X_n$ são i.i.d. com média μ e variância finita σ^2, então para qualquer $\varepsilon > 0$:

$$\lim_{n\to\infty} P\left(|\bar{X}_n - \mu| \geq \varepsilon\right) = 0$$

onde $\bar{X}_n$ é a média amostral de n observações.

A Forma Forte vai além, afirmando que não apenas a média amostral converge em probabilidade, mas converge quase certamente para a média populacional. Para a mesma sequência de variáveis aleatórias i.i.d., isso é expresso como:

$$P\left(\lim_{n\to\infty} \bar{X}_n = \mu\right) = 1$$

Esta forma é mais restritiva, mas fornece uma garantia mais forte sobre a convergência da média amostral.

A Lei dos Grandes Números tem aplicações fundamentais em diversas áreas, desde finanças até ciências naturais, onde a coleta e análise de grandes conjuntos de dados são

comuns. Essa lei é essencial para a inferência estatística e fornece uma base sólida para a compreensão do comportamento médio de variáveis aleatórias em experimentos repetidos.

2.1 Regressão linear

O objetivo da regressão linear é o desenvolvimento de um modelo estatístico que possa ser usado para prever valores de uma variável dependente y em função de uma variável independente x (FIGUEIRA, 2006) (CHEIN, 2019).

Para ilustrar a os fundamentos desenvolvidos nesta seção considere o seguinte exemplo. Em uma sala de aula com 10 alunos foi registrado o número de horas que cada aluno dedica aos estudos, seu desempenho na avaliação e o número de ocorrências de medidas disciplinares. A tabela a seguir mostra os dados coletados.

Tabela 1 – Dados fictícios para exemplo de regressão linear simples

Aluno	y **nota na avaliação**	x_1 **horas de estudo**	x_2 **número de advertências**
1	3.0	0.2	9
2	2.8	0.3	7
3	4.8	0.5	8
4	5.5	1.5	5
5	6.1	1.5	5
6	5.0	2.1	3
7	6.8	2.4	2
8	7.5	3.3	1
9	8.2	3.5	0
10	10.0	3.9	0

É possível intuir que a quantidade de horas estudadas está relacionada a um melhor desempenho na avaliação e uma maior quantidade de ocorrências está relacionada a um desempenho pior. A técnica de regressão linear pode fornecer uma regra geral de como as horas de estudo e número de ocorrências influenciam no desempenho do aluno. A regressão linear é realizada através da estimativa dos coeficientes da equação de regressão. Esses coeficientes representam as relações entre as variáveis independentes e a variável dependente. A equação de regressão linear simples tem a forma geral:

$$Y = \beta_0 + \beta_1 * X + \varepsilon \tag{2.1}$$

Onde:

- Y é a variável dependente que se deseja prever ou explicar, muitas vezes designada como variável *alvo*;

- X é a variável independente usada para fazer a previsão ou explicação;
- β_0 é o coeficiente linear, que representa o valor esperado de Y quando X é igual a zero;
- β_1 é o coeficiente angular, que representa a mudança esperada em Y quando X aumenta em uma unidade;
- ε é o termo de erro, que captura as variações não explicadas pelo modelo e é assumido como uma variável aleatória com média zero.

Para estimar os coeficientes β_0 e β_1, é preciso de um conjunto de dados observados que inclua valores para a variável dependente Y e a variável independente X. O método dos mínimos quadrados permite encontrar os valores dos coeficientes que minimizam a soma dos quadrados dos resíduos.

Assumimos que para um determinado conjunto de variáveis independentes $X_a \supset X$ que possuem mesmo valor numérico, ou seja, $x = a \ \forall \ x \in X_a$ onde $a \in \mathbb{R}$ a variável dependente y correspondente se distribui de forma normal em torno de uma medida central. O modelo linear deve representar esse valor central em que Y se distribui para cada valor de X. Para isso é comum se utilizar do método de mínimos quadrados que consiste em uma técnica usada para encontrar o melhor ajuste dos coeficientes para um conjunto de dados. Esse método consiste em minimizar a soma dos quadrados das diferenças (também chamadas de resíduos) entre os valores observados e os valores previstos pelo modelo.

Suponha que temos n pares de valores observados (x_i, y_i), onde $i = 1, 2, ;n$, e um modelo linear $y = \beta_0 + \beta_1 x$. O valor previsto para cada y_i é dado por $\hat{y}_i = \beta_0 + \beta_1 x_i$. O resíduo e_i para cada observação é a diferença entre o valor observado y_i e o valor previsto $\hat{y}_i$, ou seja, $e_i = y_i - \hat{y}_i$.

A soma dos quadrados dos resíduos (SSR) é dada por:

$$SSR = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 \tag{2.2}$$

O objetivo do método dos mínimos quadrados é encontrar os valores de β_0 e β_1 que minimizam a SSR.

Para isso deve-se minimizar a soma dos quadrados dos resíduos, SSR, que é uma função dos coeficientes β_0 e β_1. A minimização ocorre quando a derivada parcial de SSR em relação a β_0 e β_1 é igual a zero. Derivando SSR em relação a β_0:

$$\frac{\partial SSR}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = 2 \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i)) = 0 \tag{2.3}$$

Rearranjando o somatório:

$$\sum_{i=1}^{n} y_i - \beta_0 \sum_{i=1}^{n} -\beta_1 \sum_{i=1}^{n} x_i = 0 \tag{2.4}$$

$$n\beta_0 = \sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i \tag{2.5}$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i - \beta_1 \sum_{i=1}^{n} x_i}{n} \tag{2.6}$$

Seja $\bar{x}$ e $\bar{y}$ a média dos valores de x e y respectivamente dadas por:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i \tag{2.7}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{2.8}$$

Tem-se:

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \tag{2.9}$$

Tomando a derivada parcial de SSR em relação a β_1:

$$\frac{\partial SSR}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2 = -2 \sum_{i=1}^{n} x_i (y_i - (\beta_0 + \beta_1 x_i)) = 0 \tag{2.10}$$

Rearranjando os somatórios:

$$\sum_{i=1}^{n} (x_i \cdot y_i) - \beta_0 \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2.11}$$

Substituindo 2.9 em 2.11:

$$\sum_{i=1}^{n} (x_i \cdot y_i) - (\bar{y} - \hat{\beta}_1 \bar{x}) \sum_{i=1}^{n} x_i - \beta_1 \sum_{i=1}^{n} x_i^2 = 0 \tag{2.12}$$

Simplificando e isolando $\hat{\beta}_1$, temos:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} [(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{2.13}$$

As equações 2.9 e 2.13 fornecem os coeficientes de regressão que minimizam a soma dos quadrados dos resíduos. Em outras palavras, esses são os coeficientes que fornecem a

"melhor"linha reta que se ajusta aos dados observados, no sentido de minimizar a soma total das diferenças quadráticas entre os valores observados e previstos.

Aplicando o método dos mínimos quadrados nos dados do exemplo, considerando como variável independente o x_1 (horas de estudo) e como variável dependente y (nota na avaliação), é possível calcular os coeficientes que melhor se ajustam aos dados da seguinte forma.

Primeiro, calcula-se as médias das horas de estudo ($\bar{x_1}$) e das notas dos alunos ($\bar{y}$). De acordo com os dados fornecidos:

$$\bar{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1i} = \frac{1}{10}(0.2{+}0.3{+}0.5{+}1.5{+}1.5{+}2.5{+}2.4{+}3.3{+}3.5{+}3.9) = 1.92, horas \quad (2.14)$$

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{10}(3.0{+}2.8{+}4.8{+}5.5{+}6.1{+}5.0{+}6.8{+}7.5{+}8.2{+}10.0) = 5.97, pontos \quad (2.15)$$

Em seguida, calcula-se o estimador de mínimos quadrados para $hat\beta_1$ usando a Equação 2.13:

$$\hat{\beta}_1 = \frac{(3-5.97)(0.2-1.92)+(2.8-5.97)(0.3-1.92)+...+(10-5.97)(3.9-1.92)}{(0.2-1.92)^2+(0.3-1.92)^2+...+(3.9-1.92)^2} = 1.56 \quad (2.16)$$

Com $\hat{\beta}_1$ é possível determinar $\hat{\beta}_0$ através de 2.9:

$$\hat{\beta}_0 = 5.97 - 1.92 \cdot 1.56 = 2.97 \quad (2.17)$$

Ou seja a equação da reta que melhor descreve os dados é a função afim:

$$\hat{y} = 2.96 + 1.56x \quad (2.18)$$

A Figura 1 mostra a representação dessa equação em contraste com os dados.

Analogamente a regressão também pode ser calculada entre os pares x_2 e y fornecendo novos coeficientes de regressão que ajustam uma nova reta aos pares dessas variáveis. Note que a relação entre as variáveis número de advertências x_2 e nota na avaliação y é de ordem decrescente. Os gráficos a seguir ilustram o resultado da regressão linear.

Importante notar que a regressão linear não fornece o valor exato da variável dependente (Nota do aluno) em relação a variável independente (horas de estudo ou número de ocorrências). A técnica nos dá uma estimativa que pode ser interpretada

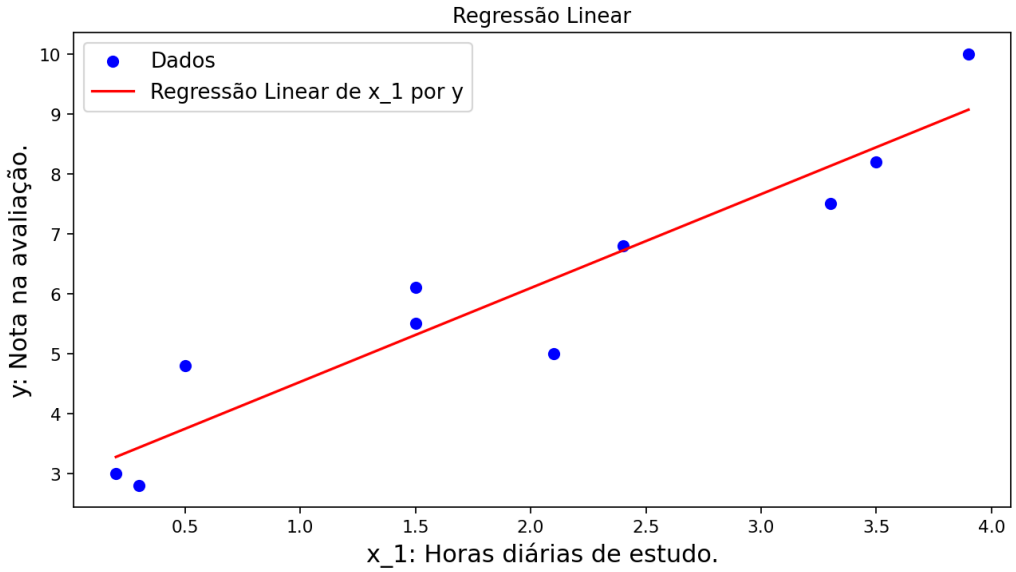


Figura 1 – Exemplo de regressão linear entre $x_1 e y$.

Fonte: Produção do próprio autor.

como uma relação geral entre as variáveis. A interpretação geométrica da regressão linear aplicada a pares de variáveis é uma linha reta que melhor se ajusta aos pontos dos dados observados. Essa técnica pode ser estendida de modo a avaliar simultaneamente duas variáveis independentes e uma dependente. Nesse caso a regressão determina um plano.

Considere uma função com k variáveis independentes como a seguir:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... \hat{\beta}_k x_k \tag{2.19}$$

De modo a operar a regressão linear simultaneamente entre mais de uma variável independente, x_1, x_2, ..., x_k em função de uma variável dependente y é possível escrever a Equação 2.19 na forma matricial:

$$Y = X\beta + \varepsilon \tag{2.20}$$

Onde:

- Y é uma matriz (n x 1) contendo os valores observados da variável dependente;
- X é uma matriz (n x k+1) contendo os valores observados das variáveis independentes, incluindo uma coluna de uns para representar o coeficiente linear β_0;
- β é uma matriz ($k + 1$ x 1) contendo os coeficientes a serem estimados;
- ε é uma matriz (n x 1) contendo os termos de erro.

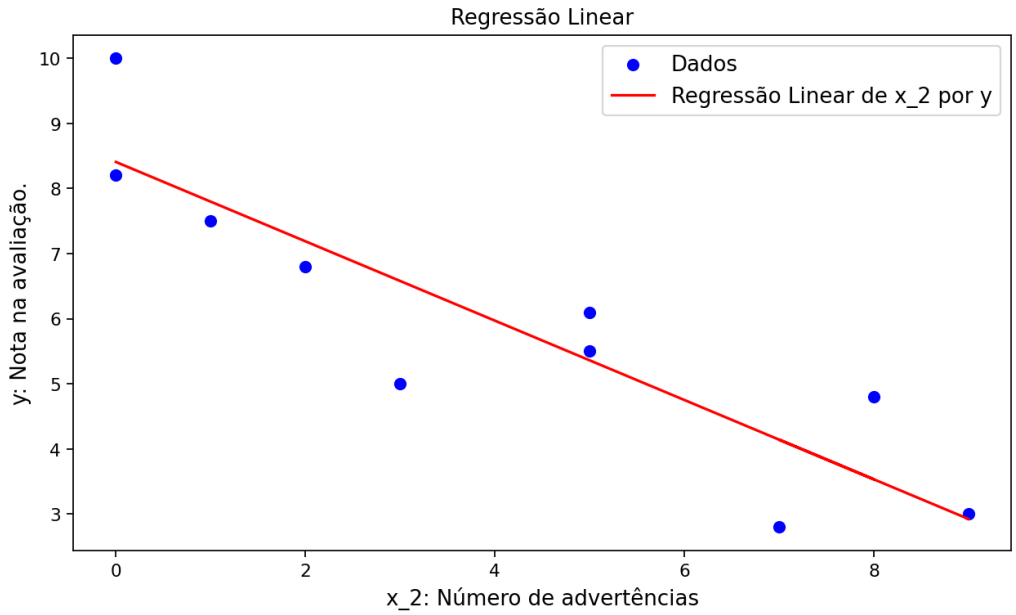


Figura 2 – Exemplo de regressão linear entre $x_2 e y$.

Fonte: Produção do próprio autor.

Nessa formulação o coeficiente $\hat{\beta}_0$ pode ser entendido como um fator que multiplica a variável independente x_0 constante e igual a um para qualquer valor de y. Esse coeficiente é comumente chamado de *bias* ou viés (HAIR, 2009). A estimativa dos coeficientes pode ser calculada através da equação:

$$\beta = (X^T X)^{-1} X^T Y \tag{2.21}$$

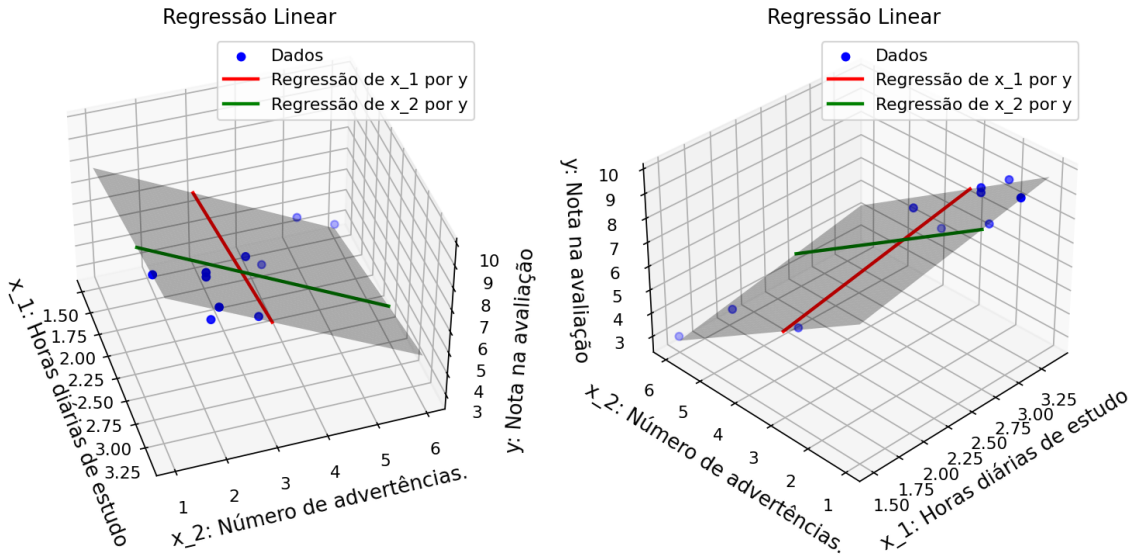
Onde:

X^T é a matriz transposta de X; $(X^T X)^{-1}$ é a matriz inversa de $X^T X$.

Ao estimar os coeficientes, obtém-se os valores de β_0, β_1, ..., β_k que melhor se ajustam os dados observados.

Os gráficos a seguir ilustram a interpretação geométrica para os dados do exemplo.

Note que o plano na interpretação em 3D da regressão linear com três variáveis pode ser reduzido as retas de regressão entre pares de variáveis. Analogamente a regressão linear pode ser estendida a conjuntos de n variáveis, no entanto a visualização geométrica se torna muito complexa devido a necessidade de mais de 3 dimensões espaciais para a representação.



Fonte: Produção do próprio autor.

Figura 3 – Exemplos de regressão linear entre três variáveis.

2.1.1 Regressão linear generalizada

A regressão linear generalizada (RLG) é um método estatístico que estende a regressão linear clássica para acomodar diferentes distribuições de resposta e relacionamentos não lineares entre as variáveis explicativas e a variável resposta. Ao contrário da regressão linear tradicional, que assume uma distribuição normal para a resposta, a RLG permite lidar com uma ampla gama de distribuições, como Poisson, binomial e gama (TURKMAN; SILVA, 2000) (NELDER; WEDDERBURN, 1972).

A RLG é composta por três componentes principais: o componente aleatório, o componente sistemático e a função de ligação. O *componente aleatório* descreve a distribuição probabilística da variável resposta e pode ser escolhido de acordo com a natureza dos dados. Por exemplo, se a resposta é uma contagem discreta, pode-se escolher uma distribuição de Poisson, binomial, de Bernoulli, entre outras.

O *componente sistemático* relaciona as variáveis explicativas à variável resposta por meio de um modelo linear. As variáveis explicativas são ponderadas por coeficientes, que são estimados a partir dos dados. O modelo linear pode incluir termos de interação e transformações não lineares das variáveis explicativas, permitindo capturar relacionamentos mais complexos.

A *função de ligação*, também conhecida como função de transformação, é um componente essencial na regressão linear generalizada (RLG) que conecta o componente aleatório (distribuição da resposta) ao componente sistemático (modelo linear das variáveis

explicativas). Essa função transforma a média da distribuição da resposta em uma combinação linear das variáveis explicativas.

A função de ligação é escolhida de acordo com a distribuição da resposta e tem a propriedade de ser uma função monótona diferenciável (AGRESTI, 2002). A monotonicidade significa que a função preserva a ordem dos valores. Por exemplo, se dois valores da média são maiores do que outros dois valores, a ordem desses valores será preservada após a aplicação da função de ligação. Isso é importante para garantir que a relação entre as variáveis explicativas e a resposta seja adequadamente capturada.

A diferenciabilidade refere-se à capacidade de calcular derivadas da função. Isso é fundamental para estimar os coeficientes do modelo e realizar inferências estatísticas. A diferenciabilidade permite calcular as derivadas parciais em relação aos coeficientes e determinar sua influência na resposta.

Uma função de ligação diferenciável também facilita a interpretação dos coeficientes do modelo. As derivadas parciais indicam como uma mudança nas variáveis explicativas se traduz em uma mudança na média da resposta. Além disso, a diferenciabilidade torna possível ajustar o modelo usando técnicas de otimização, como a maximização da verossimilhança.

$$g(\mu_i) = \sum_{j=1}^{n} \beta_j x_{ij} \tag{2.22}$$

O modelo de regressão linear generalizada é representado pela Equação 2.22 onde $g(\mu_i)$ é a função de ligação, β_j são os parâmetros a serem obtidos pela regressão e x_{ij} os valores da variável independente (também chamada de variável exógena). Se $g(\mu_i)$ é a função identidade então a regressão linear generalizada é reduzida a regressão linear clássica.

2.2 Regressão logística

A regressão logística, também é conhecida na literatura como regressão *logit*, classificação de entropia máxima (*MaxEnt*) ou classificador *log-linear*, é uma técnica estatística que permite modelar a probabilidade de um evento ocorrer, dado um conjunto de variáveis independentes, podendo estas serem contínuas ou categóricas. Esse método é comumente usado em situações em que a variável dependente (ou de resposta) é binária, ou seja, pode assumir dois possíveis estados (como "sim"ou "não", "sucesso"ou "falha", "doente"ou "saudável") (PEDREGOSA et al., 2011), (HOSSMER; LEMESHOW, 2000), (DE; MOREIRA; DAS, 2021).

O próprio ENEM usa a regressão logística para calcular a proficiência dos alunos com base nos erros e acertos realizados na prova e, dessa forma, imputar-lhes uma nota.

Para ilustrar o método de regressão logística, tomemos uma simplificação do problema de determinar a nota de alunos baseado no conjunto de dados obtidos em uma aplicação do exame segundo a TRI (teoria de resposta ao item). Um aluno pode acertar ou errar uma dada questão do exame, ou seja, a variável que informa se o aluno acertou ou não uma dada questão é categórica. Tomando os mesmos 10 alunos do exemplo de regressão linear podemos verificar os alunos que foram aprovados e os que foram reprovados no exame (IPEA, 2019), (GARCIA, 2019), (FERREIRA, 2018).

Tabela 2 – Dados fictícios para exemplo de regressão logística

Aluno	y **nota na avaliação**	x_1 **horas de estudo**	x_2 **número de advertências**	x_3 **Aprovado**
1	3.0	0.2	9	0
2	2.8	0.3	7	0
3	4.8	0.5	8	0
4	5.5	1.5	5	0
5	6.1	1.5	5	1
6	5.0	2.1	3	0
7	6.8	2.4	2	1
8	7.5	3.3	1	1
9	8.2	3.5	0	1
10	10.0	3.9	0	1

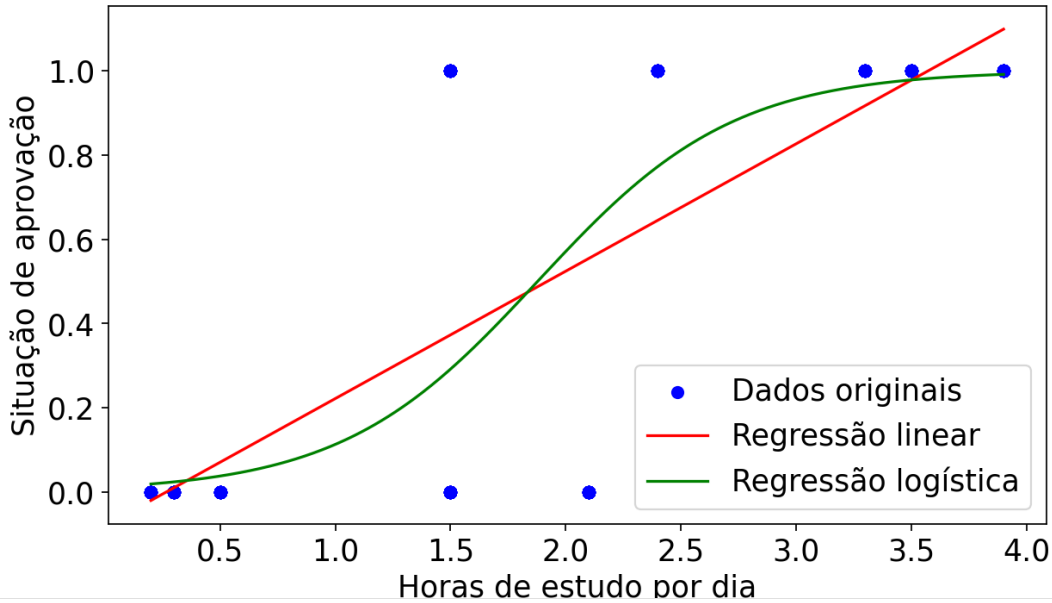
Suponha que a aprovação não dependa apenas de nota na avaliação, mas que uma nota alta contribua para a aprovação no exame, dessa forma existe uma relação entre a nota do aluno e a nova variável. Quanto maior a nota do aluno maior a chance de ele ter sido aprovado. É possível representar a variável *"Aprovado"* numericamente informando a reprovação por zero e a aprovação por 1 de modo a realizar uma regressão, no entanto essa nova variável é categórica e uma regressão linear simples não é adequada para se estabelecer um modelo que descreva a relação entre as variáveis. Em vez disso, é interessante ajustar uma curva do tipo *sigmoide*. Tomemos a aprovação do aluno como uma variável dependente e a nota como variável independente. A Figura 7 mostra uma comparação entre as regressões logística e linear com os dados do exemplo.

Repare que, dado a natureza categórica da variável *"Aprovado"*, a curva *sigmoide* se ajusta melhor aos dados, não apresenta tendência de representar os valores categóricos além do intervalo 0 e 1 e tem o comportamento desejado de mudar a representação de 0 para 1 a partir de um determinado valor da variável contínua independente. A curva *sigmoide* é dada pela equação:

$$\pi(x) = \frac{c}{1 + e^{-a(x+b)}} \tag{2.23}$$

Os parâmetros a, b e c da curva sigmoide ajustam sua posição no gráfico.

Figura 4 – Exemplo de regressão linear e logística sobre o mesmo conjunto de dados.



Fonte: Produção do próprio autor.

- **Parâmetro a:** Ajusta a inclinação máxima do gráfico, também chamada de nitidez. Quanto maior esse parâmetro mais intensamente a função muda de 0 para 1.
- **Parâmetro b:** Move a função ao longo do eixo x. É o parâmetro que indica o valor de x onde a maior diferenciação entre as variáveis categóricas ocorrem.
- **Parâmetro c:** determina o máximo da função, ou seja, a função sigmoide assume valores entre 0 e c. Se $c = 1$ então a função varia de 0 a 1.

Fazendo $c = 1$ e reorganizando os termos a e b a Equação 2.23 pode ser reescrita como 2.24 onde β_0 e β_1 são parâmetros que ajustam a curva.

$$\pi(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x)}} \tag{2.24}$$

No campo da estatística, a função sigmoide é frequentemente usada para inferir a probabilidade de uma determinada amostra pertencer a uma determinada classe, ou seja, o valor da função entre zero e um é interpretado como a probabilidade da mostra pertencer a categoria dado a variável independente.

Para ajustar os parâmetros da regressão logística usamos o modelo de regressão linear generalizada com a função de ligação *logit* dada por:

$$logit(p) = log\left(\frac{p}{1-p}\right) \tag{2.25}$$

Note que enquanto a função sigmoide tem domínio de $-\infty$ a $+\infty$ e imagem entre 0 e 1 a função *logit* é o oposto, domínio entre 0 e 1 e imagem de $-\infty$ a $+\infty$. A função sigmoide é a que se ajusta aos dados e a função logit é a função de ligação entre a relação das variáveis e o modelo linear. As Figuras 5 e 6 ilustram isso. Assim:

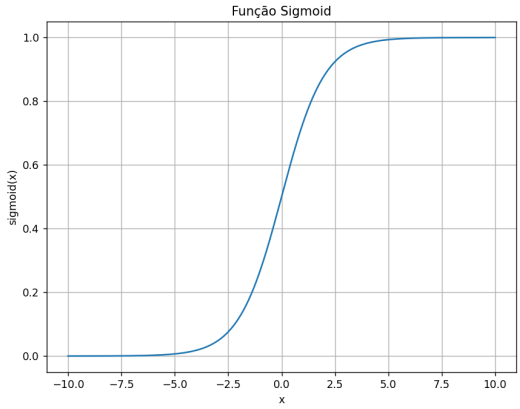


Figura 5 – Exemplo da curva sigmoide

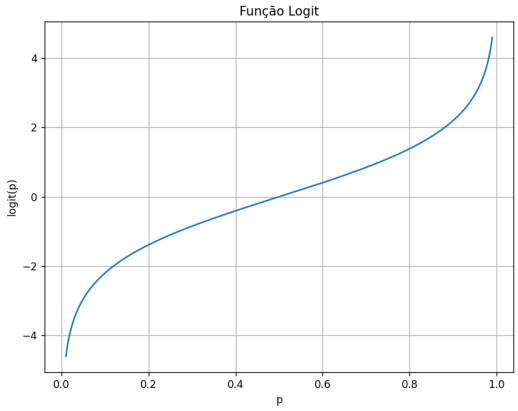


Figura 6 – exemplo da curva *logit*

$$logit(\pi(x)) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) \tag{2.26}$$

$$logit(\pi(x)) = \beta_0 + \beta_1 x \tag{2.27}$$

Onde β_0 e β_1 são coeficientes lineares que podem ser determinados de forma similar ao modelo clássico de regressão linear. O modelo de regressão logística para nosso exemplo seria:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \times (\text{horas de estudo}) \tag{2.28}$$

ou, resolvendo para p, obtemos:

$$p = \frac{1}{1+\exp(-[\beta_0 + \beta_1 \times (\text{horas de estudo})])} \tag{2.29}$$

Aqui, β_0 e β_1 são coeficientes a serem estimados a partir dos dados. β_0 é o logit da probabilidade de aprovação quando as horas de estudo são 0, e β_1 é o efeito logarítmico de uma hora adicional de estudo na probabilidade de aprovação.

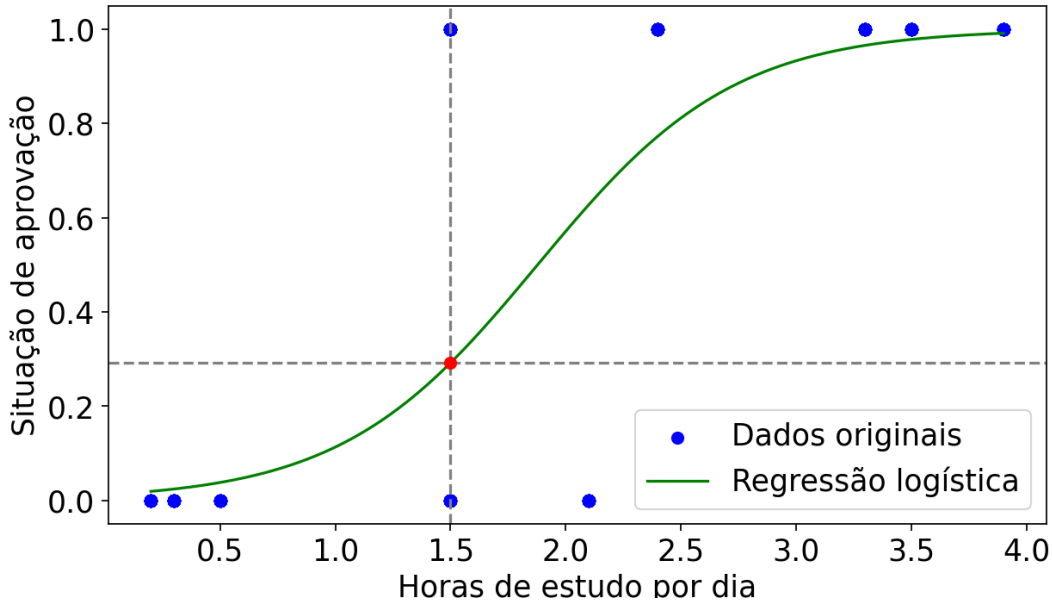
Ao estimar os coeficientes β_0 e β_1 a partir dos dados, podemos usar o modelo para prever a probabilidade de aprovação para um aluno dado o número de horas de estudo.

Para exemplificar, suponha que, após ajustar o modelo aos dados, obtenhamos $\beta_0 = -4.4057$ e $\beta_1 = 2.3458$. Se um aluno estuda 1,5 horas por dia, a probabilidade prevista de aprovação seria:

$$p = \frac{1}{1 + \exp(-[-4.4057 + 2.3458 \times 1,5])} \approx 0,2917 \tag{2.30}$$

Então, o modelo prevê que a probabilidade de aprovação para um aluno que estuda $1,5$ horas por dia é de 29%. A Figura 7 ilustra esse caso. Vale notar que o aluno 4 e o aluno 5 estudaram $1,5$ horas, entretanto apenas um deles foi aprovado. O modelo prevê uma probabilidade do aluno ser aprovado, mas não determina sua aprovação.

Figura 7 – Exemplo de regressão linear e logística sobre o mesmo conjunto de dados.



Fonte: Produção do próprio autor.

2.2.1 Estimação dos Coeficientes

A Equação 2.27 possui termos β_0 e β_1 que podem ser ajustados de modo a melhor descrever um conjunto de dados categóricos onde uma variável x pode pertencer ou não a uma categoria Y. No caso da regressão linear clássica foi desenvolvida uma forma de medir os desvios do ajuste com a Equação 2.2 e, através de suas derivadas parciais, obter os parâmetros que tornam esse desvio o menor possível, mas esse método não pode ser aplicado na regressão logística pois, segundo Maroco (2007), o método de mínimos quadrados não funciona pois o modelo logístico é binário, não segue uma distribuição de erro normal e nem apresenta variância constante.

Considere novamente os alunos 4 e 5 que estudaram $1,5$ horas por dia. O modelo nos informa que, para eles a chance de aprovação é de 29%, entretanto um foi aprovado e o outro não. Não é correto afirmar que o modelo errou uma previsão e acertou a outra já que o que ele fornece é uma informação estatística. A ideia é que se mais e mais alunos que estudam 1,5 horas por dia fossem considerados nos dados deveriam se distribuir de modo que 29% deles fossem aprovados e os demais reprovados.

Seja $X = (x_i, y_i)$ com $i = 1, 2, 3, ..., n$ observações independentes de um experimento aleatório onde y_i é uma variável dependente dicotômica (ou seja $y_i = 1$ se a observação pertence a uma dada categoria ou $y_i = 0$ se a observação não pertence). Dado um vetor de parâmetros $\beta = (\beta_0, \beta_1)$ o modelo de regressão fornece uma previsão probabilística para $y_i = 1$ dado o valor de x_i (no exemplo seria a probabilidade de ser aprovado dado as horas de estudo). Para estimar a qualidade do modelo é possível realizar a combinação das probabilidades para cada dado observado. Esse método é chamado de teste de hipótese por verosimilhança foi sugerido em 1928 por Neyman, Jerzy e Pearson, Egon Sharpe (DODGE, 2008).

Seja $\pi(\beta, x_i) = P(Y = 1|x_i, \beta)$ a probabilidade fornecida pelo modelo da variável dependente ser 1 dada a variável independente x_i e os parâmetros β e $P(Y = 0|x, \beta)$ a probabilidade dada pelo modelo de a variável dependente ser 0. Como a variável dependente é dicotômica tem-se que $P(Y = 0|x, \beta) = 1 - P(Y = 1|x, \beta)$. Assim é possível escrever essa relação de forma geral.

$$[P(Y = 1|x_i, \beta)]^{y_i}[1 - P(Y = 1|x_i, \beta)]^{1-y_i} \tag{2.31}$$

Ou ainda:

$$\pi(\beta, x_i)^{y_i}(1 - \pi(\beta, x_i))^{1-y_i} \tag{2.32}$$

Define-se a função de verossimilhança em função dos parâmetros β e das observações X como:

$$\ell(\beta, X) = \prod_{i=1}^{n}[\pi(\beta, x_i)]^{y_i}[1 - \pi(\beta, x_i)]^{1-y_i} \tag{2.33}$$

O logaritmo da verossimilhança é dado por:

$$L(\beta, X) = \ln \ell(\beta, X) = \sum_{i=1}^{n} y_i \ln(\pi(\beta, x_i)) + (1 - y_i) \ln(1 - \pi(\beta, x_i)). \tag{2.34}$$

A fim de se obter os parâmetros que melhor se ajustam aos dados $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)$ a função do logaritmo da verossimilhança é diferenciada em cada parâmetro da seguinte forma:

$$\left.\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_0}\right|_{(\hat{\beta}_0, \hat{\beta}_1)} = 0$$

$$\sum_{i=1}^{n} \left(\frac{y_i}{\pi(\beta, x_i)} \frac{\partial[\pi(\beta, x_i)]}{\partial \beta_0} + \frac{1 - y_i}{1 - \pi(\beta, x_i)} \frac{\partial[1 - \pi(\beta, x_i)]}{\partial \beta_0} \right) = 0 \tag{2.35}$$

Da Equação 2.25 pode-se escrever as derivadas parciais em β_0 e β_1.

$$\frac{\partial \pi(\beta, x_i)}{\partial \beta_0} = \frac{\partial}{\partial \beta_0} \left(\frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = \frac{e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} \tag{2.36}$$

$$\frac{\partial \pi(\beta, x_i)}{\partial \beta_1} = \frac{\partial}{\partial \beta_1} \left(\frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = \frac{x_i e^{-(\beta_0 + x_i \beta_1)}}{(1 + e^{-(\beta_0 + x_i \beta_1)})^2} \tag{2.37}$$

Substituindo 2.36 em 2.35 obtém-se:

$$\sum_{i=1}^{n} \left(y_i \frac{e^{-(\beta_0 + x_i \beta_1)}}{1 + e^{-(\beta_0 + x_i \beta_1)}} + (y_i - 1) \frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = 0$$

$$\sum_{i=1}^{n} \left(y_i - \frac{1}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = 0 \tag{2.38}$$

Além disso:

$$\left.\frac{\partial L(\beta_0, \beta_1)}{\partial \beta_1}\right|_{(\hat{\beta}_0, \hat{\beta}_1)} = 0$$

$$\sum_{i=1}^{n} \left(\frac{y_i}{\pi(\beta, x_i)} \frac{\partial}{\partial \beta_1} \pi(\beta, x_i) + \frac{1 - y_i}{1 - \pi(\beta, x_i)} \frac{\partial}{\partial \beta_1} \big(1 - \pi(\beta, x_i)\big) \right) = 0 \tag{2.39}$$

Novamente substituindo 2.37 em 2.39 obtém-se:

$$\sum_{i=1}^{n} \left(y_i x_i \frac{e^{-(\beta_0 + x_i \beta_1)}}{1 + e^{-(\beta_0 + x_i \beta_1)}} + (y_i - 1) \frac{x_i}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = 0$$

$$\sum_{i=1}^{n} \left(y_i x_i - \frac{x_i}{1 + e^{-(\beta_0 + x_i \beta_1)}} \right) = 0 \tag{2.40}$$

Por fim:

$$\sum_{i=1}^{n} (y_i - \pi(x_i)) = 0 \tag{2.41}$$

$$\sum_{i=1}^{n} x_i(y_i - \pi(x_i)) = 0 \tag{2.42}$$

Analogamente a estimação de parâmetros pela maximização da verossimilhança pode ser estendida para um número finito de variáveis dependentes de forma que $\beta = (\beta_0, \beta_1, ..., \beta_k)$ onde k é a quantidade de variáveis independentes a serem consideradas no modelo, ou seja, podemos estimar $k + 1$ parâmetros pelo método de maximização da verossimilhança (HOSSMER; LEMESHOW, 2000). Além disso as equações 2.41 e 2.42 não são lineares e exigem um método numérico iterativo para encontrar os parâmetros β que maximizam a função de verossimilhança, no entanto não é escopo deste trabalho detalhar os métodos numéricos que poderiam ser aplicados. Entre os muitos métodos disponíveis na literatura (NOCEDAL; WRIGHT, 2006) optou-se pelo *BFGS* devido ao grande volume de dados que devem ser manipulados a fim de se estimar os parâmetros.

O método BFGS (Broyden-Fletcher-Goldfarb-Shanno) é frequentemente empregado como um algoritmo de otimização para maximizar a função de verossimilhança em problemas estatísticos, provendo diversas vantagens quando aplicado em grandes volumes de dados além de escalar bem para um grande número de dados(SCHRAUDOLPH; YU; GüNTER, 2007), ou seja, é capaz de processar de forma eficiente uma grande quantidade de parâmetros e observações, tornando-se uma escolha apropriada para análises de dados extensas e complexas além de ser capaz de encontrar uma solução aproximada para a maximização da verossimilhança em um número relativamente pequeno de iterações.

2.2.2 Interpretação dos Coeficientes

Os coeficientes $\beta_0, \beta_1, \beta_2, ..., \beta_p$ fornecem informações sobre a relação entre as variáveis independentes e as probabilidades das categorias da variável dependente. A interpretação dos coeficientes pode variar dependendo da codificação utilizada para a variável dependente e das categorias de referência. Voltando ao exemplo do desempenho dos alunos da Tabela 2 o problema pode ser abordado como uma regressão logística onde horas de estudo e número de advertência são variáveis independentes e a situação de aprovação a variável dependente, logo a regressão terá um coeficiente que representa a variável horas de estudo (digamos β_1) e outra que representa número de advertências (digamos β_2). Cada coeficiente β expressa o quão intensamente o logaritmo da probabilidade de aprovação varia com a mudança da variável independente associada. Por exemplo, espera-se que um aluno com mais horas de estudo tenha maior chance de ser aprovado então o coeficiente β_1 deve ser maior que zero. Espera-se que o aluno com maior número de advertência tenha menor chance de aprovação então β_2 deve ser menor que zero. Um coeficiente com valor próximo a zero indica pouca correlação entre a variável independente e a dependente. A comparação do módulo dos coeficientes pode informar a variável que mais afeta o resultado

da variável dependente. (FIGUEIRA, 2006), (HOSSMER; LEMESHOW, 2000), (JúNIA; MARIZ, 2021), (DE; MOREIRA; DAS, 2021).

2.3 Pré processamento

Antes de executar as regressões desejadas os dados originais devem ser processados de forma a se tornarem entradas adequadas para o modelo de regressão. Para isso é utilizado técnicas de discretização, binarização e normalização.(GÉRON, 2019), (HOSSMER; LEMESHOW, 2000).

2.3.1 Discretização de Variáveis Contínuas

A discretização de variáveis contínuas é um processo essencial em muitos campos da matemática aplicada, estatística e ciência de dados. A transformação de variáveis contínuas em variáveis discretas facilita a análise, modelagem e interpretação dos dados. Neste capítulo, explora-se duas abordagens comuns para discretização: o método de intervalo de valores e o método de intervalo de quantis (MORETTIN; BUSSAB, 2017), (DEVORE, 2010) (ANDERSEN, 1997).

O método de intervalo de valores consiste em dividir o domínio da variável contínua em intervalos fixos. Essa abordagem é frequentemente utilizada quando se busca uma representação simplificada dos dados, reduzindo a complexidade computacional e facilitando a comunicação dos resultados. A escolha adequada do tamanho do intervalo é crucial, pois impacta diretamente na sensibilidade da análise.

A discretização por intervalo de valores é fundamentada na premissa de que a variação dentro de cada intervalo é insignificante em comparação com a variação total da variável contínua. Essa simplificação é útil em contextos onde a precisão é menos crítica, mas a interpretabilidade é prioritária.

O método de intervalo de quantis, por outro lado, baseia-se na divisão dos dados em intervalos que contenham aproximadamente a mesma quantidade de observações. A ideia subjacente é preservar a distribuição original dos dados e garantir que cada intervalo contenha uma proporção representativa da variabilidade total.

Os quantis, como os quartis e os percentis, são utilizados para determinar os limites dos intervalos. Essa abordagem é particularmente valiosa quando se busca manter informações detalhadas sobre a distribuição dos dados, sendo útil em análises mais sensíveis à variação local.

Como exemplo considere novamente os dados dos dez alunos da Tabela 2. Suponha que se deseja discretizar a variável *nota na avaliação*. Pode-se separar os alunos em duas ou mais categorias. Discretizando os alunos em duas categorias (A ou B, 0 ou 1, bom ou

Tabela 3 – Exemplo de discretização por intervalo de valores e de quantis.

Nota	intervalo de valores	intervalo de quantis
2.8	0	0
3.0	0	0
4.8	0	0
5.0	1	0
5.5	1	0
6.1	1	1
6.8	1	1
7.5	1	1
8.2	1	1
10.0	1	1

ruim), o critério que enquadra o aluno em cada categoria pode ser baseada por intervalo de valores ou de quantis. No caso do intervalo de valores como a nota da prova varia de zero a dez a categorização pode ser feita da seguinte forma, *se* $0 < nota < 5$ *então o aluno pertence a categoria 0, caso contrário pertence a categoria* 1.

A escolha entre os métodos de intervalo de valores e intervalo de quantis depende das metas específicas da análise e das características dos dados. O método de intervalo de valores é eficaz quando se deseja simplificar a representação dos dados e a interpretação dos resultados. Por outro lado, o método de intervalo de quantis é preferível quando se busca preservar a distribuição original e a sensibilidade à variação local é crucial.

Ambos os métodos têm aplicações em diversas áreas, incluindo economia, ciências sociais, engenharia e medicina. A escolha da abordagem mais apropriada deve considerar a natureza dos dados, os objetivos da análise e as demandas específicas do problema em questão.

2.3.2 Binarização

Uma variável binária é aquela que admite apenas um dentre dois estados geralmente representados por zero e um, mas que também pode ser interpretado como "sim"e "não"ou ainda "verdadeiro"e "falso". Esse tipo de variável é adequada para o modelo de regressão logística uma vez que, atribuindo valor zero ou um, pode-se realizar o produto da variável com seu respectivo coeficiente a ser ajustado pelo modelo e assim ter sua contribuição na função de ajuste. Por outro lado existem variáveis categóricas não binárias. Esse tipo de variável não se adéqua ao modelo, por exemplo a variável "estado civil"do conjunto de microdados ENEM2.3.2 que possui valores de 0 a 3 que não representam uma quantidade, mas o estado civil como 0 solteiro, 1 casado, 2 divorciado e 3 viúvo. A categoria 2 (divorciado) não representa uma quantidade superior a 1 (casado) logo usar esses números na regressão não faz sentido. Para isso utiliza-se o método de binarização onde variáveis

Tabela 4 – Exemplo de binarização da variável "Estado Civil".

Variável original		Após a binarização	
Estado civil	0: Solteiro	Solteiro	0: não
	1: Casado		1: sim
	2: Divorciado	Casado	0: não
	3: Viúvo		1: sim
		Divorciado	0: não
			1: sim
		Viúvo	0: não
			1: sim

multi nominais (como o caso do estado civil) que possuam mais de uma categoria não ordinal são transformadas em mais de uma variável binária. Note que esse processo produz n novas variáveis onde n é o número de categorias da variável nominal. Com isso apenas uma dentre as n variáveis possuirá valor 1 enquanto as demais possuirão valor 0 o que implica em uma dependência entre as novas variáveis que é prejudicial para o modelo de regressão pois é importante que as variáveis exógenas sejam independentes entre elas.

2.3.3 Normalização

O processo de normalização de variáveis desempenha um papel crucial no contexto de modelos de regressão. Em muitas situações, as variáveis de entrada podem apresentar escalas diferentes, o que pode afetar negativamente a convergência e a estabilidade do modelo. A normalização visa mitigar esses problemas, proporcionando um ambiente mais equilibrado para o algoritmo de regressão. Nesta subseção, exploraremos a utilidade e o funcionamento do método de normalização de variáveis em modelos de regressão.

A normalização das variáveis métricas promovem:

Estabilidade Numérica: Variáveis em escalas distintas podem resultar em operações numéricas instáveis durante o treinamento do modelo. A normalização reduz a amplitude das variáveis, evitando valores extremamente grandes ou pequenos que possam afetar a precisão numérica.

Convergência Mais Rápida: Algoritmos de otimização utilizados em modelos de regressão, como o gradiente descendente, muitas vezes convergem mais rapidamente quando as variáveis estão em uma escala similar. A normalização contribui para uma convergência mais eficiente do algoritmo, reduzindo o número de iterações necessárias para atingir um mínimo global ou local.

Interpretabilidade dos Coeficientes: A normalização das variáveis facilita a interpretação dos coeficientes do modelo. Os coeficientes normalizados representam a mudança média na variável dependente associada a uma mudança de uma unidade nas variáveis independentes, proporcionando uma interpretação mais intuitiva.

Existem vários métodos de normalização, dentre elas a normalização Min-Max que dimensiona as variáveis para um intervalo específico, como [0, 1]. A fórmula para a normalização Min-Max é dada por:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

Onde X_{min} e X_{max} são, respectivamente, o valor mínimo e máximo da variável.

2.4 Medida de qualidade do modelo

A avaliação da performance de modelos de regressão é crucial para garantir a generalização adequada a dados não observados e, consequentemente, a utilidade prática do modelo em questão. Dois elementos essenciais na validação de modelos são a divisão dos dados em conjuntos de treino e teste, e a utilização de escores para mensurar a eficácia do modelo (GÉRON, 2019).

2.4.1 Separação de Dados em Conjuntos de Treino e Teste

A separação dos dados em conjuntos de treino e teste é uma prática fundamental na construção e validação de modelos de aprendizado de máquina, incluindo a regressão logística. A ideia central é treinar o modelo em uma parte dos dados e testá-lo em outra parte, permitindo uma avaliação mais realista de como o modelo se comportará em situações não vistas anteriormente.

O conjunto de treino é utilizado durante a fase de treinamento, na qual o modelo aprende os padrões presentes nos dados. Já o conjunto de teste é reservado para avaliar o desempenho do modelo em dados não utilizados durante o treinamento, proporcionando uma estimativa mais confiável de sua capacidade de generalização.

A divisão adequada dos dados é crucial para evitar problemas de overfitting, nos quais o modelo se ajusta excessivamente aos dados de treino, mas falha em generalizar para novos dados. Usualmente utiliza-se de 10% a 20% dos dados disponíveis para se testar o modelo. Esses dados devem ser selecionados de forma aleatória da base. Ferramentas como a biblioteca *scikit-lear* oferece métodos completos para se fazer essa separação dos dados (GÉRON, 2019).

2.4.2 Medida da acurácia do modelo

A verossimilhança usada para se obter os coeficientes não é uma boa forma de se medir a qualidade do modelo uma vez que seu valor tende a zero a medida em que se adiciona mais dados. Uma forma de medir a qualidade do modelo é através da pontuação

por acurácia. A acurácia do modelo pode ser entendida como o quão perto ele se aproxima dos dados reais (DODGE, 2008). Na prática essa pontuação é calculada da seguinte forma:

$$acurácia(y, \hat{y}) = \frac{\sum_{i=0}^{n-1} 1(\hat{y}_i = y_i)}{n} \tag{2.43}$$

A função $1(\hat{y}_i = y_i)$ é uma função indicadora que assume valor 1 caso a condição seja atendida e zero para qualquer outro caso, n é o número de amostras, $\hat{y}_i$ é o resultado previsto pelo modelo para uma dada entrada X_i e y_i é o resultado esperado. Por exemplo, tomemos um candidato que realizou a prova do ENEM. Esse candidato possui um conjunto de variáveis X_0 que, por hipótese, são independentes. Esse candidato que realizou a prova obteve uma nota que o coloca em uma categoria de desempenho y_0 que, por hipótese, é a variável dependente do modelo. Dado como entrada para o modelo o conjunto de dados X_0 obtém-se uma saída $\hat{y}_0$. Se $\hat{y}_0 = y_0$ o modelo "acertou"a previsão e isso aumenta a pontuação de acurácia, caso contrário o modelo "errou"e isso diminui a pontuação.

A pontuação de acurácia (*accuracy score*) é um valor que varia entre zero e um de modo que quanto mais próximo de um melhor o modelo previu o resultado esperado.

Outra vantagem de medir a qualidade do modelo pela pontuação de acurácia é que esta permite a medição sob a mesma escala para qualquer tipo de modelo incluindo o modelo fictício.

2.4.3 Modelo fictício

Na prática estatística, a inclusão de um modelo fictício, também conhecido como "*dummy model*"ou "*baseline model*", é defendida por diversas razões relevantes (GÉRON, 2019), (PEDREGOSA et al., 2011).

- Estabelecimento de Referência:

  A incorporação de um modelo fictício se revela útil como uma linha de base simples. A comparação de modelos mais complexos com essa abordagem básica permite a avaliação de possíveis melhorias significativas proporcionadas pelos modelos mais avançados.

- Interpretação da Performance:

  A presença de um modelo fictício simplifica a interpretação da performance de modelos mais sofisticados. A ausência de superação desse modelo mais simples por parte de abordagens mais avançadas suscita questionamentos pertinentes sobre sua aplicabilidade prática. Em contrapartida, a superação do modelo fictício sugere a entrega de informações relevantes pelos modelos mais complexos.

- Avaliação de Generalização:

  Ao comparar um modelo com um modelo fictício, é possível avaliar sua capacidade de generalização para além do aprendizado dos padrões presentes nos dados de treino. Essa análise é essencial para assegurar que o modelo não esteja apenas memorizando os dados de treino, mas sim extrapolando para novos contextos.

- Simplicidade e Eficiência:

  Modelos fictícios são caracterizados por sua simplicidade e eficiência computacional. Eles proporcionam uma abordagem direta e eficaz para realizar avaliações preliminares, sendo particularmente valiosos em estágios iniciais de desenvolvimento de modelos.

- Compreensão do Problema:

  A utilização de um modelo fictício frequentemente reflete uma abordagem ingênua ou uma expectativa mínima em relação ao desempenho do modelo. Isso pode contribuir significativamente para uma compreensão mais profunda do problema em análise, fornecendo insights valiosos sobre sua complexidade intrínseca.

- Facilitação da Comunicação:

  A inclusão de um modelo fictício simplifica a comunicação com partes interessadas não técnicas. A demonstração da melhoria do modelo em relação a um ponto de referência simples torna os resultados mais acessíveis e compreensíveis, promovendo uma comunicação eficaz. Em resumo, a prática de incorporar um modelo fictício é respaldada por sua utilidade substancial na avaliação da eficácia de modelos mais avançados. Essa abordagem não apenas facilita a interpretação dos resultados, mas também fomenta a transparência e a compreensão do processo de modelagem estatística.

Para chegar a um modelo fictício adequado se faz necessário compreender os dados que se está trabalhando, por exemplo se o que se quer é prever o desempenho de um aluno baseando-se nas suas informações socioeconômicas pode-se criar um modelo simples em que alunos de renda mais alta são colocados em categorias de desempenho mais elevadas ou ainda pode-se classificar todos os alunos como sendo de uma categoria de desempenho mais frequente. A partir da medida de desempenho do modelo fictício temos uma referência para avaliar o modelo ajustado.

2.5 Dados INEP

O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) desempenha um papel fundamental na organização e disponibilização de dados sobre

educação no Brasil. Criado em 1937 como Instituto Nacional de Pedagogia e transformado em autarquia federal vinculada ao Ministério da Educação (MEC) em 1997, o INEP é responsável por fornecer informações precisas e confiáveis sobre o sistema educacional do país (SAVIANI, 2012), (FONSECA; NAMEN, 2016).

Uma das principais atividades do INEP é a organização e aplicação de avaliações e exames educacionais, sendo o Exame Nacional do Ensino Médio (ENEM) um dos destaques. O ENEM, aplicado anualmente, tem como objetivo avaliar o desempenho dos estudantes no final da educação básica e é reconhecido como um importante indicador da qualidade do ensino oferecido pelas escolas brasileiras. Os microdados do ENEM, disponibilizados pelo INEP, são uma valiosa fonte de informação para pesquisadores, gestores educacionais e demais interessados em compreender os fatores que influenciam o desempenho dos estudantes.

Além do ENEM, o INEP também é responsável pela produção e divulgação de indicadores educacionais, que fornecem uma visão abrangente do sistema educacional brasileiro. Esses indicadores abrangem diversos aspectos, como taxas de matrícula, taxas de conclusão, infraestrutura escolar, formação dos professores, entre outros. Os dados disponibilizados pelo INEP permitem o monitoramento da qualidade da educação, a identificação de desigualdades regionais e a formulação de políticas públicas mais efetivas no campo educacional.

A disponibilização desses dados pelo INEP é de extrema importância para a comunidade acadêmica e para os gestores da educação. Pesquisadores podem acessar os microdados do ENEM e realizar análises estatísticas detalhadas para investigar os determinantes do desempenho dos estudantes, identificar padrões e tendências, e contribuir para o desenvolvimento de estudos na área da educação. Os gestores educacionais podem utilizar os indicadores educacionais produzidos pelo INEP para monitorar o progresso das escolas, identificar pontos de melhoria e embasar decisões para aprimorar a qualidade da educação no país.

Dessa forma, a disponibilidade dos dados do ENEM e dos indicadores educacionais pelo INEP é fundamental para promover a transparência e o acesso à informação sobre a educação no Brasil. Esses dados são ferramentas valiosas para a realização de pesquisas, a formulação de políticas educacionais e o monitoramento do sistema educacional. Ao utilizar esses dados em estudos e análises, é possível obter *insights* relevantes que contribuem para a compreensão dos desafios e das oportunidades no campo da educação, visando sempre o aprimoramento e a equidade educacional no país.

2.5.1 ENEM

O Exame Nacional do Ensino Médio (ENEM) é uma prova realizada anualmente pelo Ministério da Educação (MEC) no Brasil, desde o ano 1998. Seu principal objetivo é avaliar a qualidade do ensino médio no país, além de oferecer aos estudantes a oportunidade de ingressar em universidades públicas e privadas, através do Sistema de Seleção Unificada (Sisu), do Programa Universidade para Todos (ProUni) e do Fundo de Financiamento Estudantil (Fies).

O ENEM é constituído por uma prova com questões objetivas e uma redação. As questões objetivas são divididas em quatro áreas de conhecimento: Linguagens, Códigos e suas Tecnologias, Matemática e suas Tecnologias, Ciências Humanas e suas Tecnologias, e Ciências da Natureza e suas Tecnologias. Cada área de conhecimento é composta por 45 questões, totalizando 180 questões objetivas, com pesos diferentes para cada área, conforme o edital do exame.

A redação do ENEM é avaliada segundo cinco critérios: domínio da norma culta da língua portuguesa; compreensão da proposta da redação e aplicação de conceitos de diversas áreas do conhecimento; organização das informações e argumentos; coerência e coesão do texto; e elaboração de uma proposta de intervenção para o problema abordado, respeitando os direitos humanos. A nota da redação varia de 0 a 1000 pontos.

Além de sua função avaliativa e seletiva, o ENEM também tem sido utilizado como ferramenta de diagnóstico e monitoramento do ensino médio no país. Os resultados do exame são divulgados em níveis nacional, estadual e municipal, permitindo a análise da qualidade do ensino médio em diferentes regiões e redes de ensino. A partir desses resultados, é possível identificar as principais dificuldades dos estudantes e traçar estratégias para a melhoria do ensino.

Desde sua criação, o ENEM tem passado por diversas mudanças, buscando se adaptar às necessidades do sistema educacional brasileiro e às demandas do mercado de trabalho. Em 2009, por exemplo, o exame foi reformulado para se tornar o principal critério de seleção para o acesso às universidades federais, substituindo o vestibular tradicional. Desde então, o ENEM tem sido cada vez mais valorizado pelas instituições de ensino superior, sendo utilizado como critério de seleção para a concessão de bolsas de estudos, financiamentos estudantis e até mesmo para a contratação de profissionais.

Em resumo, o ENEM é uma prova de grande importância para a educação brasileira, tendo como objetivos principais a avaliação da qualidade do ensino médio e o acesso ao ensino superior. Sua realização anual e a divulgação dos resultados em níveis nacional, estadual e municipal permitem a identificação das principais dificuldades dos estudantes e o desenvolvimento de estratégias para a melhoria do ensino.

2.6 Linguagem e Software

A linguagem *Python*, combinada com as bibliotecas *NumPy*, Pandas e *Statsmodels*, proporciona um ambiente robusto e versátil para análise estatística, incluindo a regressão logística ordinal.

Python é uma linguagem de programação de alto nível com uma sintaxe clara e concisa que enfatiza a legibilidade do código. Essa linguagem tornou-se uma das favoritas entre os cientistas de dados, matemáticos e estatísticos devido à sua simplicidade e ao suporte extensivo para operações numéricas e análise de dados (PAPERT, 1996).

A biblioteca *NumPy* é um dos componentes fundamentais do *stack* científico em *Python*. Ela fornece uma estrutura de dados de matriz poderosa e eficiente, juntamente com uma grande biblioteca de funções matemáticas de alto nível que operam em matrizes e *arrays* (VAART, 2000).

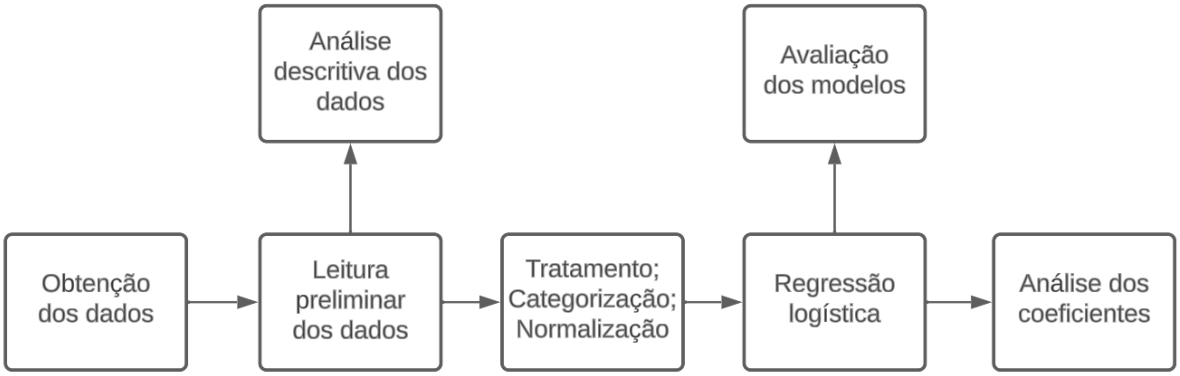
A biblioteca Pandas estende a funcionalidade do *NumPy* ao fornecer estruturas de dados de alto nível (Series e *DataFrame*) e ferramentas de manipulação de dados. Estas são especialmente úteis para organizar, transformar e analisar conjuntos de dados em formatos tabulares (MCKINNEY, 2010).

Finalmente, a biblioteca *Statsmodels* fornece uma grande quantidade de modelos estatísticos, incluindo a regressão logística ordinal por meio do módulo *statsmodels.miscmodels.ordinal_model*. *Statsmodels* é uma poderosa biblioteca para estimação de modelos estatísticos, realização de testes estatísticos e exploração de dados estatísticos. A regressão logística ordinal é particularmente útil quando a variável de resposta é ordinal, ou seja, categorias com uma ordem natural (SEABOLD; PERKTOLD, 2010).

3 Método

A Figura 8 ilustra a cadeia de procedimentos realizados.

Figura 8 – Fluxograma das etapas da pesquisa.



Fonte: Produção do próprio autor.

3.1 Obtenção dos dados

Os dados foram obtidos através do portal INEP que disponibiliza históricos das aplicações do exame nacional do ensino médio, indicadores das escolas do país entre outras informações. Para essa análise foi utilizada a base de dados históricos do ENEM dos anos de 2020, 2021 e 2022. A base de dados histórico ENEM de cada ano traz as seguintes informações:

- Itens das provas: Arquivo de extensão *csv* com a descrição das provas.
- Microdados ENEM: Arquivo de extensão *csv* com os dados dos candidatos que serão detalhados adiante.
- Dicionários: Arquivos extensão *ods* e *xlsx* com a descrição dos microdados.
- Inputs: Arquivos com dados de itens das provas e microdados em extensão *R*, *sas* e *sps*.
- Documentos técnicos: Arquivos sobre a avaliação como matriz de referência, manual de redação e edital.
- Provas e gabaritos: As provas aplicadas em cada variação de cor e seus respectivos gabaritos.

3.2 Leitura preliminar dos dados

Os dados fornecidos pelo portal do INEP são inicialmente lidos como texto através de procedimento nativo do *python* sem preocupação a respeito da interpretação de cada tipo de dado. O texto é segmentado em trechos definidos por *";"* e, dessa forma, um *dataframe* da biblioteca *pandas* é construído e exportado como documento de extensão *csv* que pode ser facilmente lido posteriormente como *data frame* através de procedimentos da mesma biblioteca.

Inicialmente os dados do ENEM 2020, 2021 e 2022 estão organizados como descritos a seguir:

- *Dados dos participantes* com 12 variáveis diferentes como número da inscrição, ano de aplicação e idade do candidato;
- *Dados da escola* com 7 variáveis que descrevem a escola onde o aluno estuda. Em anos anteriores era informado o código da escola que podia ser usado para cruzar os dados ENEM com outros indicadores do INEP, entretanto a partir de 2020 esse dado passou a não ser público;
- *Dados dos pedidos de atendimento especializado* com 13 variáveis informando as eventuais solicitações do aluno a recursos especiais para adequação de, por exemplo, baixa visão, surdez ou autismo;
- *Dados do local de prova* com 4 variáveis sobre o município onde a prova foi realizada;
- *Dados da prova objetiva* com 21 variáveis informando a presença do aluno durante a aplicação, cor, código e notas de cada uma das 4 provas objetivas bem como a escolha de língua estrangeira.
- *Dados da redação* com 7 variáveis contendo informações sobre a participação e desempenho na redação.
- *Dados do questionário socioeconômico* com 25 variáveis contendo as respostas do aluno ao questionário socioeconômico realizado no ato da inscrição na prova. Esse questionário traz informações como renda familiar, escolaridade dos familiares e itens de posse como geladeira, carro e banheiro.

3.3 Seleção de variáveis

A seleção de variáveis independentes para um modelo inferencial pode ser feita por estudo prévio do problema, segundo (HAIR, 2009). A escolha das variáveis bem como

parte da categorização feita a posteriori, foi realizada usando (JúNIA; MARIZ, 2021) como referência, mas adaptando aos dados das provas de 2020, 2021 e 2022.

O agrupamento de níveis de respostas de variáveis do modelo estão apresentadas na Tabela **??**. Antes de usar as variáveis para ajustar o modelo elas passaram por processo de normalização.

3.3.1 Sexo

A variável *TP_SEXO* pode possuir os valores *M* para candidatos do sexo masculino ou *F* para candidatos do sexo feminino. Para a aplicação do modelo de regressão ela foi transformada em uma variável categórica *Masculino* com o valor 1 para o caso do candidato ser do sexo masculino e 0 para o caso de ser do sexo feminino.

3.3.2 Estado Civil

Nos dados originais a variável *TP_ESTADO_CIVIL* pode assumir um entre 5 valores numéricos que categorizam o estado civil do candidato.

- 0: Não informado.
- 1: Solteiro(a).
- 2: Casado(a)/Mora com companheiro(a)
- 3: Divorciado(a)/Desquitado(a)/Separado(a)
- 4: Viúvo(a)

Para o modelo essa variável foi transformada na variável *Solteiro* que recede valor 1 se o candidato for solteiro e valor 0 em qualquer outro caso.

3.3.3 Cor Raça

Originalmente a variável *TP_COR_RACA* possui 6 estados discriminados por valores numéricos de 0 a 5 como a seguir.

- 0: Não declarado
- 1: Branca
- 2: Preta
- 3: Parda

Variável	**Níveis de Resposta**	**Níveis Agrupados**
Estado Civil	0 - Solteiro(a)	1 - Solteiro
	1 - Casado(a)/Mora com companheiro(a)	0 - Não Solteiro
	2 - Divorciado, etc	
	3 - Viúvo(a)	
Cor/raça	1 - Branca	1 - Branca
	2 - Preta	2 - Preta
	3 - Parda	
	4 - Amarela	0 - Outro
	5 - Indígena	
	0 - Não declarado	
Escolaridade Pai	H - Não sei.	0 - Não sei
	A - Nunca estudou.	A - Nunca estudou
	B - Não completou a 4ª série do Fundamental.	B - Completou no máximo o Ensino Médio
	C - Ensino Fundamental Incompleto.	
	D - Ensino Médio Incompleto	
	E - Graduação Incompleta.	
	F - Pós Graduação Incompleta.	C - Completou a faculdade e ou pós graduação
	G - Completou a Pós Graduação	
Escolaridade mãe	H - Não sei.	0 - Não sei
	A - Nunca estudou.	A - Nunca estudou
	B - Não completou a 4ª série do Fundamental.	B - Completou no máximo o Ensino Médio
	C - Ensino Fundamental Incompleto.	
	D - Ensino Médio Incompleto	
	E - Graduação Incompleta.	
	F - Pós-graduação Incompleta.	C - Completou a faculdade e ou pós graduação
	G - Completou a Pós graduação	
Ocupação pai	A - Grupo 1	A - Ocupação com menor requisito de formaçãoo
	F - Não sei.	
	B - Grupo 2	
	C - Grupo 3	
	D - Grupo 4	B - Ocupação com maior requisito de formaçãoo
	E - Grupo 5.	
Ocupação mae	A - Grupo 1	A - Ocupação com menor requisito de formaçãoo
	F - Não sei.	
	B - Grupo 2	
	C - Grupo 3	
	D - Grupo 4	B - Ocupação com maior requisito de formaçãoo
	E - Grupo 5.	
Renda	A - Nenhuma renda.	A - Até 1 salário mínimo
	B - Até R954, 00.	
	C - De R954, 01*até*R 1.431,00.	B - De 1 a 2,5 salários mínimos
	D - De R1.431, 01*até*R 1.908,00.	
	E - De R1.908, 01*até*R 2.385,00.	
	F - De R2.385, 01*até*R 2.862,00.	C - De 2,5 a 7 salários mínimos
	G - De R2.862, 01*até*R 3.816,00.	
	H - De R3.816, 01*até*R 4.770,00.	
	I - De R4.770, 01*até*R 5.724,00.	
	J - De R5.724, 01*até*R 6.678,00.	
	K - De R6.678, 01*até*R 7.632,00.	D - De 7 a 12 salários mínimos
	L - De R7.632, 01*até*R 8.586,00.	
	M - De R8.586, 01*até*R 9.540,00.	
	N - De R9.540, 01*até*R 11.448,00.	
	O - De R11.448, 01*até*R 14.310,00.	E - Acima de 12 salários mínimos
	P - De R14.310, 01*até*R 19.080,00.	
	Q - Mais de R19.080, 00.	
Computador em casa	A - Não.	A- Não
	B - Sim, um.	B - Sim
	C - Sim, dois.	
	D - Sim, três.	
	E - Sim, quatro ou mais.	
Conclusão do Ensino Médio	A - Já concluí o Ensino Médio.	A - Já concluí
	B - Concluirei o Ensino Médio em 2018.	B - Concluirei em 2018.
	C - Concluirei o Ensino Médio ap´os 2018.	C - Não concluirei em 2018
	D - Não concluí e não cursando	
Tipo de Escola	A - Somente em escola pública.	A - Somente em escola pública
	F - Não frequentei a escola	
	B - Escola pública e privada SEM bolsa	B - Parte em escola pública e parte em escola privada
	C - Escola pública e privada COM bolsa	
	D - Escola privada SEM bolsa	C - Somente em escola privada
	E - Escola privada COM bolsa	

- 4: Amarela
- 5: Indígena

Dessa variável foi gerado para o modelo duas variáveis categóricas:

- BRANCA: Recebe o valor 1 caso TP_COR_RACA for 1, caso contrário recebe 0.
- PRETA: Recebe valor 1 caso TP_COR_RACA for 2 ou 3, caso contrário recebe 0.

3.3.4 Situação de Conclusão

A variável *TP_ST_CONCLUSAO* da base de dado original é numérica de 1 a 4 e informa sobre a situação do candidato a respeito da conclusão do ensino médio como a seguir.

- 1 Já concluí o Ensino Médio.
- 2 Estou cursando e concluirei o Ensino Médio no ano de aplicação da prova.
- 3 Estou cursando e concluirei o Ensino Médio após o ano de aplicação da prova.
- 4 Não concluí e não estou cursando o Ensino Médio.

Para o modelo de regressão foram geradas as variáveis categóricas:

- Já concluiu o EM
- Concluirá o EM no ano de avaliação
- Não concluirá o EM até o ano de aplicação da prova

3.3.5 Tipo Escola

A base de dados original disponibilizadão pelo INEP traz a variável *TP_ESCOLA* que informa o tipo de escola que o candidato cursa ou cursou o ensino médio.

- 1 Não Respondeu
- 2 Pública
- 3 Privada

Alunos que se declararam de escola pública totalizam cerca de 70% dos candidatos em cada um dos anos de aplicação avaliados nesse trabalho. Por isso optou-se por discriminar apenas alunos de escola pública. A variável produzida para o modelo de regressão foi *Escola pública* que recebe valor 1 caso o candidato tenha se declarado de escola pública ou 0 caso contrário.

3.3.6 Treineiro

Os dados originais trazem a variável *TP_TREINEIRO* com o valor 1 caso o candidato tenha se declarado treineiro e 0 caso contrário. Essa variável foi integrada no modelo como *Treineiro* com o mesmo comportamento.

3.3.7 Escolaridade dos pais

As variáveis *Q001* e *Q002* são respostas do candidato ao questionário socioeconômico e trazem informações sobre a escolaridade do pai e da mãe respectivamente através de variável categórica discriminada a seguir:

- A: Nunca estudou.
- B: Não completou a 4ª série/5º ano do Ensino Fundamental.
- C: Completou a 4ª série/5º ano, mas não completou a 8ª série/9º ano do Ensino Fundamental.
- D: Completou a 8ª série/9º ano do Ensino Fundamental, mas não completou o Ensino Médio.
- E: Completou o Ensino Médio, mas não completou a Faculdade.
- F: Completou a Faculdade, mas não completou a Pós-graduação.
- G: Completou a Pós-graduação.
- H: Não sei.

Para o modelo de regressão essas variáveis foram agrupadas da seguinte forma:

- Escolaridade do Pai A: Baixa escolaridade.
- Escolaridade do Pai B: Escolaridade intermediária.
- Escolaridade do Pai C: Alta escolaridade

A Tabela **??** mostra detalhadamente o agrupamento dessas variáveis.

3.3.8 Ocupação dos pais

As variáveis *Q003* e *Q004* são respostas do candidato ao questionário socioeconômico e trazem informações sobre o tipo de ocupação do pai e da mãe respectivamente através de variável categórica. A pergunta feita ao candidato para a variável "Q003"foi:

> A partir da apresentação de algumas ocupações divididas em grupos ordenados, indique o grupo que contempla a ocupação mais próxima da ocupação do seu pai ou do homem responsável por você. (Se ele não estiver trabalhando, escolha uma ocupação pensando no último trabalho dele).
>
> Fonte: Dicionário de dados fornecido pelo INEP junto aos microdados.

A pergunta da questão Q004 é semelhante, mas se referindo a mãe do candidato em vez do pai. O agrupamento dessa variável para o modelo de regressão logística se deu em duas categorias, *Ocupação A* que agrupou as categorias originais A, B e C e *Ocupação B* que agrupou D e E. Ou seja, a variável *Ocupação A* indica que o pai ou a mãe tem ocupação laboral de baixa especialidade enquanto a variável *Ocupação B* indica uma ocupação menos especializada.

3.3.9 Renda Familiar

A sexta pergunta do questionário socioeconômico é sobre renda familiar. O texto da questão é a seguinte:

> Qual é a renda mensal de sua família? (Some a sua renda com a dos seus familiares.)
>
> Fonte: Dicionário de dados fornecido pelo INEP junto aos microdados.

As opções de respostas são categorias ordinais de A a Q onde A representa nenhuma renda mensal e Q uma renda superior a um valor mensal elevado. Dependendo do ano de realização das provas essa categoria é reajustada segundo o poder de compra do salário mínimo daquele ano, o que para fins de análise qualitativa ajuda o modelo a se comportar de forma consistente. A Tabela 5 mostra a faixa de valores para cada categoria nos anos de 2020, 2021 e 2022.

Para o modelo de regressão as categorias dessa variável foram agrupadas em novas variáveis binárias como a seguir.

- Renda A: 1 se a resposta a pergunta Q006 foi A ou B.
- Renda B: 1 se a resposta para a pergunta Q006 foi C, D ou E.

Tabela 5 – Valores monetários para cada categoria de renda familiar.

	Valores por categoria		
Categoria	2020	2021	2022
A	Nenhuma Renda	Nenhuma Renda	Nenhuma Renda
B	Até R1.045,00	Até R1.100,00	Até R1.212,00
C	De R1.045,01*até*R 1.567,50	De R1.100,01*até*R 1.650,00.	De R1.212,01*até*R 1.818,00.
D	De R1.567,51*até*R 2.090,00	De R1.650,01*até*R 2.200,00.	De R1.818,01*até*R 2.424,00.
E	De R2.090,01*até*R 2.612,50	De R2.200,01*até*R 2.750,00.	De R2.424,01*até*R 3.030,00.
F	De R2.612,51*até*R 3.135,00	De R2.750,01*até*R 3.300,00.	De R3.030,01*até*R 3.636,00.
G	De R3.135,01*até*R 4.180,00	De R3.300,01*até*R 4.400,00.	De R3.636,01*até*R 4.848,00.
H	De R4.180,01*até*R 5.225,00	De R4.400,01*até*R 5.500,00.	De R4.848,01*até*R 6.060,00.
I	De R5.225,01*até*R 6.270,00	De R5.500,01*até*R 6.600,00.	De R6.060,01*até*R 7.272,00.
J	De R6.270,01*até*R 7.315,00	De R6.600,01*até*R 7.700,00.	De R7.272,01*até*R 8.484,00.
K	De R7.315,01*até*R 8.360,00	De R7.700,01*até*R 8.800,00.	De R8.484,01*até*R 9.696,00.
L	De R8.360,01*até*R 9.405,00	De R8.800,01*até*R 9.900,00.	De R9.696,01*até*R 10.908,00.
M	De R9.405,01*até*R 10.450,00	De R9.900,01*até*R 11.000,00.	De R10.908,01*até*R 12.120,00.
N	De R10.450,01*até*R 12.540,00	De R11.000,01*até*R 13.200,00.	De R12.120,01*até*R 14.544,00.
O	De R12.540,01*até*R 15.675,00	De R13.200,01*até*R 16.500,00.	De R14.544,01*até*R 18.180,00.
P	De R15.675,01*até*R 20.900,00	De R16.500,01*até*R 22.000,00.	De R18.180,01*até*R 24.240,00.
Q	Acima de R20.900,00	Acima de R22.000,00.	Acima de R24.240,00.

Fonte: INEP, microdados ENEM 2020, 2021 e 2022

- Renda C: 1 se a resposta para a pergunta Q006 foi F, G ou H.
- Renda D: 1 se a resposta para a pergunta Q006 foi K, L ou M.
- Renda E: 1 se a resposta para a pergunta Q006 foi O, P ou Q.

3.3.10 Posse de computador pessoal

A questão Q024 pergunta sobre a posse de computador pessoal na residência do candidato. A variável original traz informação a respeito da quantidade de computadores na residência. Para o modelo optou-se por discriminar apenas a presença ou não de computador pessoal na residência do candidato agrupando as respostas sobre qualquer quantidade de computadores como 1 para a variável *Tem computador em casa* e zero para o caso de não possuir computador em casa.

3.3.11 Desempenho

As métricas de desempenho do candidato na avaliação são as variáveis endógenas do modelo de regressão, isto é, as variáveis dependentes que serão preditas pelo modelo. Optou-se por avaliar os modelos de regressão para o desempenho em cada prova individualmente (ciências humanas, ciências da natureza, linguagens e códigos e matemática) além da média das provas objetivas e média geral.

Tabela 6 – Relação da discretização por intervalo de valores e de quantis

Categoria	Intervalo de valores	Intervalo de quantis
0	de 0 a 199 pontos	de 0 a 462 pontos
1	de 200 a 399 pontos	de 463 a 511 pontos
2	de 400 a 599 pontos	de 512 a 558 pontos
3	de 600 a 799 pontos	de 559 a 621 pontos
4	de 800 a 1000 pontos	de 622 a 1000 pontos

- **Desempenho Ciências Humanas**: Desempenho exclusivamente na prova de ciências humanas.
- **Desempenho Ciências da Natureza**: Desempenho exclusivamente na prova de ciências da natureza.
- **Desempenho Linguagens e códigos**: Desempenho exclusivamente na prova de linguagens e códigos.
- **Desempenho Matemática**: Desempenho exclusivamente na prova de matemática.
- **Redação**: Desempenho exclusivamente na prova de redação.
- **Desempenho objetivas**: Média aritmética das provas objetivas de matemática, língua portuguesa, ciências humanas e ciências da natureza.
- **Desempenho geral**: Média aritmética das provas objetivas de matemática, língua portuguesa, ciências humanas, ciências da natureza e redação.

Para cada tipo de desempenho citado foi aplicado dois tipos de métodos de discretização, por intervalo de valores e de quantis. Em cada método a discretização foi feita de modo a se obter 5 categorias. A Tabela 3.3.11 mostra os intervalos de pontuação atribuída para cada categoria de desempenho.

3.4 Regressão logística e avaliação dos coeficientes

De posse dos dados tratados, agrupados e normalizados é feita, então, a regressão logística para cada variável endógena.

Os dados são inicialmente separados em dados de treino e teste para que se possa avaliar o modelo com dados que não contribuíram com o ajuste dos parâmetros. Para isso é utilizado o procedimento *train_test_split* da biblioteca *scikit learn*. Em cada ano avaliado 10% dos dados foram separados para testes do modelo.

A regressão foi feita se utilizando o modelo de regressão logística do *scikit learn* (PEDREGOSA et al., 2011). Um modelo foi ajustado para cada variável dependente em cada ano.

Dos modelos foram gerados pontuações que medem sua qualidade de predição e tabelas com os coeficientes de cada variável exógena colocada como entrada do modelo.

4 Resultados

Os algoritmos usados para gerar os resultados apresentados a seguir podem ser encontrados no repositório: <https://github.com/AntonioSN/Regressao_Logistica_ENEM.git>

4.1 Análise descritiva dos dados

Nos anos de 2020, 2021 e 2022 participaram do exame 5.783.109, 3.389.832 e 3.476.105 candidatos respectivamente totalizando 12.649.046 nos três anos. Os três anos apresentam uma presença maior de candidatas do sexo feminino como pode ser visto no Gráfico 9. Os anos de 2021 e 2022 apresentam menor número de candidatos prestando o exame em relação ao ano de 2020, entretanto 2020 é o ano que acumula o maior número de candidatos que não realizaram todas as provas, vide Gráfico 10. Observa-se também uma certa constância no número de alunos de até 18 anos, ou seja, a redução de candidatos em relação ao ano de 2020 por grupo etário se concentrou em candidatos com idade superior a 18 anos, além disso alunos que concluiriam o ensino médio no ano de aplicação tem maior tendencia de realizar todas as provas se comparado com alunos que já concluíram o ensino médio, que não cursam ou que irão concluir em ano posterior a aplicação da prova.

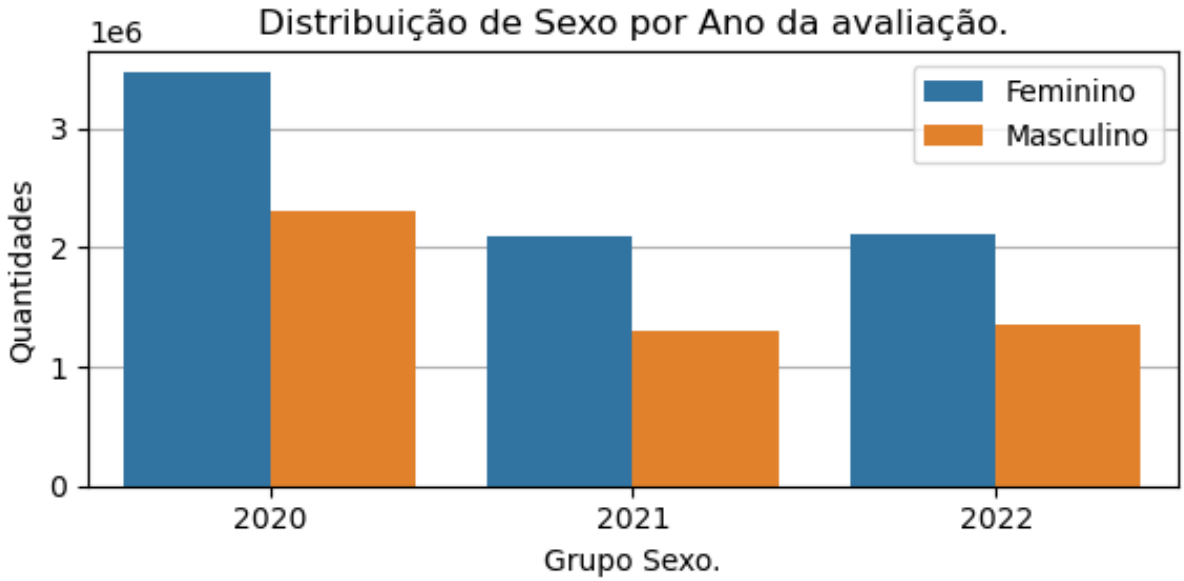
Sobre a situação de conclusão do ensino médio ilustrado no Gráfico 14 dos candidatos em cada ano observa-se pouca mudança em relação aos candidatos que concluem o ensino médio no ano de aplicação da avaliação ou posterior e uma grande redução de alunos que já concluíram o ensino médio. Dentre os alunos que responderam corretamente a escola onde concluíram ou concluirão o ensino médio a prevalência é de alunos da rede estadual e não há mudanças apreciáveis entre os anos de aplicação.

Tabela 7 – Relação de candidatos que realizaram todas as provas e situação do ensino médio

Situação do ensino médio	Percentagem	
	Realizou todas as provas	Não realizou todas as provas
Concluído no ano de aplicação	54,17%	45,83%
Concluirá após o ano da aplicação	49,27%	50,73%
Já concluído	40,02%	59,98%
Não concluiu e não cursa	43,15%	56,85%

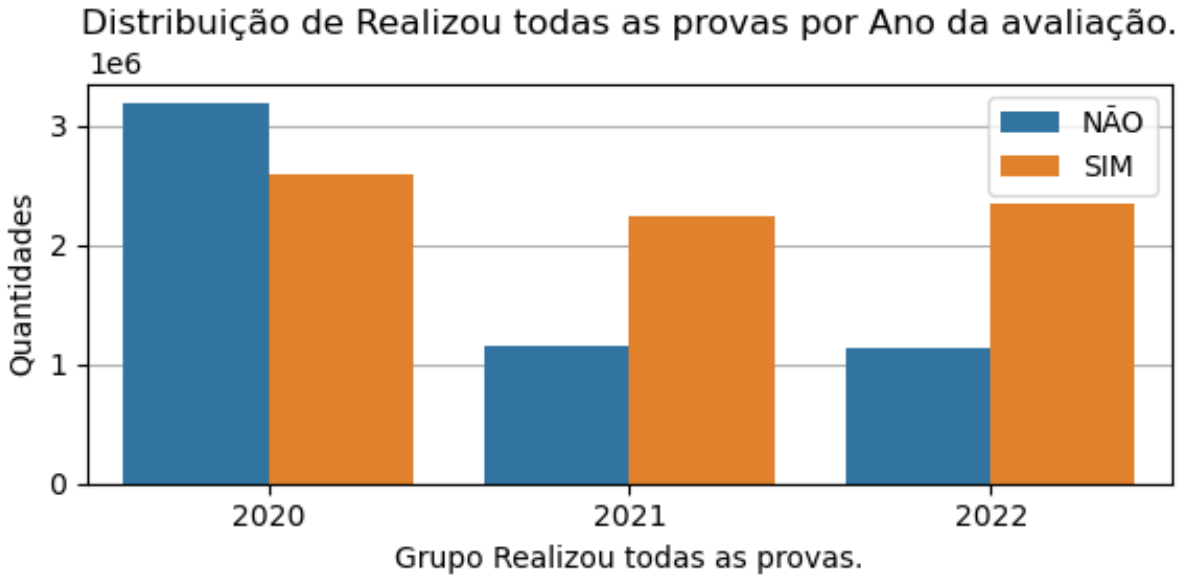
Observando a distribuição de "cor raça"no Gráfico 12 o número de candidatos que se declaram pardos diminuiu em relação aos demais nos anos de 2021 e 2022 se comparados

Figura 9 – Distribuição de sexo por ano da avaliação.



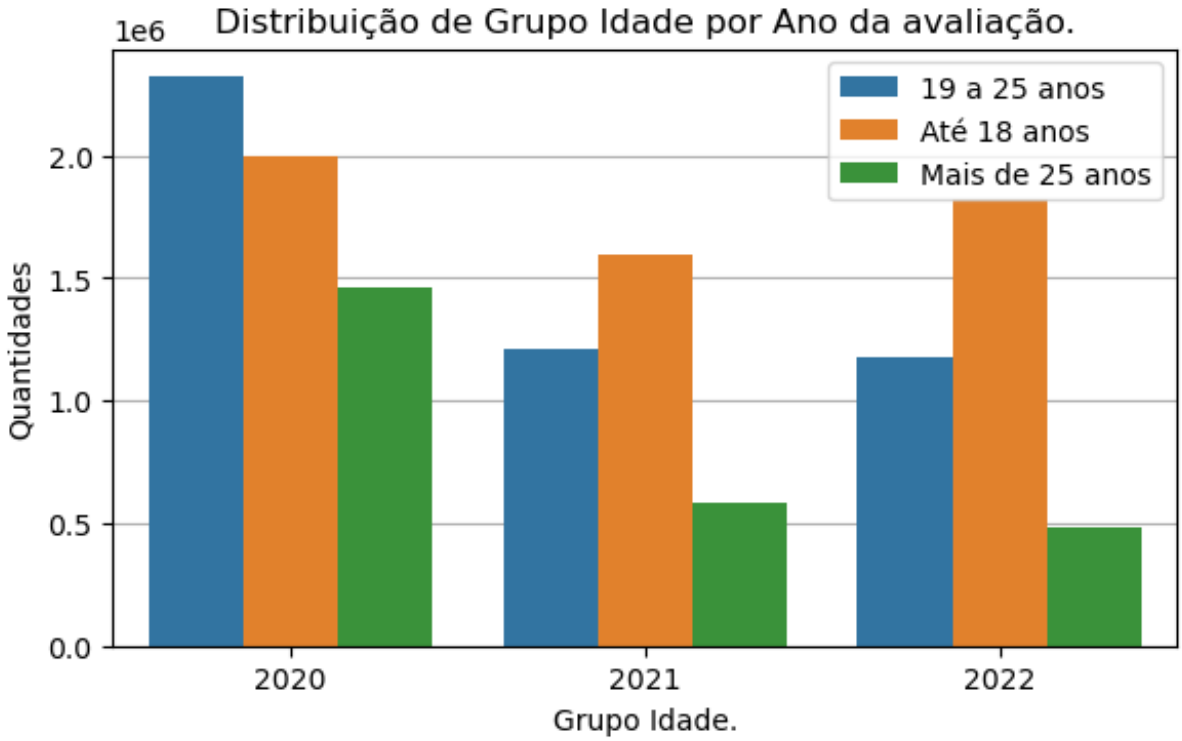
Fonte: Produção do próprio autor.

Figura 10 – Alunos que realizaram todas as provas no ano de aplicação.



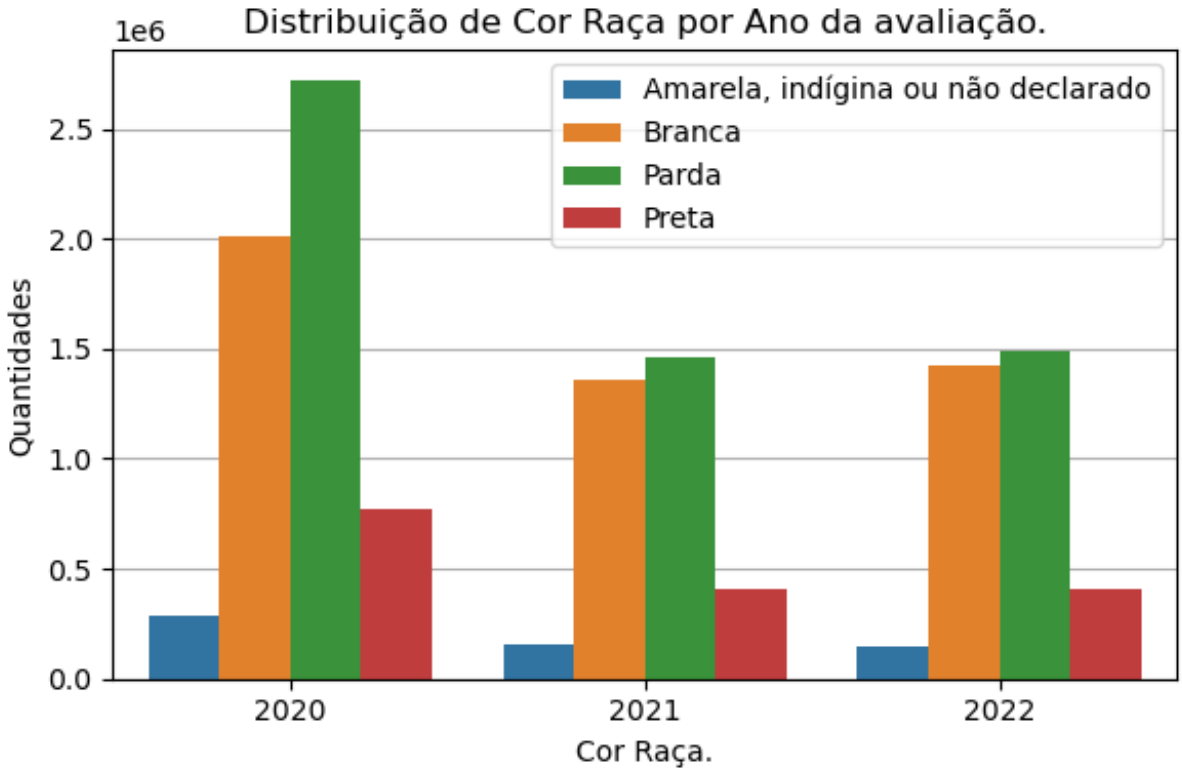
Fonte: Produção do próprio autor.

Figura 11 – Grupo etário por ano de aplicação.



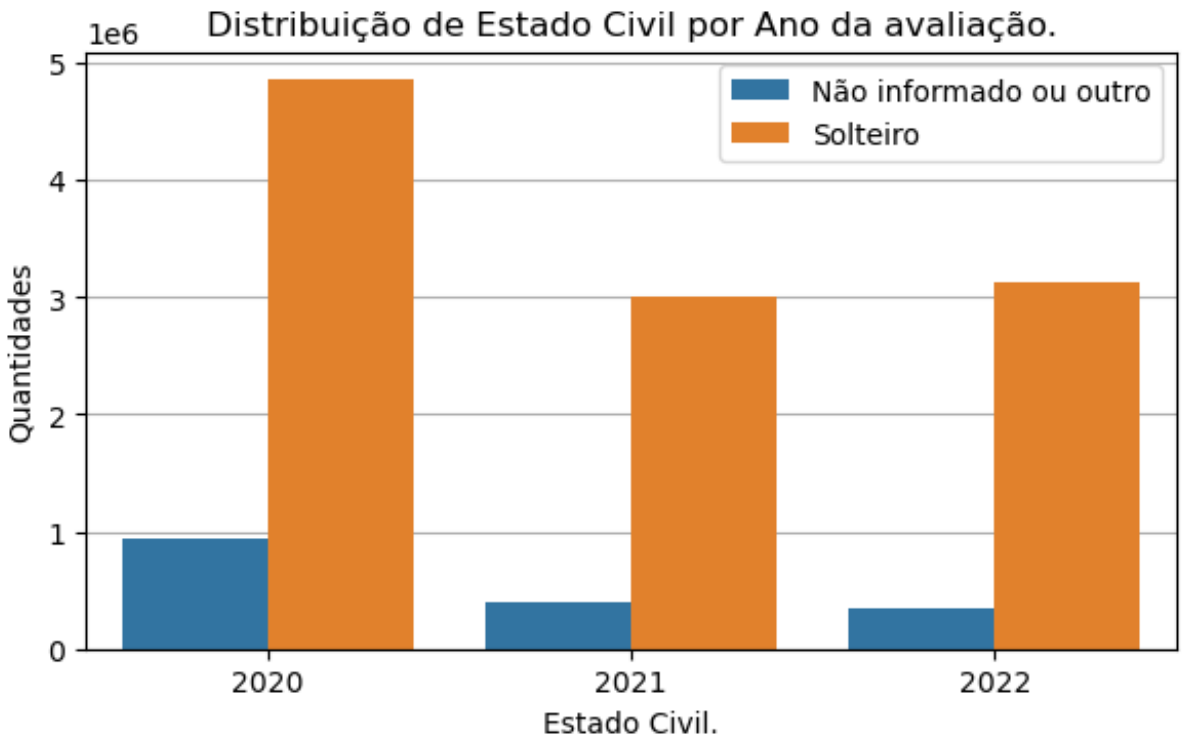
Fonte: Produção do próprio autor.

Figura 12 – Cor raça por ano de aplicação.



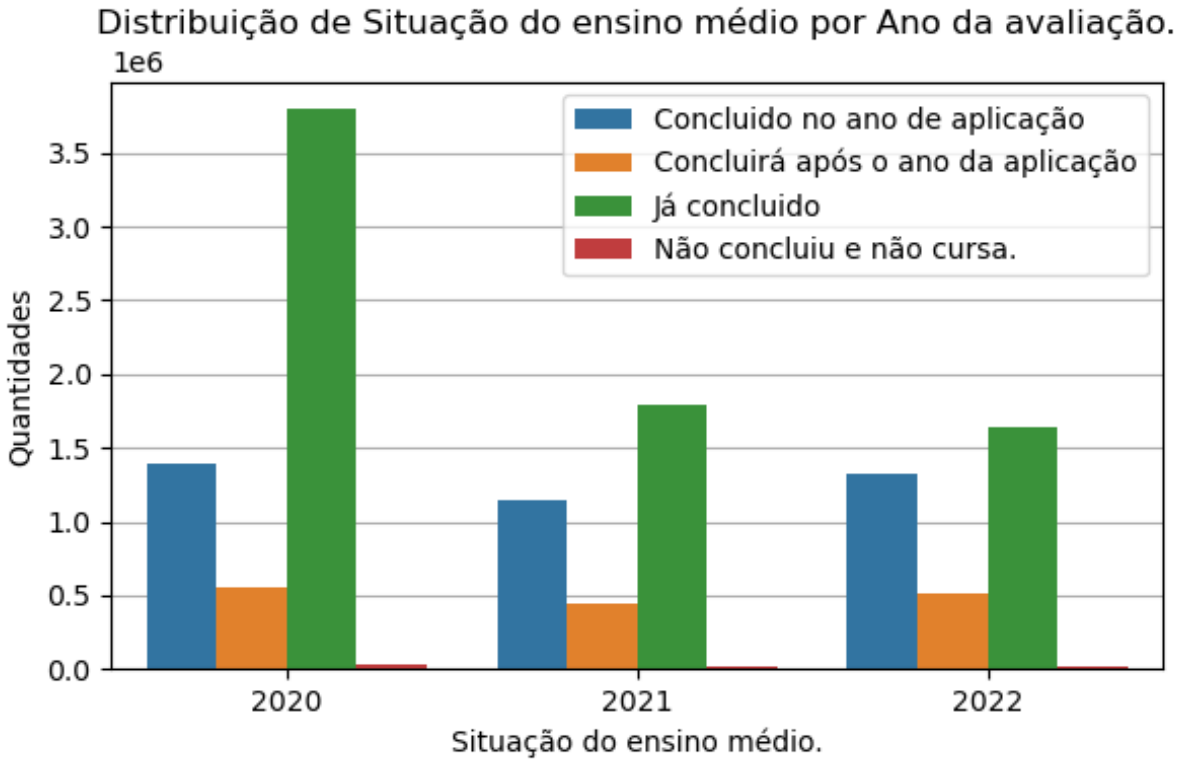
Fonte: Produção do próprio autor.

Figura 13 – Estado civil por ano de aplicação.



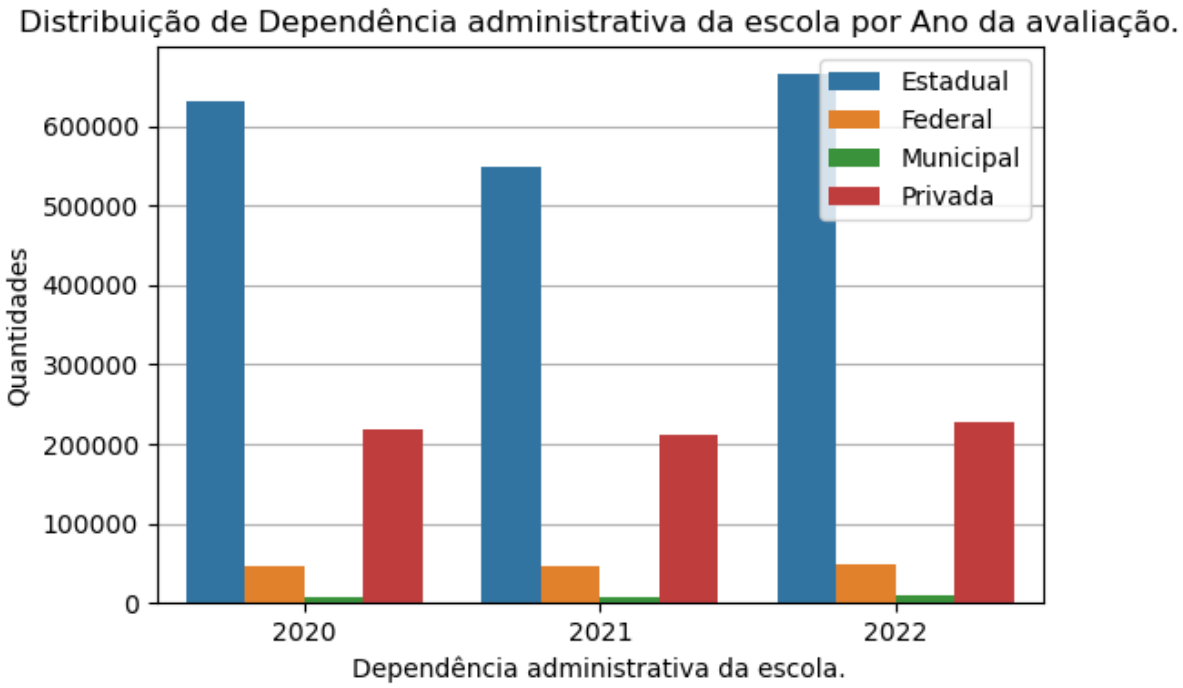
Fonte: Produção do próprio autor.

Figura 14 – Situação de conclusão do ensino médio por ano de aplicação.



Fonte: Produção do próprio autor.

Figura 15 – Dependência administrativa da escola por ano de aplicação.



Fonte: Produção do próprio autor.

	2020	2021	2022	2021 relativo a 2020	2022 relativo a 2020
Solteiro	4851310	2995915	3127949	-38,25%	-35,52%
Não solteiro	931799	393917	348156	-57,73%	-62,64%

a 2020, mas de maneira geral todas as categorias sofreram redução nos anos 2021 e 2022. Sobre o estado civil, vide Gráfico 13 dos candidatos em todos os anos a prevalência é de solteiros. A redução geral no número de inscritos nos anos de 2021 e 2022 em relação a 2020 aconteceu para as categorias de solteiro e de não solteiros, entretanto, em termos de queda percentual relativa, a redução de inscrição de candidatos não solteiros foi mais elevada como pode ser visto na Tabela 4.1.

Em relação a situação de conclusão do ensino médio mostrado na Figura 14 a redução de candidatos participantes nos anos de 2021 e 2022 em relação a 2020 aconteceu majoritariamente entre os candidatos que já concluíram o ensino médio em data anterior a da aplicação da prova. A Tabela 8 mostra os números absolutos e a variação relativa.

A distribuição do tipo de dependência administrativa da escola em que o aluno concluiu ou está cursando o ensino médio (federal, estadual, municipal ou privada), vide Figura 15 mudou pouco entre os três anos avaliados havendo uma prevalência das dependências administrativas federais.

A análise descritiva do desempenho dos candidatos foi feito com base no desempenho geral que é a média aritmética das quatro provas objetivas e redação dividido em cinco

Tabela 8 – Valores absolutos e variação percentual da situação de conclusão do ensino médio em cada ano.

	2020	2021	2022	2021 relativo a 2020	2022 relativo a 2020
Já concluído	3794279	1789372	1633253	-0,5284	-0,56955
Concluído no ano de aplicação	1395827	1150857	1317560	-0,1755	-0,05607
Concluirá após o ano da aplicação	557425	437190	512732	-0,2157	-0,08018
Não concluiu e não cursa.	35578	12413	12560	-0,6511	-0,64697

Categoria	nota miníma da categoria	nota máxima da categoria	Total pertencente a categoria	Percentual em relação ao total
Muito Ruim	0	167	166	0,002%
Ruim	168	335	5122	0,071%
Médio	336	503	3447853	48,076%
Bom	504	671	3476825	48,480%
Muito bom	672	1000	241645	3,369%

categorias de igual intervalo de pontos segundo descrito no Capítulo 3.3.11. Nessa análise chamamos as categorias de desempenho geral de *muito ruim*, *ruim*, *médio*, *bom* e *muito bom*, vide Tabela 4.1.

A maioria dos candidatos (cerca de 96%) se concentram nas categorias *médio* e *bom*.

O desempenho geral de cada aluno segundo a declaração de cor raça mostra que as melhores notas do exame se concentram em candidatos declarados brancos, além de ser a única categoria de cor que apresenta mais resultados "bom"do que no "médio"é a que tem maior número de resultados "muito bom", vide Tabela 4.1.

O desempenho por tipo de escola em que o candidato concluiu o ensino médio mostra a concentração dos melhores resultados em escolas privadas, essa categoria também é a única que apresenta mais resultados "Bom"do que "Médio". Escolas de dependência administrativa estadual são mais numerosas que os demais. Comparando percentualmente as dependências federais e estaduais as federais mostram um desempenho melhor concentrando relativamente mais resultados "Muito bom", vide Figura 4.1.

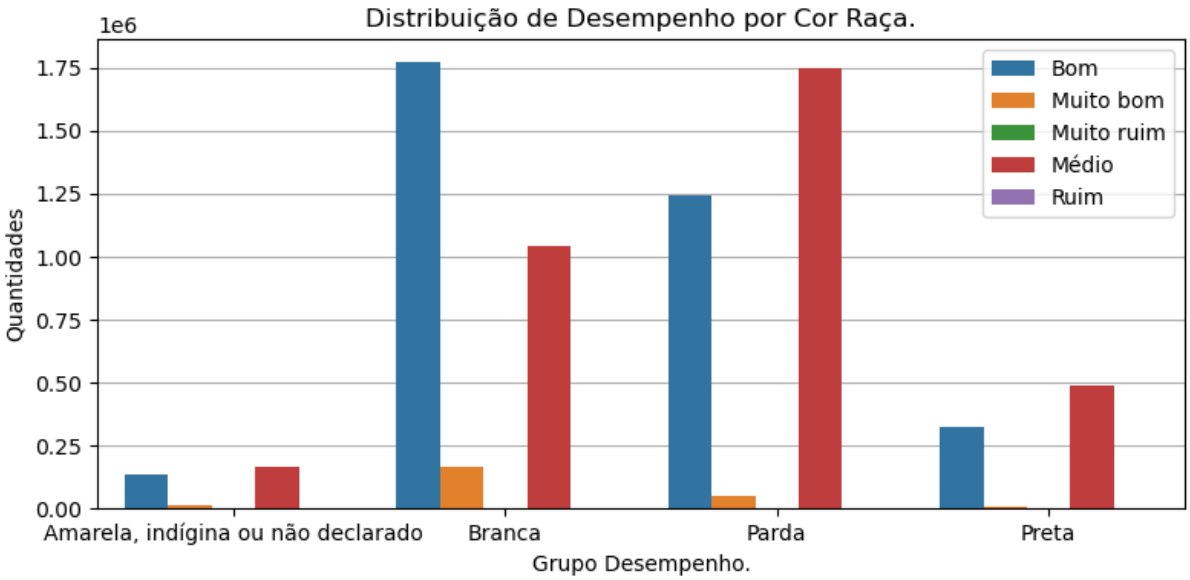
Por grupo etário o desempenho é melhor entre alunos de até 18 anos, vide Figura 4.1.

4.2 Resultado das regressões

As regressões fornecem os coeficientes que podem ser interpretados como pesos que o modelo deu para cada variável exógena. As regressões feitas de forma independente com os dados de cada ano mostraram consistência entre elas. Os coeficientes para as regressões feitas em cada ano se baseando nas categorias de desempenho definidas por quintis são apresentadas nas tabelas 9, 10 e 11

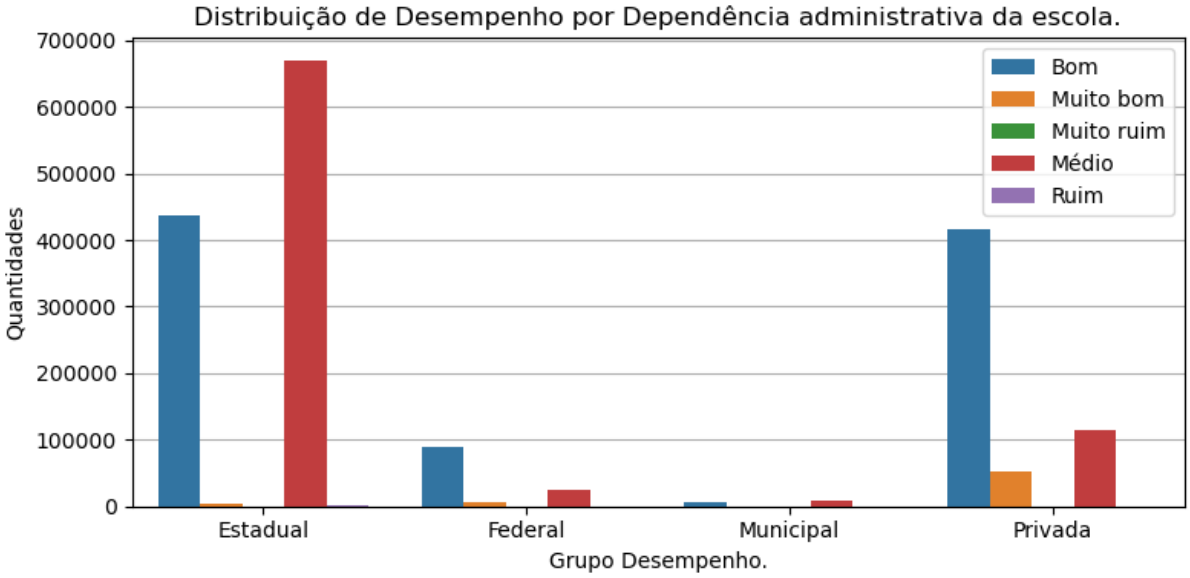
O coeficiente de maior peso na categoria 0 foi o número de pessoas que moram na mesma residência do candidato apresentando $2,1832$ para a regressão de 2020, $2,2264$

Figura 16 – Desempenho por cor raça.



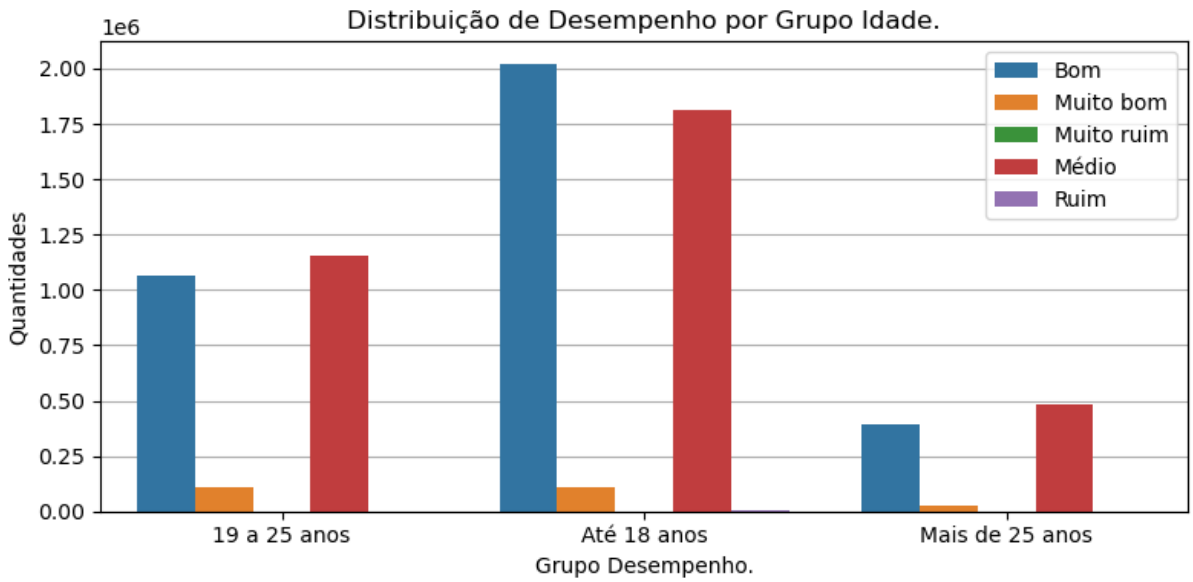
Fonte: Produção do próprio autor.

Figura 17 – Desempenho por escola.



Fonte: Produção do próprio autor.

Figura 18 – Desempenho por idade



Fonte: Produção do próprio autor.

para a de 2021 e 2, 2519 para 2020, ou seja, a quantidade de pessoas que moram na mesma residência que o candidato tem grande importância em determinar que esse candidato pertencerá ao primeiro quintil de desempenho geral no exame nacional do ensino médio. Em segundo lugar influenciando positivamente o candidato a pertencer ao primeiro quintil no desempenho geral está a idade, isso indica que alunos mais velhos evitam mais facilmente pertencer ao grupo de pior desempenho. Renda familiar de categoria A, que representa a renda mais baixa, também é um fator relevante mantendo um valor próximo de 0, 9 nas três regressões.

Interessante notar que os coeficiente mais relevantes da categoria 0 são também relevantes e inversos para a categoria 4, ou seja, se um número grande de pessoas que residem com o candidato aumentam as chances dele pertencer ao primeiro quintil também diminuem as chances de pertencer ao quinto quintil de desempenho.

Também de forma consistente para os três anos estudados o número de pessoas que residem na mesma casa do candidato mostrou ser, dentre as variáveis do modelo, o fator mais importante no sentido de aumentar as chances do aluno em pertencer a categoria 1, ou seja, o segundo quintil.

As variáveis *Renda A*, *Renda B*, *Escola Pública* e *Idade* alternam suas posições em cada regressão. De maneira curiosa a variável *Idade* apresentou grande relevância no modelo ajustado com os dados de 2020, ano que se observou uma maior participação de candidatos mais velhos. Das variáveis com maior peso no sentido de reduzir as chances do candidato a pertencer no grupo 1 e de forma consistente em todos os anos, aparece "*EM concluído*"e *"Renda E"*, ou seja, pertencer a uma categoria de renda familiar mais elevada

Tabela 9 – Coeficientes da regressão ajustada com dados de 2020 para o desempenho geral dividido por quantis.

	0	1	2	3	4
IDADE	1.6321	0.5543	-0.0991	-0.6781	-1.4092
MASCULINO	0.1623	0.1691	-0.0113	-0.1738	-0.1462
SOLTEIRO	-0.0374	-0.1179	-0.1122	-0.0287	0.2962
BRANCA	-0.2785	-0.0865	0.0838	0.1791	0.1021
PRETA	0.0719	0.0976	0.1009	0.0306	-0.3009
EM CONCLUIDO	-1.2840	-0.7092	-0.1842	0.3384	1.8390
EM NO ANO	-0.4440	-0.2512	-0.0263	0.1834	0.5381
EM APOS	-0.2565	-0.1862	-0.0158	0.1229	0.3356
Escola Pública	0.2908	0.2724	0.1417	-0.0148	-0.6900
Treineiro	-0.2565	-0.1862	-0.0158	0.1229	0.3356
Escolari. Pai A	0.3074	0.1328	-0.0025	-0.1523	-0.2855
Escolari. Pai B	0.0060	-0.0217	-0.0351	-0.0169	0.0677
Escolari. Pai C	-0.2187	-0.1563	-0.0834	0.0694	0.3891
Escolari. Mae A	0.2475	0.1555	0.0488	-0.0921	-0.3597
Escolari. Mae B	-0.1382	-0.0148	0.0654	0.1146	-0.0271
Escolari. Mae C	-0.2626	-0.1195	0.0011	0.1460	0.2349
Ocupação Pai A	-0.0569	-0.0225	0.0089	0.0284	0.0421
Ocupação Pai B	-0.4185	-0.2108	-0.0360	0.1562	0.5091
Ocupação Mae A	-0.0440	0.0076	0.0286	0.0257	-0.0179
Ocupação Mae B	-0.2302	-0.1056	-0.0040	0.0918	0.2479
N pessoas	2.1839	1.1926	0.2040	-1.0068	-2.5738
Renda A	0.8795	0.4991	0.0923	-0.3577	-1.1131
Renda B	0.3036	0.2566	0.1229	-0.0962	-0.5868
Renda C	-0.2219	-0.0355	0.0513	0.1272	0.0789
Renda D	-0.5388	-0.2813	-0.0229	0.2472	0.5959
Renda E	-0.5755	-0.4177	-0.1282	0.2481	0.8733
Tem computador	0.3435	0.1390	-0.0386	-0.1798	-0.2641

e já ter concluído o ensino médio em anos anteriores a aplicação da prova reduz a chances de pertencer ao segundo quintil de desempenho. Interessante notar que essas variáveis também reduzem as chances de pertencer ao primeiro quintil além de serem as variáveis que mais contribuem para o pertencimento do quarto e quinto quintil.

De maneira consistente nas três regressões os coeficientes que descrevem as chances do candidato de pertencer ao grupo 2 de desempenho (terceiro quintil) apresentam baixos módulos tendo a maioria módulos inferiores a $0,2$. Nesse grupo de desempenho nenhuma variável exógena se destacou muito, mas ter o ensino médio concluído antes da aplicação da prova se mostrou a variável de maior peso para as três regressões no sentido de reduzir as chances de pertencer a categoria 2. As variáveis *Solteiro*, *Renda E* e *Idade* alternam entre as variáveis mais negativamente influente influentes depois de *EM concluído*.

Sobre a regressão ao quarto quintil (categoria de desempenho 3) de desempenho

Tabela 10 – Coeficientes da regressão ajustada com dados de 2021 para o desempenho geral dividido por quantis.

	0	1	2	3	4
IDADE	1.1913	0.2199	-0.2264	-0.5212	-0.6636
MASCULINO	0.1189	0.1166	0.0191	-0.1175	-0.1372
SOLTEIRO	-0.1089	-0.1520	-0.1000	0.0208	0.3402
BRANCA	-0.3506	-0.0664	0.0964	0.1920	0.1286
PRETA	0.0172	0.1134	0.1116	0.0331	-0.2754
EM CONCLUIDO	-1.3816	-0.6705	-0.2621	0.3722	1.9420
EM NO ANO	-0.3324	-0.1415	-0.0531	0.1124	0.4146
EM APOS	-0.2672	-0.1283	-0.0523	0.0957	0.3521
Escola Pública	0.3931	0.2797	0.1974	-0.0530	-0.8172
Treineiro	-0.2672	-0.1283	-0.0523	0.0957	0.3521
Escolari. Pai A	0.2788	0.1034	0.0016	-0.1225	-0.2613
Escolari. Pai B	-0.0233	-0.0498	-0.0241	0.0090	0.0881
Escolari. Pai C	-0.2480	-0.1717	-0.0596	0.0982	0.3811
Escolari. Mae A	0.3432	0.1980	0.0625	-0.1270	-0.4768
Escolari. Mae B	-0.0702	0.0239	0.0754	0.0710	-0.1001
Escolari. Mae C	-0.2064	-0.0788	0.0103	0.1178	0.1572
Ocupação Pai A	-0.0564	-0.0065	0.0069	0.0188	0.0372
Ocupação Pai B	-0.3948	-0.1734	-0.0233	0.1400	0.4516
Ocupação Mae A	-0.0769	-0.0004	0.0165	0.0330	0.0280
Ocupação Mae B	-0.2375	-0.0856	-0.0042	0.0780	0.2493
N pessoas	2.2264	1.0291	0.0158	-1.0149	-2.2563
Renda A	0.9239	0.4669	0.0808	-0.3687	-1.1029
Renda B	0.3560	0.2527	0.1182	-0.1170	-0.6099
Renda C	-0.1316	-0.0220	0.0546	0.0933	0.0056
Renda D	-0.4573	-0.2085	-0.0183	0.2114	0.4727
Renda E	-0.6245	-0.3459	-0.0871	0.2428	0.8147
Tem computador	0.3434	0.1095	-0.0624	-0.1769	-0.2136

geral a variável mais influente foi o número de pessoas que moram junto com o candidato e influencia de forma inversa as chances de pertencer a esse grupo. Ter mais idade e pertencer ao grupo de renda familiar A (renda mais baixa) também reduz a chance de pertencer a esse quintil. Ter o ensino médio concluído foi a variável mais significativa no sentido de aumentar as chances do candidato em pertencer a essa categoria de desempenho seguido de renda familiar D e E (as categorias mais altas de renda familiar) e se declarar da cor branca.

Sobre as variáveis que influenciam o quinto quintil o número de pessoas que residem junto com o candidato é a variável de maior influência com um coeficiente de $-2,57$, $-2,25$ e $-2,27$ nos anos de 2020, 2021 e 2022 respectivamente. Entre as variáveis que mais influenciam as chances do candidato a pertencer esse grupo estão *EM concluído* e *Renda E*. De forma geral para os três anos avaliados pode-se esperar melhor desempenho

Tabela 11 – Coeficientes da regressão ajustada com dados de 2022 para o desempenho geral dividido por quantis.

	0	1	2	3	4
IDADE	1.2931	0.3354	-0.2290	-0.6010	-0.7985
MASCULINO	0.0565	0.0733	-0.0058	-0.0569	-0.0671
SOLTEIRO	-0.1340	-0.1572	-0.0965	0.0274	0.3603
BRANCA	-0.4113	-0.1099	0.1049	0.2276	0.1887
PRETA	-0.0249	0.0679	0.1123	0.0687	-0.2241
EM CONCLUIDO	-1.1840	-0.6259	-0.3262	0.2869	1.8492
EM NO ANO	-0.2994	-0.1193	-0.0566	0.1034	0.3719
EM APOS	-0.2397	-0.1279	-0.0789	0.0927	0.3538
Escola Pública	0.3355	0.3029	0.2370	-0.0584	-0.8170
Treineiro	-0.2397	-0.1279	-0.0789	0.0927	0.3538
Escolari. Pai A	0.2606	0.0873	-0.0345	-0.1096	-0.2039
Escolari. Pai B	-0.0529	-0.0573	-0.0471	0.0177	0.1396
Escolari. Pai C	-0.2776	-0.1890	-0.0825	0.0962	0.4529
Escolari. Mae A	0.3331	0.2133	0.0566	-0.1358	-0.4672
Escolari. Mae B	-0.0736	0.0501	0.0952	0.0629	-0.1346
Escolari. Mae C	-0.2233	-0.0797	0.0285	0.1179	0.1566
Ocupação Pai A	-0.0762	-0.0425	0.0171	0.0405	0.0610
Ocupação Pai B	-0.4179	-0.2220	-0.0219	0.1771	0.4847
Ocupação Mae A	-0.0854	-0.0269	0.0111	0.0408	0.0605
Ocupação Mae B	-0.2534	-0.1160	-0.0124	0.0935	0.2884
N pessoas	2.2519	1.0244	-0.0077	-0.9943	-2.2743
Renda A	0.9041	0.4807	0.0905	-0.3663	-1.1091
Renda B	0.3171	0.2642	0.1285	-0.1073	-0.6024
Renda C	-0.1668	-0.0124	0.0667	0.0966	0.0159
Renda D	-0.4546	-0.2105	-0.0080	0.2050	0.4682
Renda E	-0.5636	-0.3383	-0.0973	0.2171	0.7822
Tem computador	0.3957	0.1305	-0.0531	-0.1993	-0.2738

do candidato que reside com um número menor de pessoas, já concluiu o ensino médio, pertence a um grupo de idade maior que 18 anos e a um grupo de renda familiar elevado.

4.3 Análise das regressões

Como descrito no capítulo 3 vários modelos de regressão logística foram ajustados e testados. As tabelas apresentadas neste capítulo organizam os valores dos escores das regressões segmentado por prova. Em todos os casos a regressão logística se saiu melhor que os modelos simples. Em todas as provas os modelos ajustados tiveram pontuações semelhantes entre si mesmo quando testado com dados de outros anos o que demonstra uma robustez do modelo para prever resultados futuros. Os modelos ajustados com a discretização do desempenho com base em intervalo de valores da pontuação da prova mostraram um escore mais alto que a discretização por quantis, mas as hipóteses simples

Matemática	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.618	0.673	0.626	0.439	0.427	0.434
Regressão com dados de 2021	0.617	0.675	0.626	0.438	0.428	0.434
Regressão com dados de 2022	0.618	0.673	0.627	0.439	0.427	0.435
Modelo simples 1	0.581	0.635	0.584	0.131	0.133	0.126
Modelo simples 2	0.585	0.625	0.601	0.154	0.151	0.145

Linguagens e códigos	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.798	0.785	0.790	0.359	0.274	0.294
Regressão com dados de 2021	0.795	0.791	0.792	0.274	0.362	0.340
Regressão com dados de 2022	0.797	0.791	0.793	0.308	0.355	0.347
Modelo simples 1	0.791	0.789	0.789	0.252	0.228	0.240
Modelo simples 2	0.689	0.593	0.647	0.315	0.258	0.276

Ciências Humanas	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.692	0.689	0.764	0.367	0.359	0.296
Regressão com dados de 2021	0.691	0.694	0.769	0.362	0.366	0.300
Regressão com dados de 2022	0.690	0.693	0.770	0.324	0.320	0.326
Modelo simples 1	0.668	0.671	0.751	0.181	0.167	0.227
Modelo simples 2	0.628	0.600	0.666	0.215	0.211	0.259

também mostram valores elevados de escore. Observando as regressões feitas com a discretização de desempenho por quantis entre as provas objetivas o modelo de regressão que se saiu melhor em prever as notas foram a de matemática e ciências da natureza em qualquer ano. Os modelos de regressão para a prova de redação também apresenta uma boa nota, apesar de 2020 ter se mostrado um ano particularmente mais difícil de se prever o desempenho que os demais. Curioso notar que essa pontuação mais baixa nos modelos de regressão para redação ocorrem com os dados de 2020 para modelos ajustados em quaisquer anos. Além disso o modelo para redação ajustado com os dados de 2020 também prevê melhor os dados de 2021 e 2022 corroborando com a existência de uma dificuldade particular de predição nesse ano que foi, dentre os três, o de participação etária mais variada.

Ciências da Natureza	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.774	0.781	0.849	0.395	0.403	0.411
Regressão com dados de 2021	0.774	0.782	0.850	0.384	0.416	0.432
Regressão com dados de 2032	0.774	0.781	0.850	0.377	0.414	0.434
Modelo simples 1	0.770	0.777	0.848	0.187	0.168	0.162
Modelo simples 2	0.629	0.588	0.659	0.222	0.187	0.169

Redação	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.446	0.484	0.445	0.463	0.500	0.555
Regressão com dados de 2021	0.445	0.486	0.447	0.437	0.525	0.580
Regressão com dados de 2022	0.430	0.471	0.462	0.440	0.524	0.583
Modelo simples 1	0.400	0.423	0.354	0.130	0.096	0.141
Modelo simples 2	0.422	0.461	0.417	0.147	0.132	0.142

Média das objetivas	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.816	0.817	0.835	0.355	0.348	0.345
Regressão com dados de 2021	0.815	0.822	0.838	0.335	0.364	0.349
Regressão com dados de 2022	0.815	0.822	0.838	0.339	0.357	0.360
Modelo simples 1	0.798	0.808	0.825	0.209	0.191	0.211
Modelo simples 2	0.716	0.669	0.709	0.242	0.214	0.228

Média geral	Intermalo de valores			Intervalo de quantis		
	2020	2021	2022	2020	2021	2022
Regressão com dados de 2020	0.773	0.772	0.764	0.347	0.337	0.341
Regressão com dados de 2021	0.773	0.776	0.763	0.343	0.343	0.342
Regressão com dados de 2022	0.773	0.772	0.765	0.344	0.340	0.344
Modelo simples 1	0.738	0.740	0.717	0.203	0.205	0.208
Modelo simples 2	0.715	0.690	0.712	0.231	0.234	0.237

4.4 Discussão dos Resultados

Os resultados obtidos nas análises dos modelos de regressão logística proporcionam insights valiosos sobre a capacidade preditiva em relação ao desempenho nas diferentes provas. Conforme discutido no Capítulo 3, foram realizados diversos ajustes e testes, destacando-se a superioridade dos modelos de regressão logística em comparação com abordagens mais simples.

Ao analisar as tabelas apresentadas neste capítulo, organizadas de acordo com as diferentes provas, fica evidente que, em todos os casos, os modelos ajustados superaram os modelos simples. As pontuações dos modelos ajustados foram consistentemente semelhantes entre si, mesmo quando submetidos a dados de anos diferentes, revelando a robustez do modelo na previsão de resultados futuros.

Uma abordagem interessante consistiu na discretização do desempenho com base em intervalos de valores da pontuação da prova. Nesse contexto, os modelos apresentaram escores mais elevados em comparação com a discretização por quantis. No entanto, é importante ressaltar que as hipóteses simples também demonstraram valores elevados de escore, indicando a relevância desse método na análise.

Ao observar as regressões feitas com a discretização por quantis entre as provas objetivas, destaca-se o desempenho superior do modelo de regressão na previsão das notas

de Matemática e Ciências da Natureza em todos os anos. Os modelos de regressão para a prova de redação também apresentaram bons resultados, embora o ano de 2020 tenha se mostrado particularmente desafiador para prever o desempenho, em comparação com os demais anos.

Curiosamente, a pontuação mais baixa nos modelos de regressão para redação ocorreu especificamente com os dados de 2020, independentemente do ano de ajuste. Além disso, o modelo para redação ajustado com os dados desse ano demonstrou uma capacidade superior na previsão dos dados de 2021 e 2022, corroborando a existência de uma dificuldade única de predição nesse ano, que se destacou por apresentar uma participação etária mais variada em comparação aos anos anteriores.

Esses resultados contribuem para a compreensão da eficácia dos modelos de regressão logística na previsão de desempenho em diferentes provas, permitindo insights valiosos para otimização e aprimoramento de estratégias educacionais.

5 Conclusão

O presente trabalho teve como objetivo explorar e analisar a eficácia de modelos de regressão logística na predição do desempenho em diferentes provas de avaliação. A partir dos resultados obtidos e das análises realizadas, algumas conclusões importantes podem ser destacadas.

A utilização de modelos de regressão logística mostrou-se uma abordagem robusta e eficaz para prever o desempenho dos estudantes nas avaliações. A superioridade desses modelos em relação a abordagens mais simples reforça a importância de considerar a complexidade das relações entre as variáveis envolvidas.

A discretização do desempenho com base em intervalos de valores da pontuação da prova proporcionou resultados animadores em um primeiro momento, com escores mais elevados em comparação com a discretização por quantis, no entanto, é necessário ponderar sobre as características específicas de cada prova, uma vez que a eficácia do modelo discretizado por intervalo de valores não se mostrou muito melhor que o modelo que faz sua previsão se baseando simplesmente na frequência de cada categoria. Na verdade isso diz mais sobre a forma de se realizar a discretização dos valores do que o modelo de regressão propriamente dito.

Destaca-se a capacidade notável do modelo de regressão logística na previsão das notas de Matemática e Ciências da Natureza em todos os anos analisados. A prova de redação também apresentou resultados positivos, embora o ano de 2020 tenha se mostrado um desafio singular na predição do desempenho, evidenciando a importância de considerar contextos particulares.

Em síntese, os resultados obtidos contribuem para o avanço do entendimento sobre a aplicação de modelos de regressão logística na predição do desempenho educacional. Sugere-se que futuros trabalhos explorem aspectos mais detalhados das características específicas das provas e considerem a inclusão de outras variáveis relevantes para aprimorar ainda mais a precisão dos modelos.

Espera-se que as descobertas deste trabalho forneçam insights valiosos para educadores, pesquisadores e profissionais envolvidos no desenvolvimento de estratégias educacionais, contribuindo para uma abordagem mais eficaz na compreensão e previsão do desempenho dos estudantes.

Referências

AGRESTI, A. *Categorical Data Analysis*. 2°. ed. [S.l.]: Wiley Interscience, 2002. ISBN 0-471-36093-7. Citado na página 25.

ALPAYDIN, E. *Introduction to machine learning*. [S.l.]: MIT press, 2020. Citado na página 14.

AMéRICO, B. L.; LACRUZ, A. J. *Contexto e desempenho escolar: Análise das notas na Prova Brasil das escolas capixabas por meio de regressão linear múltipla*. [S.l.]: Fundacao Getulio Vargas, 2017. 854-878 p. Citado na página 14.

ANDERSEN, E. B. *Introduction to the Statistical Analysis of Categorical Data*. [S.l.]: Springer, 1997. ISBN 978-3-540-62399-1. Citado na página 33.

ARAUJO, E. A. C. d.; ANDRADE, D. F. d.; BORTOLOTTI, S. L. V. Teoria da resposta ao item. *Revista da Escola de Enfermagem da USP*, SciELO Brasil, v. 43, p. 1000–1008, 2009. Citado na página 13.

ASSUNçãO, M. V. D. de; ARAúJO, A. G. de; ALMEIDA, M. R. de. The influence of family background on the access to technical and vocational education. *Revista de Administracao Publica*, Fundacao Getulio Vargas, v. 53, p. 542–559, 5 2019. ISSN 19823134. Citado na página 14.

BESSA, J. D. T. S. *A IMPORTÂNCIA DO RANKING DO ENEM PARA A SOCIEDADE E AS DIFERENÇAS ENTRE AS REDES DE ENSINOMOSSORÓ –RN2016*. 2016. Citado na página 14.

CHEIN, F. *Introdução aos modelos de regressão linear Metodologias COLEÇÃO*. [S.l.]: Enap, 2019. ISBN 978-85-256-0115-5. Citado na página 18.

DE, A.; MOREIRA, O.; DAS, C. *ESTUDO SOBRE REGRESSÃO LINEAR E REGRESSÃO LOGÍSTICA*. 2021. Citado 4 vezes nas páginas 13, 16, 25 e 33.

DEVORE, J. L. *PROBABILIDADE E ESTATISTICA PARA ENGENHARIA E CIENCIAS*. [S.l.]: Cengage Learning Edicoes Ltda, 2010. 704 p. ISBN 9788522109241. Citado na página 33.

DODGE, Y. *The Concise Encyclopedia of Statistics*. [S.l.: s.n.], 2008. ISBN 978-0-387-32833-1. Citado 2 vezes nas páginas 30 e 37.

ECCLE, C. Correlação entre horas de estudo e desempenho em avaliações. 2002. Citado na página 16.

ESPIRITU, F. D.; FILHO, S. *Teoria da resposta ao item: influência do tamanho da amostra na estimação dos parâmetros dos itens utilizando os microdados do Enem*. 2020. Citado na página 13.

FARIA, L. et al. *Ciência de dados: algoritmos e aplicações*. [s.n.], 2021. ISBN 9786589124474. Disponível em: <www.impa.br>. Citado 2 vezes nas páginas 13 e 14.

FERREIRA, E. A. *TEORIA DE RESPOSTA AO ITEM - TRI, Análise de algumas questões do ENEM - habilidades 24 a 30*. 2018. Citado 2 vezes nas páginas 13 e 26.

FIGUEIRA, C. V. *MODELOS DE REGRESSÃO LOGÍSTICA*. 2006. Citado 2 vezes nas páginas 18 e 33.

FONSECA, S. O. d.; NAMEN, A. A. Mineração em bases de dados do inep: uma análise exploratória para nortear melhorias no sistema educacional brasileiro. *Educação em Revista*, SciELO Brasil, v. 32, p. 133–157, 2016. Citado na página 39.

GARCIA, J. L. *Busca de novos indicadores de aprendizado em Matemática através da comparação entre as bases de dados do Enem 2017 e Saeb 2017*. 2019. Citado na página 26.

GÉRON, A. Hands-on machine learning with scikit-learn. *Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*, O'Reilly Media, v. 1, 2019. Citado 3 vezes nas páginas 33, 36 e 37.

HAIR, J. J. F.; BLACK, W. C.; SANT'ANNA, A. S. *Analise multivariada de dados (6a. ed.)*. [S.l.]: Grupo A - Bookman, 2000. 689 p. ISBN 9788577805341. Citado na página 16.

HAIR, R. Correlação entre variáveis e causalidade. 2009. Citado 4 vezes nas páginas 16, 17, 23 e 43.

HOSSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. Segunda. [S.l.: s.n.], 2000. Citado 5 vezes nas páginas 13, 14, 25, 32 e 33.

IPEA. Políticas sociais: acompanhamento e análise. *2019*, 2019. Citado 2 vezes nas páginas 15 e 26.

JúNIA, E.; MARIZ, B. *Aplicação do modelo de Regressão Logística em Dados de Rendimento do Exame Nacional do Ensino Médio*. 2021. Citado 3 vezes nas páginas 13, 33 e 44.

KARINO, C. A.; SOUSA, E. C. *Guia do participante ENEM*. 2012. Citado na página 16.

KLEIN, R. Utilização da teoria de resposta ao item no sistema nacional de avaliação da educação básica (saeb)*. *2003*. Citado na página 16.

MAROCO, J. *Análise Estatística Com utilização do SPSS*. 3°. ed. [S.l.]: Edições Sílabo, 2007. Citado na página 29.

MCKINNEY, W. Data structures for statistical computing in python. In: *Proceedings of the 9th Python in Science Conference*. [S.l.: s.n.], 2010. p. 51–56. Citado na página 41.

MEYER, D. Modelagem matemática de variáveis correlacionadas. 2017. Citado na página 16.

MISES, R. V. *Probability, Statistics and Truth*. [S.l.]: Dover Publications, 1957. Citado na página 17.

MONTGOMERY, S. Tipos de regressão para diferentes relações entre variáveis. 2012. Citado na página 16.

MORETTIN, P. A.; BUSSAB, W. O. *Estatística básica*. [S.l.]: Saraiva Educação SA, 2017. Citado 2 vezes nas páginas 16 e 33.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, Oxford University Press, v. 135, n. 3, p. 370–384, 1972. Citado na página 24.

NOCEDAL, J.; WRIGHT, S. *Numerical optimization*. [S.l.]: Springer Science Business Media, 2006. Citado na página 32.

PAPERT, S. An exploration in the space of mathematics educations. *International Journal of Computers for Mathematical Learning*, Springer, v. 1, n. 1, p. 95–123, 1996. Citado na página 41.

PASQUALI, L.; PRIMI, R. Fundamentos da teoria da resposta ao item: Tri. *Avaliação Psicológica: Interamerican Journal of Psychological Assessment*, Instituto Brasileiro de Avaliação Psicológica (IBAP), v. 2, n. 2, p. 99–110, 2003. Citado na página 13.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, v. 12, p. 2825–2830, 2011. Citado 3 vezes nas páginas 25, 37 e 50.

PEREIRA, L. Relação bidirecional entre variáveis dependentes e independentes na regressão. 2009. Citado na página 17.

PIRES, A. Renda familiar e escolaridade dos pais: ReflexÕes a partir dos microdados do enem 2012 do estado de sÃo paulo. *ETD*, v. 17, p. 523–541, 2015. Citado na página 14.

RODRIGUES, R. L.; MEDEIROS, F. P. D.; GOMES, A. S. Modelo de regressão linear aplicado à previsão de desempenho de estudantes em ambiente de aprendizagem. In: *Brazilian symposium on computers in education (simpósio brasileiro de informática na educação-sbie)*. [S.l.: s.n.], 2013. v. 24, n. 1, p. 607. Citado na página 13.

SAVIANI, D. O inep, o diagnóstico da educação brasileira e a rbep. *Revista brasileira de estudos pedagógicos*, INEP, v. 93, n. 234, p. 291–322, 2012. Citado na página 39.

SCHRAUDOLPH, N. N.; YU, J.; GüNTER, S. A stochastic quasi-newton method for online convex optimization. In: *Artificial Intelligence and Statistics*. [S.l.: s.n.], 2007. Citado na página 32.

SEABOLD, S.; PERKTOLD, J. Statsmodels: Econometric and statistical modeling with python. In: AUSTIN, TX. *Proceedings of the 9th Python in Science Conference*. [S.l.], 2010. v. 57, n. 61, p. 10–25080. Citado na página 41.

SILVEIRA, F. L. d.; BARBOSA, M. C. B.; SILVA, R. d. *Exame Nacional do Ensino Médio (ENEM): uma análise crítica*. [S.l.]: SciELO Brasil, 2015. 1101 p. Citado na página 13.

TADEU, S.; COSTA, S. *Teoria de Resposta ao Item aplicada no ENEM*. 2017. Citado na página 13.

TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados-da teoria à prática. *Sociedade Portuguesa de Estatística, Lisboa*, 2000. Citado na página 24.

VAART, A. W. Van der. *Asymptotic statistics*. [S.l.]: Cambridge university press, 2000. Citado na página 41.

VIGGIANO, E.; MATTOS, C. O desempenho de estudantes no enem 2010 em diferentes regiões brasileiras. *RBEP*, p. 417–438, 2009. Citado na página 14.

APÊNDICE A – Glossário de termos da estatística

Coeficiente de correlação (correlação de Bravais-Pearson) Coeficiente que mede o quão fortemente duas variáveis aleatórias estão relacionadas linearmente. O coeficiente de correlação é um valor numérico no intervalo [-1;1] onde os extremos do intervalo representam uma correspondência perfeita (positiva ou negativa) entre as variáveis e zero significa que as variáveis não se relacionam linearmente.

Variável aleatória Uma variável cujo valor é determinado por um experimento aleatório.

Experimento aleatório Operação de obtenção de dados conduzida sobre determinada condição de controle onde o resultado não pode ser previsto com antecedência. Por exemplo, entrevistar pessoas de uma região perguntando-lhes suas idades. Não há como saber de antemão a idade da próxima pessoa que será entrevistada.

Variável Característica mensurável que pode ser atribuída aos elementos da amostra. Por exemplo, idade é uma variável que pode ser atribuída a um conjunto de pessoas de uma certa região. Uma variável pode ser quantitativa, como idade ou altura, mas também pode ser qualitativa, como cor dos olhos ou gênero.

Amostra Subconjunto de uma população. Por exemplo, um estudo que coleta dados de 100 pessoas que moram em um bairro onde residem 1.000 pessoas tem uma amostra de 100 em uma população de 1.000 pessoas.

População Coleção das unidades estatísticas de interesse em um dado estudo estatístico.

Variável dependente Também conhecida como variável resposta ou variável endógena. Em um modelo de regressão é a variável que supomos ser dependente de outra variável.

Variável independente É a variável que supomos influenciar a variável dependente. Por exemplo é muito comum descrever sistemas mecânicos através de sua função no tempo, nesses casos o tempo é a variável independente e a posição do sistema é a variável dependente do tempo.

Modelo Formulação abstrata que pode descrever parte do comportamento de uma situação real. Em estatística um modelo pode ser uma função matemática que relaciona a variável dependente com a variável independente.

Regressão O ato de ajustar parâmetros de um modelo baseado nos dados observados. Existem várias técnicas de regressão e a mais comum é a regressão linear.

De acordo com Angrist e Pischke (2009), os modelos de regressão podem ser vistos como um dispositivo computacional para estimação de diferenças entre um grupo de tratados e um grupo de controle, com ou sem covariadas.

Grandezas Característica que pode ser medida. Pode ser quantitativa (massa, temperatura) ou qualitativa (gênero, espécie).

Espaço Amostral Conjunto de todos os resultados possíveis em um experimento aleatório.

Quantil O quantil é uma medida estatística que divide um conjunto de dados ordenados em partes iguais, expressas em termos percentuais. Em outras palavras, os quantis representam pontos de corte que dividem uma distribuição em porcentagens específicas.

O quantil mais comum é o quartil, que divide os dados em quatro partes iguais. Outros quantis também são usados, como os quintis (dividindo os dados em cinco partes iguais) e os percentis (dividindo em cem partes iguais).