Universidade de Brasília
Departamento de Estatística

Estudo sobre a Evasão Acadêmica no curso de Bacharelado em Ciência da Computação da Universidade de Brasília

Uma aplicação a Modelos de Regressão Logística

Ana Carolina Gomez Valenzuela Vianna

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Ana Carolina Gomez Valenzuela Vianna

Estudo sobre a Evasão Acadêmica no curso de Bacharelado em Ciência da Computação da Universidade de Brasília

Uma aplicação a modelos de Regressão Logística

Orientadora: Profa. Maria Teresa Leão Costa

Relatório final apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2023

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais, Claudia e Fernando, por todo amor, cuidado, suporte e por sempre me proporcionarem a melhor educação possível. Aos meus irmãos, Heloísa e Arthur, que sempre estiveram comigo me apoiando em todos os momentos. Sem vocês nada disso seria possível.

Estendo meu agradecimento a toda a minha família - avós, avôs, tias, tios, madrinhas - que torcem e vibram por mim em todos os momentos.

Agradeço a todos os meus amigos de curso, com os quais tive o privilégio de compartilhar experiências e aprender um pouco com cada um. Gostaria de destacar com carinho a minha dupla, Beatriz e Daniel, que estiveram ao meu lado diariamente, em todos os momentos de alegria, tristeza e conquistas.

À minha orientadora, Professora Maria Teresa Leão Costa, expresso minha gratidão por todo o apoio, compartilhamento de conhecimento e pela oportunidade de me orientar na conclusão deste trabalho.

Por fim, agradeço a todos que estiveram ao meu lado durante essa jornada. Cada contribuição, palavra de incentivo e apoio foram fundamentais para o meu crescimento pessoal e acadêmico.

Resumo

A evasão acadêmica representa um dos grandes desafios enfrentados pelas instituições de ensino superior, tanto no Brasil quanto no restante do mundo. No contexto deste trabalho, o termo evasão refere-se ao fenômeno em que um estudante deixa o curso no qual se matriculou, de maneira diferente da formatura.

O estudo em questão trata da análise de dados sobre a evasão acadêmica especificamente no curso de Bacharelado em Ciência da Computação da Universidade de Brasília, cuja amostra foi de 764 discentes. O conjunto de dados é composto com as características dos alunos que se matricularam no curso durante o período de 2012/1 a 2019/2.

Com os dados coletados, foi utilizado o modelo de Regressão Logística e o *software* R para desenvolver dois modelos independentes destinados a descrever o fenômeno da evasão. Esses modelos foram construídos com base em variáveis associadas ao desempenho dos alunos, tais como o Índice de Rendimento Acadêmico (IRA) e a Taxa de Reprovação. Dentre os fatores considerados, destacaram-se o currículo vigente no ingresso do aluno e a participação em semestres de verão. Ambos os modelos demonstraram um ajuste satisfatório, com uma especificidade superior a 80%.

Palavras-chaves: Evasão, Ensino Superior, Ciência da Computação, Regressão logística.

Lista de Tabelas

1 Formas de saída do curso de Bacharelado em Ciência da Computação - UnB 35

2 Distribuição dos alunos segundo Formas de Ingresso do curso de Bacharelado em Ciência da Computação - UnB, 2012-2019. 37

3 Distribuição dos alunos segundo Formas de Ingresso do curso de Bacharelado em Ciência da Computação - UnB, 2012-2019. (Forma Agrupada) . . 37

4 Resultado da disciplina 37

5 Regiões administrativas segundo nível de renda, Distrito Federal - 2018 . . 40

6 Distribuição dos alunos por Região Administrativa. Bacharelado em Ciência da Computação-UnB, 2012-2019. 42

7 Distribuição dos alunos residentes no Goiás por Cidade. Bacharelado em Ciência da Computação-UnB, 2012-2019. 42

8 Distribuição dos alunos segundo Quantidade de Ingressos. Bacharelado em Ciência da Computação-UnB, 2012-2019. 48

9 Distribuição dos alunos segundo Formas de saída do curso de Bacharelado em Ciência da Computação-UnB, 2012-2019 49

10 Análise bivariada por evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019. 50

11 Associação das variáveis com evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019 54

12 Distribuição dos alunos segundo Sistema de Cotas e Tipo de Escola. Bacharelado em Ciência da Computação-UnB, 2012-2019. 56

13 Teste inicial com a base de construção para todas variáveis do modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019 57

14 Teste inicial com a base de construção para todas variáveis do modelo com Taxa de Reprovação MAT. Bacharelado em Ciência da Computação-UnB, 2012-2019 58

15 Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019 59

16 Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019 59

17 Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019 60

18 Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019. 60

19 Razão de chance e IC de 95% para o modelo IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019. 61

20 Razão de chance e IC de 95% para o modelo Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019. 62

21 Teste de adequabilidade para o modelo IRA. 63

22 Teste de adequabilidade para o modelo IRA. 65

Lista de Figuras

1 Exemplo da curva da função de regressão logística Fonte: https://goo.gl/nwec4Q 12

2 Exemplo do gráfico AIC_p Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 585). 20

3 Exemplo do gráfico BIC_p Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 585). 21

4 Exemplo do gráfico de resíduos Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 595). 29

5 Exemplo do gráfico da curva ROC Fonte: Agresti, An Introduction to Categorical Data Analysis (p. 112). 33

6 Distribuição dos alunos segundo Gênero (a) e Idade (b). Bacharelado em Ciência da Computação-UnB, 2012-2019. 41

7 Distribuição dos alunos segundo Nível de Renda. Bacharelado em Ciência da Computação-UnB, 2012-2019. 43

8 Distribuição dos alunos segundo Período de Ingresso (a), Escola (b) e Forma de Ingresso (c). Bacharelado em Computação-UnB, 2012-2019. 44

9 Distribuição dos alunos segundo Sistema de Cotas (a) e Tipo de Cota (b). Bacharelado em Ciência da Computação-UnB, 2012-2019. 44

10 Distribuição dos alunos segundo Integralização (a) e IRA (b). Bacharelado em Ciência da Computação-UnB, 2012-2019. 45

11 Distribuição dos alunos segundo Taxa de Reprovação - Geral (a), Ciência da Computação (b) e Matemática (c). Bacharelado em Ciência da Computação-UnB, 2012-2019. 46

12 Distribuição dos alunos segundo Menções SR. Bacharelado em Ciência da Computação-UnB, 2012-2019. 47

13 Distribuição dos alunos segundo Trancamento. Bacharelado em Ciência da Computação-UnB, 2012-2019. 47

14 Distribuição dos alunos segundo Cursou Verão. Bacharelado em Ciência da Computação-UnB, 2012-2019. 48

15 Distribuição dos alunos segundo evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019 49

16 Distribuição dos alunos segundo Idade ao ingressar (a) e Semestres cursados (b) em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019. 52

17 Distribuição dos alunos segundo Quantidade de Menções SR (a) e Quantidade de Trancamentos (b)em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019. 52

18 Distribuição dos alunos segundo Taxa de Reprovação (a) e IRA (b) em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019. 53

19 Correlação entre IRA e Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019 55

20 Correlação entre a Taxas de Reprovação Geral com as Taxas de Reprovação CIC (a) e MAT (b). Bacharelado em Ciência da Computação-UnB, 2012-2019 55

21 Resíduos para o modelo IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019. 63

22 Distância de Cook para o modelo IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019 64

23 Curva ROC para o modelo IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019 65

24 Resíduos modelo Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019. 66

25 Distância de Cook para o modelo Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019 66

26 Curva ROC para o modelo Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019 67

Sumário

1 Introdução 8

2 Referencial Teórico 11

2.1 Regressão Logística. 11

2.2 Estimação dos Parâmetros 13

2.3 Inferência sobre os Parâmetros do Modelo 15

2.3.1 Teste de Wald 16

2.3.2 Intervalo de Confiança para os Parâmetros do Modelo 17

2.3.3 Teste da Razão de Verossimilhança 18

2.4 Seleção do Modelo 19

2.4.1 Métodos de Seleção Automáticos 22

2.5 Avaliação do Modelo 23

2.5.1 Teste χ^2 de Pearson 23

2.5.2 Teste de Hosmer-Lemeshow 24

2.5.3 Teste *Deviance* de Adequabilidade 24

2.6 Análise de Resíduos 26

2.6.1 Tipos de Resíduos 27

2.6.2 Detecção de Observações Influentes 29

2.7 Estimação da Probabilidade de Sucesso 30

2.8 A Curva ROC 32

3 Metodologia 34

3.1 Banco de Dados 34

3.2 Criação de Variáveis 35

4 Resultados 41

4.1 Análise Descritiva 41

4.1.1 Dados Pessoais 41

4.1.2 Ingresso na Universidade 43

4.1.3 Vida Acadêmica 45
4.1.4 Saída do curso 49
4.2 Análise Bivariada 50
4.2.1 Evasão 50
4.2.2 Correlação entre variáveis 54
4.3 Modelagem 56
4.4 Interpretação dos Parâmetros 61
4.4.1 Modelo com IRA 61
4.4.2 Modelo com Taxa de reprovação MAT 62
4.5 Teste de ajuste e diagnóstico dos modelos 62
4.5.1 Modelo com IRA 63
4.5.2 Modelo com Taxa de Reprovação MAT 65
5 Conclusão 68
Referências 70

1 Introdução

O ensino superior é uma oportunidade valiosa para se obter uma formação adequada para o mundo da atualidade. As universidades oferecem uma ampla gama de cursos de graduação e pós-graduação, abrangendo campos que vão desde ciências exatas e humanas até saúde e tecnologia. Porém, os cursos superiores representam um desafio não só no ingresso, como também na conclusão.

Segundo o relatório apresentado pela Comissão Especial de Estudos sobre a Evasão nas Universidades Públicas Brasileiras, nomeada pela Secretaria de Educação Superior do Ministério da Educação (SESu/MEC), há indícios que existem vários obstáculos vividos pelos alunos até o término da sua jornada na universidade. Algumas das principais questões que os estudantes vivenciam são: situação socioeconômica desfavorável; baixa qualidade do ensino médio; dificuldades de adaptação à universidade; vulnerabilidade; e, até mesmo, a mobilidade (ANDIFES; ABRUEM; SESU/MEC, 1996).

No percurso da vida acadêmica do aluno, esses obstáculos costumam resultar em baixa frequência às aulas, o que por sua vez tende a impactar significativamente o tempo para conclusão do curso e até mesmo levar à desistência.

Tais situações são grandes mazelas na educação brasileira, uma vez que, diante das elevadas taxas de desistência, os investimentos acabam sendo pouco aproveitados. Isso se torna uma preocupação crítica para as universidades públicas, as quais enfrentam limitações financeiras para atender a todas as demandas existentes. Portanto, nesse trabalho de pesquisa, o foco está na evasão, fenômeno que se dá em todos os níveis de ensino, sendo mais comum no ensino médio e no ensino superior, por envolver estudantes pertencentes a faixa etária jovem ou adulta.

Considerando a importância da formação de jovens para o crescimento econômico do país, a evasão do ensino superior deixa de ser apenas uma estatística e passa a ser um fator preocupante. Para tratar este assunto, será importante compreender o que se entende por evasão no ensino superior engloba as seguintes modalidades indicadas pela Associação Nacional dos Dirigentes das Instituições Federais de Ensino Superior (ANDIFES):

> Existem [...] três modalidades principais de evasão, sendo elas as seguintes: a) evasão do curso: desligamento do curso superior em razão do abandono, o que pode ocorrer por não realização da matrícula, transferência de instituição de ensino, mudança de curso, trancamento ou exclusão por desatendimento a alguma norma institucional; b) evasão da instituição, que se caracteriza pelo desligamento da instituição na qual o aluno está matriculado; c) evasão do

> sistema, que configura o abandono, definitivo ou temporário, do sistema de educação superior (ROSA, 2014, p. 247).

De acordo com o site Desafios da Educação (2022) que analisa o Mapa do Ensino Superior divulgado pela Secretaria de Modalidades Especializadas de Educação (SEMESP) em 2021, a taxa nacional de evasão no ensino superior chegou aos 36,6% nas modalidades de ensino a distância e presencial, correspondendo a 3,42 milhões de alunos. Embora esse número por si só pareça impressionante, se reduzirmos a amostra apenas aos cursos de ciências exatas e/ou tecnologia, os resultados são ainda mais alarmantes.

Aplicando este recorte para a evasão no ensino superior de ciências exatas e de tecnologia, Garcia e Gomes (2022) realizaram um levantamento histórico interessante, que destacou os cursos de Ciências Exatas como aqueles com maiores índices de evasão. No estudo, realizado em 53 Instituições de Ensino Superior (IES), foram selecionadas oito áreas do conhecimento e, entre elas, as Ciências Exatas e da Terra superaram todas as outras áreas e registraram a maior taxa de evasão (59%).

No mesmo artigo de Garcia e Gomes (2022), foi registrada uma pesquisa de Silva Filho (2007) que corroborou a primeira pesquisa aqui apresentada. Neste levantamento foi revelado que os cursos das áreas de ciências (agrupados em Ciências, Matemática e Computação), registraram a segunda maior taxa de evasão (29%) do total de cursos avaliados, sendo que, também foi identificado que os cursos de Matemática, Computação, Física e Processamento de Informações estão entre os dez cursos com maior evasão.

Como elemento de estudo sobre a evasão nas áreas de exatas, será interessante investigar o curso de Ciência da Computação, especificamente porque o curso está entre aqueles com maior índice de evasão, listado entre os dez com maior saída de alunos no país.

É importante também considerar o impacto negativo da evasão de alunos neste curso. A evasão do ensino superior em cursos de exatas está longe de ser apenas uma mais uma estatística de insucesso do sistema educacional pois existem impactos tangíveis e intangíveis como consequência desta situação, de médio e longo prazo. É sabido que as ciências exatas e de tecnologia são ferramentas de promoção da inovação nos países e, por consequência, de crescimento e desenvolvimento econômico.

Atualmente, o curso de Ciências da Computação é um dos mais visados pelos mercados público e privado, pois trata de temáticas inovadoras e atuais, necessárias para o desenvolvimento do país. A falta de profissionais nestas áreas coloca o Brasil numa situação de desigualdade e reduz a competitividade da nação frente a outras que investem

mais nesse conhecimento (CNI, 2021).

Porém, há impactos que podem ser vislumbrados mais de perto, como por exemplo, no campus da Universidade. A Universidade de Brasília é um local de excelência na área de tecnologia (UNB, 2023) contando inclusive com um Parque Tecnológico em suas dependências. Para manter toda esta infraestrutura de tecnologia e inovação, os custos de investimento são altos, os quais devem ser feitos pela Universidade para todos os alunos matriculados.

Além do impacto causado na universidade, a decisão pelo abandono de cursos também acarreta consequências significativas para os próprios alunos. A evasão via de regra representa um golpe em suas perspectivas de futuro profissional e pessoal, gerando desorientação e afetando a autoestima e a autoconfiança. Além disso, a evasão tende a afunilar as oportunidades de desenvolvimento de carreira, já que a formação superior é considerada critério decisivo pelo mercado de trabalho tanto para a selecionar novos funcionários quanto para promovê-los.

Considerando todo o exposto, o presente estudo foi direcionado a estudar a evasão universitária dos alunos do curso de Bacharelado em Ciência da Computação na Universidade de Brasília, considerando diversas variáveis a serem definidas ao longo deste documento. Com o intuito de atingir esse objetivo, será conduzida uma análise por meio de um modelo de regressão logística.

2 Referencial Teórico

2.1 Regressão Logística

A regressão logística é uma técnica estatística amplamente utilizada para modelar e prever valores de uma variável categórica com base em variáveis explicativas. Diferentemente da regressão linear, que trata de variáveis resposta contínuas, a regressão logística é aplicada quando a variável de interesse é qualitativa, ou seja, possui dois ou mais resultados possíveis. Esse modelo estima a probabilidade de ocorrência de um evento em relação às variáveis explicativas, permitindo compreender a relação entre essas variáveis e a variável categórica em análise.

Uma das principais características do modelo logístico é a presença de uma variável "resposta" ou "dependente" que é categórica e frequentemente binária. No contexto específico da variável "resposta binária", são consideradas duas categorias: "sucesso", indicando a ocorrência do evento, e "fracasso", referente à não ocorrência do evento. No presente caso, o evento em questão é a evasão de alunos. Assim, a variável "resposta" segue uma distribuição Bernoulli, representada por:

$$Y_i = \begin{cases} 1, & \text{sucesso} \\ 0, & \text{fracasso} \end{cases}, \qquad i = 1, 2, \ldots, n.$$

A probabilidade de sucesso é determinada por π, em que $0 \leq \pi \leq 1$. Por sua vez, a probabilidade de fracasso é complementar à probabilidade de sucesso e é representada por $1 - \pi$. Pode-se expressar a probabilidade de Y assumir o valor 1 como $P(Y_i = 1) = \pi_i$ e a probabilidade de Y assumir o valor 0 como $P(Y_i = 0) = 1 - \pi_i$. Além disso, a média da distribuição é igual a $E(Y_i) = \pi_i$, sendo que π_i varia entre 0 e 1.

Para expressar o modelo de regressão é utilizada a função logística, definida da seguinte maneira:

$$E[Y_i] = \pi(X_1, \ldots, X_p) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}, \quad i = 1, 2, \ldots, n \qquad (2.1.1)$$

Nesse caso, $X_1, \ldots, X_p$ são as variáveis explicativas incluídas no modelo e $\beta_0, \ldots, \beta_p$ são os parâmetros do modelo.

A função $\pi(X_1, \ldots, X_p)$ representa a probabilidade de ocorrência do evento de interesse. A forma da função logística garante que o resultado sempre esteja entre 0 e 1, o que é adequado para modelar uma variável resposta binária.

A curva da regressão logística simples pode ser representada da seguinte forma:

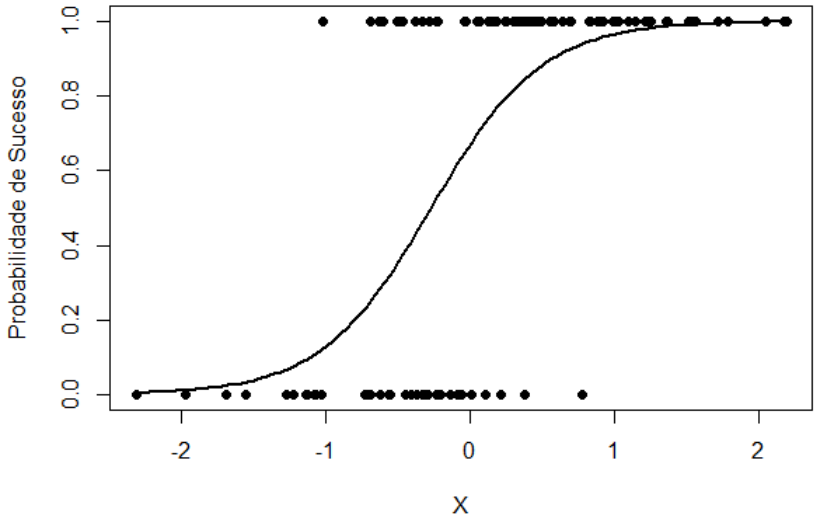


Figura 1: Exemplo da curva da função de regressão logística
Fonte: https://goo.gl/nwec4Q

E ainda, é possível conectar a estrutura linear do modelo ao parâmetro utilizando uma função de ligação. A transformação conhecida como *logito* destaca-se entre as diversas opções de função de ligação devido à sua característica linear, simplificando a interpretação dos coeficientes. Essa transformação corresponde ao logaritmo natural da *odds*, que representa a chance de sucesso e é expressa por:

$$logito(\pi) = \ln(odds) = \ln\left(\frac{\pi(X_1, \ldots, X_p)}{1 - \pi(X_1, \ldots, X_p)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p, \qquad (2.1.2)$$

em que β_k mede o efeito de X_k sobre o logaritmo da chance de sucesso ($Y = 1$), sendo $k = 0, 1, \ldots, p$.

Utilizando notação matricial, é possível escrever a expressão 2.1.1 a partir dos seguintes vetores:

$$\boldsymbol{\beta_{p\times 1}} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{bmatrix}, \qquad \boldsymbol{X_{p\times 1}} = \begin{bmatrix} 1 \\ X_1 \\ X_2 \\ \vdots \\ X_{p-1} \end{bmatrix}, \qquad \boldsymbol{X_{i_{p\times 1}}} = \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{i.p-1} \end{bmatrix}. \tag{2.1.3}$$

Ao multiplicar os vetores, obtém-se as expressões a seguir:

$$\boldsymbol{X'\beta} = \beta_0 + \beta_1 X_1 + \cdots + \beta_{p-1} X_{p-1}, \tag{2.1.4}$$

$$\boldsymbol{X_i'\beta} = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1} X_{i.p-1}. \tag{2.1.5}$$

Portanto, a forma matricial do modelo de regressão logística é representada por:

$$E[Y] = \frac{\exp(\boldsymbol{X'\beta})}{1 + \exp(\boldsymbol{X'\beta})}. \tag{2.1.6}$$

Adicionalmente, a função *logito* apresentada em 2.1.2 também pode ser escrita matricialmente.

$$\pi' = \ln\left(\frac{\pi}{1-\pi}\right) = \boldsymbol{X'\beta}. \tag{2.1.7}$$

2.2 Estimação dos Parâmetros

A fim de ajustar um modelo de regressão, é essencial estimar os parâmetros $\beta_0, \ldots, \beta_p$ do modelo. Isso é feito utilizando o método de máxima verossimilhança. Esse método busca encontrar os estimadores $\hat{\beta}_0, \ldots, \hat{\beta}_p$ que maximizam a função de verossimilhança, a partir dos dados da amostra, ou seja, o conjunto de observações. A estimação por máxima verossimilhança permite encontrar os valores dos parâmetros do modelo que têm maior probabilidade de reproduzir o padrão de observações na amostra de dados.

Como mencionado anteriormente, a variável Y_i segue uma distribuição Bernoulli e, portanto, a função de probabilidade conjunta é dada por:

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} \pi_i^{Y_i} (1-\pi_i)^{1-Y_i}. \tag{2.2.1}$$

Assim, a melhor forma para encontrar as estimativas de máxima verossimilhança é aplicando o logaritmo da função de probabilidade conjunta:

$$\begin{aligned}\ln g(Y_1,\ldots,Y_n) &= \ln \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n}\left[Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n}\ln(1-\pi_i). \end{aligned} \tag{2.2.2}$$

Sabendo que

$$1-\pi_i = [1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)]^{-1}, \tag{2.2.3}$$

a função de log verossimilhança é definida como:

$$\ln(L(\beta_0,\beta_1,\ldots,\beta_p)) = \sum_{i=1}^{n} Y_i(\beta_0+\beta_1X_1+\cdots+\beta_pX_p) - \sum_{i=1}^{n}\ln[1+\exp(\beta_0+\beta_1X_1+\cdots+\beta_pX_p)]. \tag{2.2.4}$$

Em termos matriciais, tem-se:

$$\ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n} Y_i(\boldsymbol{X_i'\beta}) - \sum_{i=1}^{n}\ln[1+\exp(\boldsymbol{X_i'\beta})]. \tag{2.2.5}$$

Para encontrar as estimativas de máxima verossimilhança, é preciso derivar log verossimilhança em relação a cada parâmetro do modelo.

$$\frac{\partial \ln(L(\beta_0,\beta_1,\ldots,\beta_p))}{\partial\beta_k}, \quad k=0,1,2,\ldots,p. \tag{2.2.6}$$

Como não há uma fórmula fechada para os valores de β_k que maximizem a função de máxima verossimilhança, é necessário utilizar o método de Newton-Raphson para encontrar as estimativas $b_0, b_1, \ldots, b_{p-1}$. Dessa forma, o vetor das estimativas de máxima verossimilhança é dado por:

$$\boldsymbol{b_{p\times 1}} = \begin{bmatrix} b_0 \\ b_1 \\ \vdots \\ b_{p-1} \end{bmatrix} \tag{2.2.7}$$

- **Interpretação dos Parâmetros**

No caso da regressão linear simples, a função resposta da regressão logística é denotada por:

$$\hat{\pi} = \frac{\exp(b_0 + b_1 X)}{1 + \exp(b_0 + b_1 X)}, \tag{2.2.8}$$

sendo, a chance de x:

$$odds_x = \frac{\pi(x)}{1 - \pi(x)} = \exp(b_0 + b_1 X) = e^{b_0}(e^{b_1})^x, \tag{2.2.9}$$

e a chance de $x + 1$:

$$odds_{x+1} = \frac{\pi(x+1)}{1 - \pi(x+1)} = \exp(b_0 + b_1(X+1)) = e^{b_0}(e^{b_1})^x e^{b_1}. \tag{2.2.10}$$

A razão de chances entre x e $x + 1$ é dada por:

$$\frac{odds_{x+1}}{odds_x} = \frac{e^{b_0}(e^{b_1})^x e^{b_1}}{e^{b_0}(e^{b_1})^x} = e^{b_1}. \tag{2.2.11}$$

Assim, a interpretação de b_1 pode ser encontrada a partir do *logito*, definido como:

$$logito(\pi(x)) = \ln\left(\frac{odds_{x+1}}{odds_x}\right) = \ln(e^{b_1}) = b_1. \tag{2.2.12}$$

Portanto, a chance estimada de sucesso para o nível $x + 1$ é igual a do nível x multiplicada por e^{b_1}.

A mesma lógica se aplica a qualquer parâmetro β_k do modelo de regressão logística apresentado em 2.1.1.

2.3 Inferência sobre os Parâmetros do Modelo

Após estimar os parâmetros do modelo, tem-se interesse em avaliar a significância das variáveis no modelo.

Os métodos de inferência que serão abordados nessa Seção baseiam-se em tamanhos de amostra grandes. Em tais amostras, e sob condições geralmente aplicáveis, os estimadores de máxima verossimilhança para a regressão logística são distribuídos de

maneira aproximadamente normal, apresentando viés mínimo ou inexistente. Além disso, suas variâncias e covariâncias aproximadas são determinadas pelas derivadas parciais de segunda ordem do logaritmo da função de verossimilhança.

Seja $\boldsymbol{G}$ a matriz das segundas derivadas parciais da função de verossimilhança (apresentada em 2.2.5) em relação aos parâmetros $\beta_0, \beta_1, \ldots, \beta_{p-1}$:

$$\boldsymbol{G_{p\times p}} = [g_{ij}], \quad i = 0, 1, \ldots, p-1 \quad \text{e} \quad j = 0, 1, \ldots, p-1, \tag{2.3.1}$$

em que:

$$g_{00} = \frac{\partial \ln(L(\boldsymbol{\beta}))}{\partial \beta_0^2},$$
$$g_{01} = \frac{\partial^2 \ln(L(\boldsymbol{\beta}))}{\partial \beta_0 \partial \beta_1},$$

etc.

Este conjunto de derivadas parciais de segunda ordem na matriz $\boldsymbol{G}$ é conhecido como a matriz Hessiana. Quando essas derivadas são estimadas em $\boldsymbol{\beta = b}$, ou seja, nas estimativas de máxima verossimilhança, a matriz de variância-covariância dos coeficientes de regressão estimados para a regressão logística pode ser obtida da seguinte forma:

$$s^2\{\boldsymbol{b}\} = ([-g_{ij}]_{\beta=b})^{-1}. \tag{2.3.2}$$

As inferências sobre os coeficientes de regressão para o modelo de regressão logística são baseadas no seguinte resultado aproximado quando o tamanho da amostra é grande:

$$\frac{b_k - \beta_k}{s(b_k)} \sim z, \quad k = 0, 1, \ldots, p-1, \tag{2.3.3}$$

em que z é uma variável aleatória normal padrão e $s(\hat{\beta}_k)$ é o desvio padrão aproximado estimado de b_k obtido a partir de 2.3.2.

2.3.1 Teste de Wald

O teste de Wald é empregado na análise de regressão logística para avaliar a relevância estatística dos coeficientes calculados. Ele verifica se cada coeficiente é signifi-

cativamente diferente de zero, o que indica se uma variável independente tem uma relação estatisticamente significativa com a variável dependente. As hipóteses a serem testadas, então, são as seguintes:

$$\begin{cases} H_0 : \beta_k = 0 \\ H_1 : \beta_k \neq 0 \end{cases}, \quad k = 0, \ldots, p.$$

A estatística do teste é obtida pela comparação entre a estimativa de máxima verossimilhança do parâmetro ($\hat{\beta}_k$) e a estimativa de seu desvio padrão ($s(\hat{\beta}_k)$):

$$Z^2 = \left(\frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \right)^2 \sim \chi^2_{(1)\ g.l}. \tag{2.3.4}$$

Sob a hipótese nula H_0, a estatística do teste segue aproximadamente uma distribuição Qui-quadrado com 1 grau de liberdade.

2.3.2 Intervalo de Confiança para os Parâmetros do Modelo

Baseado na estimativa pontual do parâmetro, é possível criar uma estimativa intervalar para o parâmetro com um nível de confiança de $1 - \alpha$. Dessa maneira, o intervalo de confiança de $1 - \alpha$ para um parâmetro β_k é calculado da seguinte forma:

$$\hat{\beta}_k \pm z_{1-\frac{\alpha}{2}} s(\hat{\beta}_k), \quad k = 0, \ldots, p, \tag{2.3.5}$$

em que:

- $z_{1-\frac{\alpha}{2}}$ é o percentil $(1 - \frac{\alpha}{2})100$ da distribuição Normal Padrão;
- $s(\hat{\beta}_k)$ é a estimativa do erro padrão do estimador $\hat{\beta}_k$.

Intervalo de confiança para a *odds*

Uma vez obtido o intervalo de confiança para β_k no modelo de regressão logística, é possível determinar o intervalo para a *odds ratio* aplicando a função exponencial aos limites do intervalo para $\hat{\beta}_k$:

$$\exp\left[\hat{\beta}_k \pm z_{1-\frac{\alpha}{2}} s(\hat{\beta}_k)\right], \quad k = 0, \ldots, p. \tag{2.3.6}$$

A análise do intervalo da *odds ratio* fornece insights adicionais sobre a relação entre a variável preditora e a variável de resposta. Ao considerar o intervalo de confiança para β_k, juntamente com a correspondente *odds ratio*, é possível obter uma compreensão mais clara tanto da relação entre as variáveis quanto da incerteza associada à estimativa dos parâmetros do modelo de regressão logística. Essa análise conjunta fornece informações relevantes para a interpretação dos resultados e auxilia na tomada de decisões.

2.3.3 Teste da Razão de Verossimilhança

O teste da razão de verossimilhança é aplicado para verificar a significância dos parâmetros estimados das variáveis independentes em um modelo. Seu objetivo é avaliar se todos os parâmetros β's associados às variáveis independentes são iguais a zero, o que indicaria a ausência de relação linear entre as variáveis independentes e a variável dependente. As hipóteses a serem testadas são as seguintes:

$$\begin{cases} H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0; \\ H_1 : \text{Existe pelo menos um } \beta_k \neq 0, \qquad k = 1, \ldots, p. \end{cases}$$

A estatística do teste é dada por:

$$G^2 = -2(L_0 - L_1) \sim \chi^2_{(v)\ g.l}, \tag{2.3.7}$$

em que

- L_0 é o máximo do log da função de verossimilhança sob a hipótese H_0;
- L_1 é o máximo do log da função de verossimilhança sob a hipótese H_1.

Para uma amostra de tamanho n grande, a estatística G^2 segue uma distribuição $\chi^2_{(v=p-q)\ g.l}$ em que o grau de liberdade corresponde à diferença no número de parâmetros entre o modelo reduzido (sob a hipótese H_0) e o completo (sob a hipótese H_1).

Dessa forma, a regra de decisão é definida por:

$$\text{Se } G^2 \leq \chi^2(1-\alpha; v), \text{ aceita-se } H_0;$$
$$\text{Se } G^2 \leq \chi^2(1-\alpha; v), \text{ rejeita-se } H_0.$$

Também é possível utilizar o teste de Razão de Verossimilhança para determinar se diversos β's do modelo são iguais a zero. Nesse contexto, as hipóteses testadas são as seguintes:

$$\begin{cases} H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0; \\ H_1 : \text{Nem todos } \beta_k \text{ em } H_0 \text{ são iguais a } 0, \qquad k = q, \ldots, p-1. \end{cases}$$

Por conveniência, o modelo é organizado de modo que os últimos $p-q$ coeficientes sejam aqueles a serem testados. Nesse caso, p representa o número total de parâmetros no modelo completo e q representa o número de parâmetros no modelo reduzido.

2.4 Seleção do Modelo

A seleção do modelo desempenha um papel crucial na análise estatística, buscando escolher o modelo mais parcimonioso, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados, considerando a melhor combinação de variáveis independentes e a forma funcional que melhor descreve a relação com a variável dependente.

1. **Critério de Informação de Akaike (AIC)**

   O Critério de Informação de Akaike é uma medida que busca encontrar o equilíbrio entre o ajuste do modelo aos dados e sua complexidade. Ele é calculado levando em consideração a função de verossimilhança do modelo e o número de parâmetros estimados. O objetivo é minimizar o valor do AIC, pois isso indica um modelo bem ajustado. Importante ressaltar que o AIC incorpora uma penalização automática ao número de parâmetros, promovendo a preferência por modelos mais parcimoniosos. Essa penalização é crucial para evitar a escolha de modelos excessivamente comple-

xos. Assim, ao comparar modelos, opta-se pelo menor valor de AIC, considerando não apenas o ajuste aos dados, mas também a penalização associada ao número de parâmetros, o que promove a escolha de modelos mais eficazes.

O cálculo do AIC é dado por:

$$AIC_p = -2 \ln L(b) + 2p. \tag{2.4.1}$$

A análise gráfica desse critério é feita da seguinte maneira:

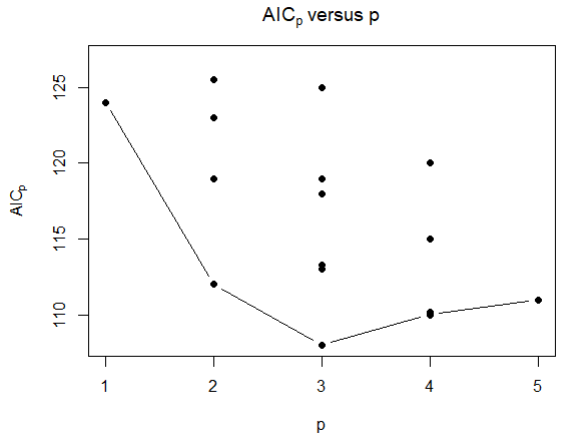


Figura 2: Exemplo do gráfico AIC_p

Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 585).

Com base na Figura 2, o critério de seleção de modelos indicou que aqueles com três parâmetros são os mais adequados para o fenômeno do exemplo. Entre esses modelos, a escolha final foi determinada pelo menor valor de AIC_p.

2. **Critério de Informação Bayesiano (BIC)**

O Critério de Informação Bayesiano é uma alternativa ao AIC que incorpora a teoria da decisão bayesiana. Assim como o AIC, o BIC leva em consideração tanto o ajuste dos dados quanto a complexidade do modelo, mas penaliza a complexidade mais rigorosamente do que o AIC. Isso significa que ele favorece modelos mais simples, mesmo em casos em que o ajuste dos dados é semelhante. Ao escolher entre modelos, selecionamos aquele com o menor valor de BIC.

O cálculo do BIC é dado por:

$$BIC_p = -2 \ln L(b) + p \ln (n). \tag{2.4.2}$$

A análise gráfica desse critério é feita da seguinte maneira:

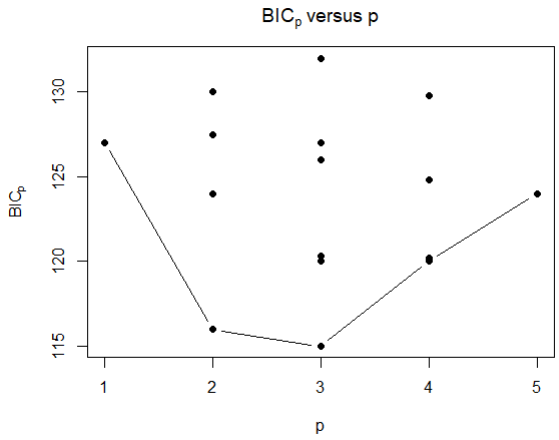


Figura 3: Exemplo do gráfico BIC_p

Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 585).

Com base na Figura 5, o critério de seleção de modelos indicou que aqueles com três parâmetros são os mais adequados para o fenômeno do exemplo. Entre esses modelos, a escolha final é determinada pelo menor valor de BIC_p.

3. **Pseudo** R^2

O Pseudo R^2 é uma medida de ajuste do modelo que busca quantificar a proporção da variabilidade explicada pela combinação de variáveis independentes selecionadas. Diferente do R^2 tradicional, o Pseudo R^2 é utilizado em modelos de regressão que não seguem os pressupostos lineares, como a regressão logística. Em geral, um valor mais alto de Pseudo R^2 indica um melhor ajuste do modelo aos dados.

O cálculo do Pseudo R^2 é dado por:

$$R_p^2 = \frac{\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y}_i)^2}. \tag{2.4.3}$$

2.4.1 Métodos de Seleção Automáticos

Existem diversos métodos de seleção de variáveis amplamente utilizados, tais como os métodos *Backward*, *Forward* e *Stepwise*. A escolha do método mais adequado deve levar em consideração a complexidade do modelo, a interpretabilidade dos resultados e a capacidade de generalização dos resultados para novos dados.

1. **Eliminação *Backward***

   O método de eliminação *backward* começa com um modelo completo contendo todas as variáveis independentes e, em seguida, remove iterativamente as variáveis menos significativas até que reste apenas um conjunto de variáveis estatisticamente significativas para a variável dependente. A remoção é realizada com base em critérios como o p-valor do teste para cada variável ou critérios de informação, como AIC e BIC, que penalizam modelos mais complexos.

2. **Seleção *Forward***

   O método de seleção *forward* começa com um modelo sem nenhuma variável explicativa e, em seguida, adiciona iterativamente as variáveis independentes que têm o maior impacto na variável dependente, até que nenhuma outra variável possa ser adicionada para melhorar significativamente o modelo. A remoção de variáveis é realizada utilizando critérios semelhantes ao método *backward*, considerando a significância estatística (p-valor) ou critérios de informação.

3. **Regressão *Stepwise***

   O método de regressão *stepwise* combina os dois métodos anteriores. Ele começa com um modelo sem nenhuma variável explicativa e adiciona as variáveis independentes que são estatisticamente significativas. Em seguida, remove iterativamente as variáveis menos significativas e adiciona novas variáveis que melhorem o modelo. A remoção e adição de variáveis são guiadas pelos mesmos critérios mencionados anteriormente, visando um equilíbrio entre a simplicidade do modelo e a inclusão das variáveis mais relevantes.

2.5 Avaliação do Modelo

A análise feita sobre o ajustamento do modelo de regressão logística é realizada através de testes de adequação, em que o interesse é assegurar a significância das variáveis no modelo. Medidas de adequabilidade de ajuste ajudam a determinar se o modelo de regressão logística escolhido é apropriado para descrever a relação entre as variáveis independentes e a variável dependente.

As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \text{O Modelo de Regressão Logística ajusta-se aos dados;} \\ H_1 : \text{O Modelo de Regressão Logística não se ajusta aos dados.} \end{cases}$$

2.5.1 Teste χ^2 de Pearson

O teste Qui-Quadrado de Pearson é utilizado para determinar se o modelo logístico é adequado para o conjunto de dados.

As hipóteses a serem testadas são:

$$\begin{cases} H_0 : E(Y) = \dfrac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}; \\ H_1 : E(Y) \neq \dfrac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}. \end{cases}$$

E a estatística do teste é definida por :

$$\chi^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(O_{jk} - E_{jk})^2}{E_{jk}}, \tag{2.5.1}$$

em que:

- O_{jk} é o número observado de eventos na categoria j do conjunto k;
- E_{jk} é o número esperado de eventos na categoria j do conjunto k.

Sob a hipótese nula H_0, a estatística do teste segue aproximadamente uma distribuição Qui-quadrado com $c - p$ graus de liberdade, quando o tamanho amostral é grande e o número de parâmetros do modelo é menor que o número de categorias, ou seja, $p < c$ em que:

- p é o número de parâmetros do modelo;
- c é número conjuntos de valores distintos das variáveis explicativas.

A regra de decisão então é dada por:

$$\text{Se } G^2 \leq \chi^2(1-\alpha; c-p), \text{ aceita-se } H_0;$$
$$\text{Se } G^2 \leq \chi^2(1-\alpha; c-p), \text{ rejeita-se } H_0.$$

2.5.2 Teste de Hosmer-Lemeshow

O Teste de Hosmer-Lemeshow é utilizado para avaliar a adequação de um modelo de regressão logística aos dados observados. A ideia central do teste é dividir os dados em grupos com base nas probabilidades preditas pelo modelo de regressão logística. Esses grupos são chamados de "grupos de decil"e são formados classificando as observações de acordo com suas probabilidades preditas e dividindo-as em 10 grupos com aproximadamente a mesma quantidade de observações em cada grupo.

A estatística do teste é a mesma representada em 2.5.1, onde c é o número total de grupos.

É importante destacar que, em determinadas situações, o teste de Hosmer-Lemeshow é mais informativo do que o teste Qui-Quadrado de Pearson. Isso se deve ao fato de que o teste Qui-Quadrado de Pearson avalia a adequação global do modelo, enquanto o teste de Hosmer-Lemeshow adota uma abordagem mais refinada com a divisão dos grupos.

2.5.3 Teste *Deviance* de Adequabilidade

O Teste *Deviance* de Adequabilidade é utilizado para comparar dois modelos: o modelo completo e o modelo restrito, com o objetivo de verificar se o modelo restrito é uma melhor escolha em termos de ajuste aos dados.

Modelo Completo:

O Modelo Completo é caracterizado pela estimação separada de cada categoria da variável dependente. Essa abordagem resulta em um número excessivo de parâmetros, o que pode acarretar problemas de *overfitting*. Por essa razão, o Modelo Completo é considerado saturado.

$$E(Y_{ij}) = \pi_j, \quad i = 1, \ldots, n \quad \text{e} \quad j = 1, 2, \ldots, c. \tag{2.5.2}$$

Modelo Restrito (sob H_0):

O Modelo Restrito é uma versão simplificada do Modelo Completo, no qual são aplicadas restrições aos parâmetros, impostas pela hipótese nula. Na regressão logística, o Modelo Restrito representa a relação entre a variável dependente e as variáveis independentes por meio da função logística:

$$E[Y_{ij}] = \pi(X_1, \ldots, X_p) = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p\right)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}. \tag{2.5.3}$$

Esse Modelo Restrito é considerado mais apropriado, pois oferece uma interpretação mais simplificada e evita problemas de *overfitting*.

Para realizar o cálculo da estatística do teste *deviance*, é necessário considerar a estatística do teste da razão de verossimilhança:

$$G^2 = -2 \ln \left(\frac{L(R)}{L(F)} \right) = -2[\ln\left(L(R)\right) - \ln\left(L(F)\right)], \tag{2.5.4}$$

em que:

- $L(R)$ é o valor da função de verossimilhança para o modelo restrito;
- $L(F)$ é o valor da função de verossimilhança para o modelo completo;

E ainda, é preciso obter os valores das máximas verossimilhanças para os modelos restrito e completo. A máxima verossimilhança de $L(R)$ é obtida ao ajustar o modelo restrito, enquanto as estimativas de máxima verossimilhança dos c parâmetros no modelo $L(F)$ são calculadas com base nas proporções amostrais apresentadas em:

$$p_j = \frac{Y_{.j}}{n_j}, \quad j = 1, 2, \ldots, c \tag{2.5.5}$$

Sendo $\hat{\pi_j}$ a estimativa de π_j pelo Modelo Restrito para cada X_j, $j = 1, 2, \ldots, c$, a estatística do teste é dada por:

$$\begin{aligned} G^2 &= -2\sum_{j=1}^{c}\left[Y_j \ln\left(\frac{\hat{\pi_j}}{p_j}\right) + (n_j - Y_j)\ln\left(\frac{1-\hat{\pi_j}}{1-p_j}\right)\right] \\ &= DEV(X_0, X_1, \ldots, X_{p-1}). \end{aligned} \tag{2.5.6}$$

Sob a hipótese nula H_0, a estatística do teste segue aproximadamente uma distribuição Qui-quadrado com $c - p$ graus de liberdade, quando n_j é grande e o número de parâmetros do modelo é menor que o número de categorias, ou seja, $p < c$.

- p é o número de parâmetros do modelo;
- c é número conjuntos de valores distintos das variáveis explicativas.

A regra de decisão então é dada por:

$$\text{Se } DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1-\alpha; c-p), \text{ aceita-se } H_0;$$
$$\text{Se } DEV(X_0, X_1, \ldots, X_{p-1}) \leq \chi^2(1-\alpha; c-p), \text{ rejeita-se } H_0.$$

2.6 Análise de Resíduos

A Análise de Resíduos para regressão logística é mais desafiadora em comparação com os modelos de regressão linear, devido ao fato de que as respostas, representadas por Y_i, assumem apenas os valores 0 e 1. Isso implica que o resíduo ordinário do caso i, denotado por e_i, assumirá um de dois valores possíveis:

$$e_i = \begin{cases} 1 - \hat{\pi}_i, & \text{para } Y_i = 1 \\ -\hat{\pi}_i, & \text{para } Y_i = 0 \end{cases}, \qquad i = 1, \ldots, n.$$

Os resíduos comuns não seguem uma distribuição normal e, de fato, sua distribuição sob a suposição de que o modelo ajustado está correto é desconhecida. Portanto, os gráficos dos resíduos comuns em relação aos valores ajustados ou às variáveis preditoras geralmente não fornecem informações úteis.

2.6.1 Tipos de Resíduos

1. **Resíduos de Pearson Padronizados**

   Os resíduos comuns podem ser padronizados e tornados mais comparáveis ao dividir cada um deles pelo desvio padrão estimado de Y_i, que é igual a, $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}$. Essa transformação resulta nos resíduos de Pearson, dado por:

$$r_{P_i} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}}, \quad i = 1, \ldots, n. \tag{2.6.1}$$

2. **Resíduos de Pearson Studentizados**

   Os Resíduos de Pearson não possuem variância unitária, uma vez que não foi considerada a variação inerente ao valor ajustado $\hat{\pi}_i$. Um método mais apropriado é dividir os resíduos comuns pelo seu desvio padrão estimado. Essa estimativa é aproximada por $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$, em que h_{ii} é o i-ésimo elemento da matriz diagonal $n \times n$ estimada para a regressão logística:

$$H = \hat{W}^{\frac{1}{2}} X (X'\hat{W}X)^{-1} X'\hat{W}^{\frac{1}{2}}. \tag{2.6.2}$$

   Nesse caso, a matriz $\hat{W}$ é uma matriz diagonal de dimensão $n \times n$, na qual os elementos são calculados como $\hat{\pi}_i(1-\hat{\pi}_i)$. A matriz X é a matriz usual de dimensão $n \times p$ e a matriz $\hat{W}^{\frac{1}{2}}$ é uma matriz diagonal de dimensão $n \times n$, em que os elementos diagonais são iguais às raízes quadradas dos elementos correspondentes em $\hat{W}$. Os Resíduos de Pearson Studentizados resultantes são definidos por:

$$r_{SP_i} = \frac{r_{P_i}}{\sqrt{1-h_{ii}}}, \quad i = 1, \ldots, n. \tag{2.6.3}$$

   É importante lembrar que, para a regressão linear múltipla, a matriz H satisfaz a expressão matricial $\hat{Y} = HY$. No caso da regressão logística, a matriz é desenvolvida de forma análoga e se aproxima da expressão $\hat{\pi'} = HY$, em que π' é um vetor de preditores lineares de dimensão $(n \times 1)$.

3. **Resíduos *Deviance***

Os Resíduos *deviance* representam a diferença entre as respostas observadas e as respostas previstas pelo modelo de regressão logística. São utilizados para avaliar o ajuste do modelo aos dados e identificar possíveis padrões não explicados pelo modelo. No caso de dados binários, o modelo é dado por:

$$DEV(X_0, X_1, \ldots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i \ln(\hat{\pi}_i) + (1 - Y_i)\ln(1 - \hat{\pi}_i)], \quad i = 1, \ldots, n. \tag{2.6.4}$$

O Resíduo *deviance* para o caso i é definido como a raiz quadrada da contribuição do caso i para a *deviance* do modelo DEV em 2.6.4:

$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\sum_{i=1}^{n}[Y_i \ln(\hat{\pi}_i) + (1 - Y_i)\ln(1 - \hat{\pi}_i)]}, \quad i = 1, \ldots, n. \tag{2.6.5}$$

O Resíduo é positivo quando $Y_i \geq \hat{\pi}_i$ e negativo quando $Y_i < \hat{\pi}_i$. Portanto, a soma dos resíduos *deviance* ao quadrado corresponde à *deviance* do modelo em 2.6.4.

$$\sum_{i=1}^{n}(dev_i)^2 = DEV(X_0, \ldots, X_{p-1}). \tag{2.6.6}$$

4. **Gráfico de Resíduos**

Nos modelos de regressão logística, os gráficos de resíduos são empregados para avaliar a adequação do ajuste do modelo. A Figura 4 representa dois exemplos de gráfico de resíduos.

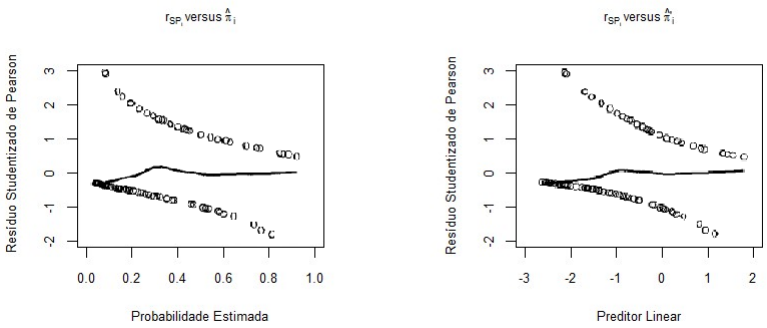


Figura 4: Exemplo do gráfico de resíduos

Fonte: NETER, J. et al. Applied Linear Statistical Models (p. 595).

Se o modelo estiver correto, um ajuste suavizado (*lowess smooth*) do gráfico dos resíduos em relação à probabilidade estimada $\hat{\pi}_i$ resultará aproximadamente em uma linha horizontal com intercepto no zero. Qualquer desvio significativo desse padrão sugere que o modelo pode ser inadequado.

2.6.2 Detecção de Observações Influentes

1. **Estatísticas Qui-Quadrado de Pearson e *Deviance* para Valores Influentes**

Considerando as estatísticas de Pearson (χ^2) e *Deviance* (DEV) mencionadas previamente, calculadas a partir do conjunto de dados completo, utiliza-se $\chi^2_{(i)}$ e $DEV_{(i)}$ para representar os valores dessas estatísticas de teste quando o caso i for excluído. As estatísticas $\triangle\chi^2_i$ e $\triangle dev_i$ do *i-ésimo* caso serão definidas como a mudança nas estatísticas Qui-quadrado e *deviance* quando o caso i for removido:

$$\triangle\chi^2_i = \chi^2 - \chi^2_{(i)} \tag{2.6.7}$$

$$\triangle dev_i = DEV - DEV_{(i)} \tag{2.6.8}$$

Calcular as n estatísticas $\triangle\chi^2_i$ ou as n estatísticas $\triangle dev_i$ requer realizar n maximizações da verossimilhança, o que pode ser um processo demorado. Com o objetivo de obter um cálculo mais rápido, foram desenvolvidas as seguintes aproximações:

$$\triangle\chi^2_i = r^2_{SP_i} \tag{2.6.9}$$

$$\triangle dev_i = h_{ii} r^2_{SP_i} + dev^2_i \tag{2.6.10}$$

A interpretação dessas estatísticas não é simples. Decidir se um caso é ou não influente geralmente envolve uma avaliação visual através de gráficos. As estatísticas $\triangle\chi^2_i$ e $\triangle dev_i$ são comumente representadas em gráficos em relação ao padrão i ou em relação a $\hat{\pi}_i$. Valores extremos são identificados como picos no gráfico em relação ao padrão i e como outliers nas regiões superiores do gráfico em relação a $\hat{\pi}_i$.

2. **Distância de *Cook***

A Distância de *Cook* é uma medida utilizada para identificar observações influentes em uma análise de regressão. No caso da regressão logística, a distância de *Cook* mede a mudança padronizada no preditor linear ajustado $\hat{\pi}_i$ quando o i-ésimo caso é excluído. Assim como as estatísticas delta mencionadas acima, obter esses valores requer n maximizações da verossimilhança. Portanto, utiliza-se a seguinte aproximação para calcular a distância de *Cook*:

$$D_i = \frac{r^2_{P_i} h_{ii}}{p(1 - h^2_{ii})} \tag{2.6.11}$$

Para identificar observações influentes, podem ser construídos gráficos de índice dos valores de influência h_{ii} no espaço X. Além disso, gráficos de índice de D_i podem ser utilizados para identificar os casos que têm um grande efeito no preditor linear ajustado. No entanto, não existem regras fixas para interpretar a magnitude desses diagnósticos, sendo necessário confiar em uma avaliação visual adequada por meio dos gráficos.

2.7 Estimação da Probabilidade de Sucesso

1. **Estimador Pontual**

Denota-se o vetor de níveis das variáveis X para as quais π é estimado por X_h:

$$\boldsymbol{X_{h_{p\times 1}}} = \begin{bmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{h.p-1}, \end{bmatrix}, \tag{2.7.1}$$

e a resposta média de interesse por π_h:

$$\pi_h = [1 + \exp(-\boldsymbol{X_h'\beta})]^{-1}. \tag{2.7.2}$$

O estimador pontual de π_h, denotado por $\hat{\pi_h}$, é definido da seguinte forma:

$$\hat{\pi_h} = [1 + \exp(-\boldsymbol{X_h'b})]^{-1}, \tag{2.7.3}$$

em que $\boldsymbol{b}$ é o vetor de coeficientes de regressão estimados na expressão 2.2.7.

2. **Estimativa de Intervalo**

Um intervalo de confiança para π_h é obtido em duas etapas. Inicialmente, é necessário calcular os limites de confiança para a resposta média *logito* π_h'. Posteriormente, utiliza-se a relação 2.7.4 para obter os limites de confiança para a resposta média π_h. Para uma compreensão clara, considera-se a equação 2.7.4 para $\boldsymbol{X = X_h}$.

$$E[Y] = \frac{\exp(\boldsymbol{X_h'\beta})}{1 + \exp(\boldsymbol{X_h'\beta})}. \tag{2.7.4}$$

Reformulando a expressão acima utilizando a informação de que $E[Y_h] = \pi_h$ e $\boldsymbol{X_h'\beta} = \pi_h'$ tem-se que:

$$\pi_h = \frac{\exp(\pi_h')}{1 + \exp(\pi_h')}. \tag{2.7.5}$$

A relação expressa em 2.7.5 é utilizada para converter os limites de confiança para π_h' em limites de confiança para π_h.

O estimador pontual da resposta média *logito*, $\pi_h' = \boldsymbol{X_h'\beta}$, é calculado como $\hat{\pi_h'} = \boldsymbol{X_h'b}$, enquanto a variância estimada de $\hat{\pi_h'} = \boldsymbol{X_h'b}$ é definida como:

$$s^2\left\{\hat{\pi_h'}\right\} = s^2\left\{\boldsymbol{X_h'b}\right\} = \boldsymbol{X_h'}s^2\left\{\boldsymbol{b}\right\}\boldsymbol{X_h}, \tag{2.7.6}$$

em que $s^2\{\boldsymbol{b}\}$ denota a matriz variância-covariância dos coeficientes estimados de regressão conforme descrito em 2.3.2 quando o tamanho da amostra é grande.

Os limites de confiança aproximados de $1-\alpha$ para a resposta média *logito* π_h em amostras grandes são então obtidos da maneira usual:

$$L = \hat{\pi'_h} - z\left(1-\frac{\alpha}{2}\right) s\{\pi'_h\}, \tag{2.7.7}$$

$$U = \hat{\pi'_h} + z\left(1-\frac{\alpha}{2}\right) s\{\pi'_h\}, \tag{2.7.8}$$

sendo L e U os limites inferior e superior para π'_h, respectivamente.

Por fim, utiliza-se a relação monótona entre π'_h e π_h, apresentada em 2.7.5, para converter os limites de confiança L e U em limites de confiança aproximados L^* e U^* para a resposta média π_h.

$$L^* = [1+\exp(-L)]^{-1} \tag{2.7.9}$$

$$U^* = [1+\exp(-U)]^{-1} \tag{2.7.10}$$

2.8 A Curva ROC

Uma forma de validar a confiabilidade do modelo é analisando a Curva ROC (*Receiver Operating Characteristic*). A Curva ROC é uma ferramenta utilizada para avaliar a performance de modelos de classificação, como a regressão logística, em prever resultados binários. Ela representa graficamente a taxa de verdadeiros positivos (sensibilidade) em função da taxa de falsos positivos (1 - especificidade) para diferentes valores de ponto de corte do modelo.

A sensibilidade representa a proporção de verdadeiros positivos em relação ao total de casos positivos. Já a especificidade, representa a proporção de verdadeiros negativos em relação ao total de casos negativos.

A representação gráfica da curva ROC é dada da seguinte forma:

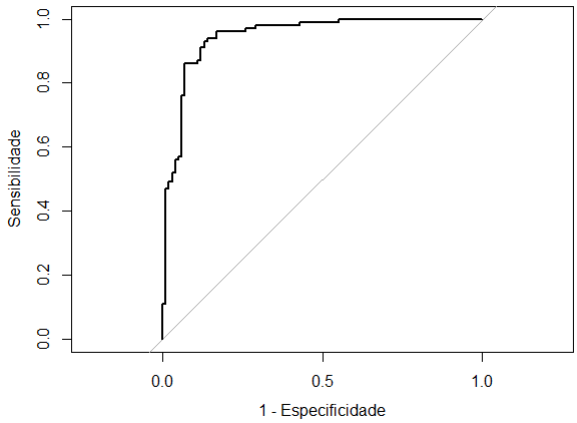


Figura 5: Exemplo do gráfico da curva ROC

Fonte: Agresti, An Introduction to Categorical Data Analysis (p. 112).

A área sob a curva, que varia de zero a um, é uma medida que avalia a capacidade do modelo em discriminar entre dois conjuntos: aqueles que experimentaram o evento de interesse e aqueles que não o experimentaram.

Denotando R como o valor correspondente à área sob a curva, segundo Hosmer e Lemeshow (2019) pode-se adotar a seguinte regra geral:

- Se $R = 0,5$, não há discriminação;
- Se $0,7 \leq R < 0,8$, a discriminação é aceitável;
- Se $0,8 \leq R < 0,9$, a discriminação é excelente;
- Se $R \geq 0,9$, a discriminação é excepcional.

3 Metodologia

3.1 Banco de Dados

Com o objetivo de entender de que maneira a evasão ocorre no curso de Ciência da Computação da Universidade de Brasília, foi realizado um estudo com dados extraídos do Sistema de Informações Acadêmicas de Graduação (SIGRA) e do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA), fornecidos pela universidade. O banco de dados contém informações acadêmicas e algumas sociodemográficas dos discentes do curso desde o primeiro semestre de 1991 até o segundo semestre de 2019. As variáveis presentes no banco de dados são:

1. Índice de Rendimento Acadêmico - IRA;
2. Gênero;
3. Data de Nascimento;
4. CEP;
5. UF de Nascimento;
6. Sistema de Cotas;
7. Tipos de Cota;
8. Escola;
9. Chamada de Ingresso na UnB;
10. Período de Ingresso na UnB;
11. Período de Ingresso no Curso;
12. Forma de Ingresso na UnB;
13. Período de Saída do Curso;
14. Forma de Saída do Curso;
15. Período que Cursou Disciplina;
16. Média Semestral do Aluno;
17. Mínimo de Créditos faltantes para Formatura;
18. Créditos no Período;
19. Total de Créditos Cursados;
20. Créditos Aprovados no Período;
21. Código da Disciplina;
22. Nome da Disciplina;
23. Créditos da Disciplina;
24. Menção na Disciplina.

O banco de dados original tem 123.126 observações e 27 variáveis, contendo informações do histórico dos estudantes. Cada linha representa uma disciplina cursada pelo estudante ao longo de sua trajetória acadêmica, refletindo assim a totalidade das disciplinas cursadas por ele durante o curso. Para manipular esses dados, foi necessário avaliar duplicações de informações, dado que as informações dos alunos se repetem de acordo com o histórico. Por exemplo, um aluno que cursou três disciplinas é registrado no banco três vezes, e assim por diante. Adicionalmente, ocorrem casos de replicação de históricos, em que alguns desses registros aparecem mais de uma vez.

Após a remoção dessas duplicidades e redundâncias, obteve-se um novo banco de dados com 86.345 observações. Nesse novo conjunto, foi aplicado um filtro para restringir o período de estudo (2012/1 até 2019/2), considerando as alterações no currículo do curso nesse intervalo. Essa filtragem resultou em um banco de dados reduzido, contendo 18.700 observações.

Posteriormente à limpeza dos dados, foi constituído o banco de dados utilizado no trabalho, onde cada linha representa um aluno único. Esse processo resultou em um banco com 764 alunos.

3.2 Criação de Variáveis

- **Evasão**

  Neste estudo, o principal objetivo é analisar a ocorrência de evasão acadêmica no curso de Ciência da Computação da Universidade de Brasília e identificar os fatores que contribuem para esse fenômeno.

  A variável "evasão" foi construída com base na variável "forma de saída do curso", considerando evasão como a saída definitiva do aluno do curso de origem sem concluí-lo, por qualquer motivo. Assim, a variável "evasão" é binária, com os valores "Não" indicando a ausência de evasão e "Sim" indicando a ocorrência de evasão do curso. A Tabela 1 oferece detalhes sobre a classificação das formas de saída do curso, utilizando o conceito de evasão. Vale ressaltar que apenas os alunos "ativos", ainda matriculados, e os "formados", que concluíram o curso, não foram categorizados como casos de evasão.

Tabela 1: Formas de saída do curso de Bacharelado em Ciência da Computação - UnB

Formas de saída	Evasão
Alunos ativos	Não
Formatura	
Reprovar 3x a mesma disciplina obrigatória	Sim
Desligamento - Não cumpriu condição	
Desligamento - Abandono	
Desligamento - Voluntário	
Desligamento - Decisão Judicial	
Mudança de Curso	
Novo vestibular	

- **Idade ao ingressar**

  Para a criação desta variável, foi utilizada a a variável "data de nascimento", considerando a idade em anos no momento do ingresso no curso. Como complemento a esse processo, empregou-se a variável "período de ingresso no curso" como referência, estabelecendo o dia 1º de março para alunos que iniciaram no primeiro semestre e o dia 1º de agosto para aqueles que ingressaram no segundo semestre.

- **Quantidade de ingressos**

  A variável "quantidade de ingressos" foi desenvolvida para contabilizar o número de vezes que um aluno ingressou no curso. No banco de dados, alguns alunos

evadiram por algum motivo e posteriormente reingressaram de diferentes formas, resultando em novas matrículas. Mesmo no caso da reintegração, onde a matrícula deve ser mantida, ocorreu esse problema. Diante dessa inconsistência, a análise da quantidade de ingressos foi realizada reunindo as matrículas associadas a um mesmo identificador do aluno.

- **Semestres cursados**

  A variável "semestres cursados" foi elaborada para determinar a quantidade de semestres que um aluno frequentou antes de sair do curso. Essa medida é obtida pela diferença entre as variáveis "período de saída do curso" e "período de entrada no curso". No caso dos alunos que permanecem ativos, o período de saída foi designado como 2019/2, correspondendo ao último semestre disponível no banco de dados. Para aqueles que saíram durante o período de verão, foi considerado o primeiro semestre subsequente como o "período de saída do curso".

- **Integralização**

  A variável "integralização" é uma métrica que calcula a porcentagem de créditos cursados pelo aluno em relação ao total mínimo necessário para a formatura. Essa análise abrange o período até a saída do aluno do curso. Para a construção dessa variável utilizou-se as variáveis "total de créditos cursados pelo aluno" e "mínimo de créditos para a formatura". Assim, a seguinte equação foi considerada como integralização:

$$\text{Integralização} = \frac{\text{Total de créditos cursados}}{\text{Mínimo de créditos para a formatura}} \tag{3.2.1}$$

  É importante observar que essa métrica tem a capacidade de exceder 100%, visto que um aluno pode cursar créditos além do mínimo estabelecido para a conclusão do curso.

- **Forma de ingresso**

  A variável "Forma de ingresso" no contexto do curso de Bacharelado em Ciência da Computação na Universidade de Brasília é uma categoria que já existe no banco de dados original e é inicialmente classificada de acordo com a Tabela 2.

Tabela 2: Distribuição dos alunos segundo Formas de Ingresso do curso de Bacharelado em Ciência da Computação - UnB, 2012-2019.

Formas de ingresso	Percentual
Vestibular	42,02%
Programa de Avaliação Seriada - PAS	30,63%
Sistema de Seleção Unificada - SISU	13,87%
Enem UnB	0,65%
Transferência Obrigatória	3,8%
Transferência Facultativa	3,8%
Portador de Diploma de Curso Superior	3,66%
Matrícula Cortesia	0,52%
Acordo Cultural-PEC-G	0,26%
Convênio-Int	0,65%
Convênio - Andifes	0,13%

Ao observar a Tabela 2, percebe-se a presença de categorias com baixas frequências ou semelhantes. Dessa forma, para simplificar a análise, realizou-se um agrupamento de algumas formas de ingresso, como apresentado na Tabela 3.

Tabela 3: Distribuição dos alunos segundo Formas de Ingresso do curso de Bacharelado em Ciência da Computação - UnB, 2012-2019. (Forma Agrupada)

Formas de ingresso	Percentual
Vestibular	42,02%
Programa de Avaliação Seriada - PAS	30,63%
ENEM	14,53%
Transferência	7,59%
Portador de Diploma de Curso Superior	3,66%
Convênios e Outros	1,57%

- **Quantidade de reprovações**

  A variável "quantidade de reprovações" foi desenvolvida a partir da categorização da variável "Menção na Disciplina" em dois grupos: Reprovado e Aprovado. O agrupamento pode ser visualizado na Tabela 4 abaixo:

Tabela 4: Resultado da disciplina

Menção	Resultado
SR	
II	Reprovado
MI	
MM	
MS	Aprovado
SS	

Essa variável representa o número de disciplinas com reprovação por aluno ao longo de seu percurso acadêmico. Essa análise é crucial para identificar padrões de desempenho e situações críticas que podem levar ao desligamento ou evasão do curso,

como, por exemplo, reprovar três vezes uma disciplina obrigatória.

- **Taxa de reprovação**

  A variável "taxa de reprovação" é uma métrica que calcula a proporção de reprovações acumuladas pelo aluno em relação ao total de créditos cursados. Essa avaliação engloba todo o período de sua permanência no curso. Para a construção dessa variável foi necessário utilizar a variável "resultado da disciplina" criada anteriormente. Assim, a seguinte equação foi considerada como taxa de reprovação:

$$\text{Taxa de reprovação} = \frac{\text{Total de créditos com reprovação}}{\text{Total de créditos cursados}} \tag{3.2.2}$$

  Nesse contexto, "Total de créditos com reprovação" representa a soma das decisões de menção classificadas como "Reprovado" por aluno. Além disso, a variável taxa de reprovação varia de 0 a 1, visto que os créditos com reprovação fazem parte do total de créditos cursados.

- **Quantidade de menções SR**

  A variável "quantidade de menções SR" visa quantificar o número de vezes que um aluno recebeu a menção SR (Sem Rendimento) em alguma disciplina. Essa métrica é relevante, pois a menção SR é atribuída no caso de abandono da disciplina, seja por não atingir o mínimo de presenças ou por obter nota zero. Assim, a quantidade de menções SR para cada aluno é calculada como a soma das vezes que essa menção foi atribuída.

  É fundamental destacar que consideramos abandono quando o número de faltas ultrapassa 25% do total de aulas da disciplina em questão. Dado que essa é a menor menção possível e, em muitos casos, resulta do abandono por parte do aluno, a análise busca compreender com qual frequência essa situação ocorre no ambiente acadêmico.

- **Quantidade de Trancamentos**

  Duas variáveis foram geradas com base nas informações sobre o número de solicitações de trancamento de disciplinas por parte dos alunos. Para essa análise, a definição de trancamento levou em conta as categorias TR e TJ da variável "menção de disciplina". Essas categorias são descritas da seguinte forma:

  1. **TR (Trancamento):** é um recurso que os alunos podem solicitar até a metade do semestre. Antigamente, realizar tal ação implicava na redução do Indice de Rendimento Acadêmico do aluno (IRA). Contudo, atualmente, essa prática não influencia mais no IRA, sendo apenas registrada no histórico acadêmico.

Entretanto, importante ressaltar que disciplinas obrigatórias não podem ser trancadas mais de uma vez;

2. **TJ (Trancamento Justificado):** ocorre quando o aluno solicita o trancamento da disciplina ou semestre mediante a apresentação de uma justificativa documentada aceita pela universidade.

- **Cursou Verão**

  Essa variável foi desenvolvida com o intuito de verificar se os alunos frequentaram alguma disciplina oferecida durante os semestres de verão ao longo de sua vida acadêmica. Para isso, a análise se baseou na variável "período que cursou a disciplina" para identificar os alunos que participaram de disciplinas durante os semestres de verão. A variável é dicotômica, atribuindo "Sim" àqueles que cursaram e "Não" àqueles que não cursaram disciplinas de verão. Essa atribuição é realizada para cada período que o estudante cursou.

- **Quantidade Verão**

  A variável "quantidade verão" foi desenvolvida para contabilizar o número de vezes que um aluno cursou algum semestre de verão. Para fazer essa contagem, foi utilizada como base variável "cursou verão".

- **Currículo**

  A variável "currículo" foi criada para determinar qual formato de currículo estava em vigor no momento do ingresso do aluno no curso. No período abrangido pela amostra do banco de dados, que compreende de 2012/1 a 2019/2, foram observadas algumas mudanças na configuração do currículo.

  A elaboração dessa variável teve como base a informação do "período de ingresso no curso" e sua categorização foi realizada da seguinte forma:

  - **Antigo:** ingressos de 2012/1 até 2015/1;
  - **Novo:** ingressos de 2015/2 até 2019/2.

- **Local de Residência**

  Para fins de identificação geográfica dos alunos presentes no estudo, foi criada a variável "Local de residência". Essa variável foi construída mediante a utilização da informação contida na variável "CEP" do banco de dados, sendo cruzada com a base nacional de CEPs. Esse processo de cruzamento possibilita a identificação das localidades de origem dos alunos, abrangendo tanto as regiões administrativas (RAs) do Distrito Federal quanto outras localizações.

Com o objetivo de captar um indicativo socioeconômico, utilizaram-se os dados da pesquisa sobre emprego e desemprego realizada pelo DIEESE para classificar as Regiões Administrativas (RAs) em quatro categorias, considerando padrões de renda média. Os dados da pesquisa podem ser verificados no link: https://www.dieese.org.br/analiseped/2018/201804pedbsb.html.

A classificação foi realizada da seguinte forma:

Tabela 5: Regiões administrativas segundo nível de renda, Distrito Federal - 2018

Nível de renda	Regiões Administrativas
Alta	Jardim Botânico, Lago Norte, Lago Sul, Plano Piloto, Park Way e Sudoeste/Octogonal
Média-alta	Águas Claras, Candangolândia, Cruzeiro, Gama, Guará, Núcleo Bandeirante, Sobradinho, Sobradinho II, Taguatinga e Vicente Pires
Média-baixa	Brazlândia, Ceilândia, Planaltina, Riacho Fundo, Riacho Fundo II, SIA, Samambaia, Santa Maria e São Sebastião
Baixa	Fercal, Itapoã, Paranoá, Recanto das Emas, SCIA – Estrutural e Varjão

Fonte: Dados sobre a pesquisa de emprego e desemprego 2018 - DIEESE

As cidades do estado de Goiás foram agrupadas de acordo com sua média de renda em 2018, conforme as categorias mencionadas anteriormente. Novo Gama, Luziânia, Cidade Ocidental e Águas Lindas de Goiás foram classificadas como pertencentes à faixa de baixa renda, enquanto Formosa foi enquadrada na categoria de média-baixa renda, já Goiânia foi designada como de média alta renda.

É importante observar que há registros de residência em outras cidades do Brasil. Nestes casos, não é possível categorizá-las com base nos níveis de renda mencionados.

Por meio dos dados coletados e criados, foram realizadas análises descritivas trançando o perfil dos estudantes para obter um panorama de como ocorre a evasão e quais são suas características. A partir dessas informações, foi possível construir dois modelos para identificar os fatores associados à evasão.

Como o estudo principal é baseado em verificar se o aluno evadiu ou não do curso (variável categórica binária), uma das metodologias mais adequada para se utilizar é o modelo de Regressão Logística.

Para a realização das etapas de preparação, análise descritiva e modelagem dos dados, foi utilizada a ferramenta computacional RStudio na versão 4.3.1.

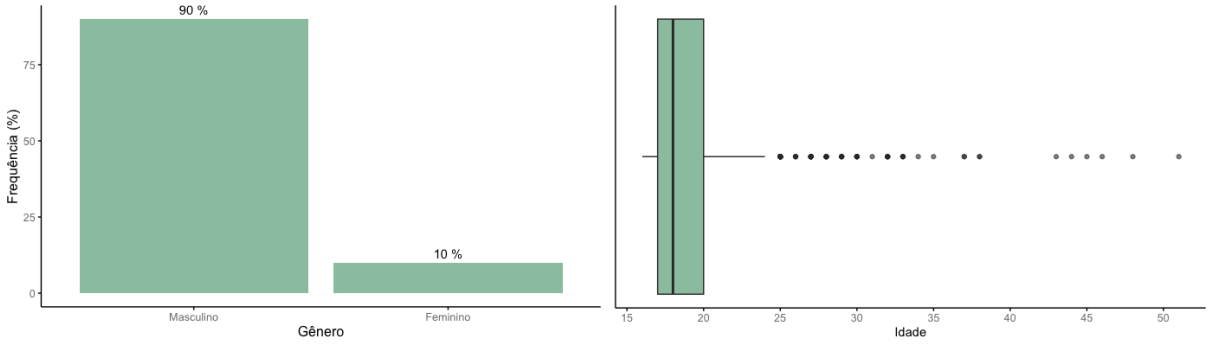
4 Resultados

4.1 Análise Descritiva

Antes de investigar os fatores relacionados à evasão no curso de Bacharelado em Ciência da Computação, é fundamental realizar uma análise descritiva para compreender o perfil dos estudantes, evidenciando suas características acadêmicas e sociodemográficas.

4.1.1 Dados Pessoais

Para traçar um perfil dos estudantes utilizou-se variáveis de dados pessoais, como gênero, idade, região administrativa do Distrito Federal onde os alunos residiam no momento do ingresso (e também no entorno do DF, Goiás) e seu nível de renda.



(a) (b)

Figura 6: Distribuição dos alunos segundo Gênero (a) e Idade (b).
Bacharelado em Ciência da Computação-UnB, 2012-2019.

Os resultados do estudo indicam que os alunos pertencentes à amostra são majoritariamente do gênero masculino (90%), sendo a representação feminina de apenas 10% da totalidade. Pode-se verificar que metade dos alunos tinham até 18 anos no momento do ingresso e que 75% tinham até 20 anos, sendo que a menor idade registrada é de 16 anos e a maior é de 51 anos. A média geral identificada é de 19,59 anos, com um desvio-padrão de 4,36 anos.

Sobre as Unidades Federativas (UF) onde nasceram os alunos matriculados no curso de Bacharelado em Ciência da Computação, nota-se que eles vêm de 26 lugares distintos. A maioria expressiva, cerca de 65%, nasceu no Distrito Federal. Minas Gerais, Goiás e Rio de Janeiro são os UF's subsequentes com maior representação de alunos. Além disso, há casos (2,5%) em que as informações sobre o local de nascimento não foram preenchidas.

Da totalidade dos alunos, 96,46% residiam no Distrito Federal (DF) no momento de sua matrícula na Universidade de Brasília. A Tabela 6 detalha a distribuição desses estudantes nas diversas regiões administrativas do DF.

Tabela 6: Distribuição dos alunos por Região Administrativa.
Bacharelado em Ciência da Computação-UnB, 2012-2019.

Período	Percentual	Período	Percentual
RA I - Plano Piloto	29.39%	RA XIV - São Sebastião	2,81%
RA II - Gama	1,13%	RA XV - Recanto das Emas	0,98%
RA III - Taguatinga	4,64%	RA XVI - Lago Sul	2,81%
RA IV - Brazlândia	0,84%	RA XVII - Riacho Fundo I	1,13%
RA V - Sobradinho	7,31%	RA XVIII - Lago Norte	4,64%
RA VI - Planaltina	1,12%	RA XIX - Candangolândia	0,56%
RA VII - Paranoá	0,56%	RA XX - Águas Claras	7,88%
RA VIII - Núcleo Bandeirante	1,55%	RA XXI - Riacho Fundo II	1,27%
RA IX - Ceilândia	4,36%	RA XXII - Sudoeste/Octogonal	4,78%
RA X - Guará	4,92%	RA XXIV - Park Way	1,41%
RA XI - Cruzeiro	2,25%	RA XXVII - Jardim Botânico	3,09%
RA XII - Samambaia	2,53%	RA XXVIII - Itapoã	0,56%
RA XIII - Santa Maria	1,69%	RA XXX - Vicente Pires	2,25%

Registra-se que quase 30% dos alunos vivem no Plano Piloto, que engloba as Asas sul e norte, o Setor Militar Urbano (SMU), o Noroeste, a Granja do Torto, a Vila Planalto e a Vila Telebrasilia. A segunda maior população (7.88%) reside em Águas Claras e o terceiro local de residência mais frequente (7.31%) é a RA V, de Sobradinho, que inclui Sobradinho, além da área do Colorado, Grande Colorado e Núcleo Rural Lago Oeste. Adicionalmente, ao considerar a soma das regiões do Plano Piloto com Lago Sul e Lago Norte, ambas situadas no centro de Brasília, observa-se que esse conjunto representa um percentual significativo de 36,84% do total de alunos.

Também há um reduzido registro de alunos que residem no estado de Goiás, em localidades consideradas como entorno do Distrito Federal. A Tabela 7 detalha a distribuição dos estudantes por cada uma dessas cidades.

Tabela 7: Distribuição dos alunos residentes no Goiás por Cidade.
Bacharelado em Ciência da Computação-UnB, 2012-2019.

Cidade	Frequência
Águas Lindas de Goiás	1
Cidade Ocidental	1
Formosa	2
Luziânia	2
Novo Gama	3

Além dos estudantes que residem no Distrito Federal e no estado de Goiás, identificou-se alunos que, ao ingressarem na UnB, eram provenientes de outros estados brasileiros, totalizando 2,24%, juntamente com 5 alunos para os quais não foi possível

identificar o local de residência devido a problemas nas informações relacionadas ao CEP.

Ao analisar as regiões administrativas e cidades com uma perspectiva socioeconômica, conforme apresentado na Tabela 5, verificou-se que os alunos do curso de Ciência da Computação são majoritariamente indivíduos que residem em áreas de renda alta. Pôde-se verificar que aproximadamente 81% da amostra tem nível de renda alto (alto e média-alta), e que ao redor de 19% dos alunos são provenientes de famílias de baixa renda (média-baixa ou baixa).

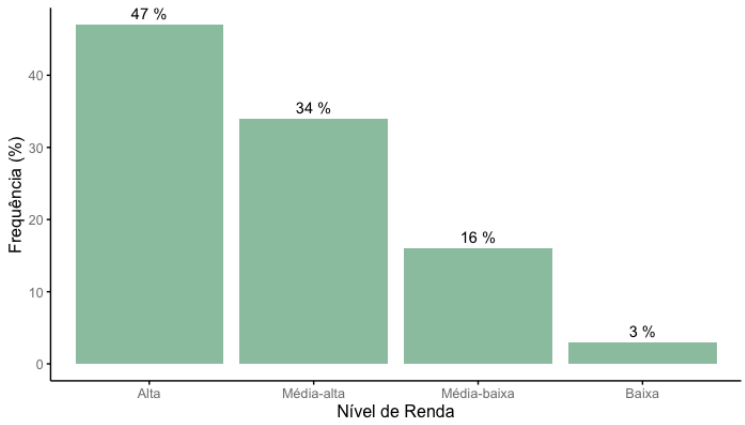
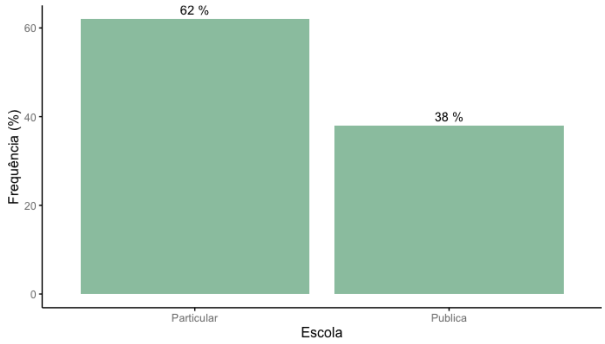


Figura 7: Distribuição dos alunos segundo Nível de Renda. Bacharelado em Ciência da Computação-UnB, 2012-2019.

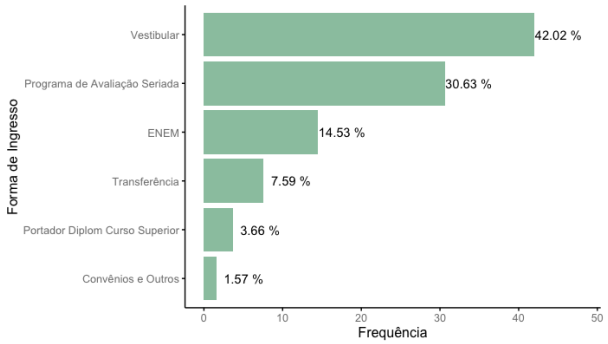
4.1.2 Ingresso na Universidade

Para analisar os aspectos da evasão de alunos no curso também é importante conhecer informações sobre o ingresso dos alunos na UnB. Dessa forma, utilizou-se dados sobre o período de ingresso, a origem da escola de Ensino Médio, a forma de ingresso na universidade e se o aluno é cotista ou não.

(a)



(b)



(c)

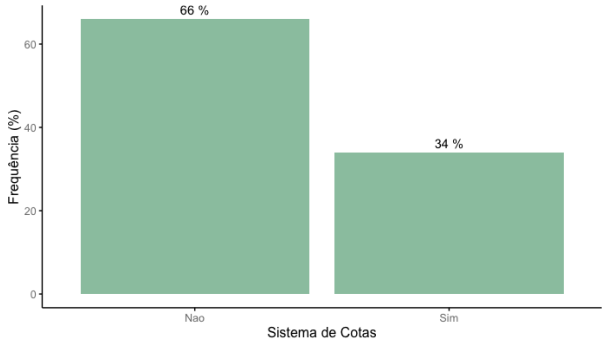
Figura 8: Distribuição dos alunos segundo Período de Ingresso (a), Escola (b) e Forma de Ingresso (c). Bacharelado em Computação-UnB, 2012-2019.

Na Figura 8a, é evidente que os semestres 2012/2, 2014/1, 2015/1, 2016/2 e 2017/2 apresentam uma menor quantidade de alunos ingressantes, variando entre 42 e 44 alunos, em comparação com os demais semestres, nos quais a entrada de alunos oscila entre 48 e 54.

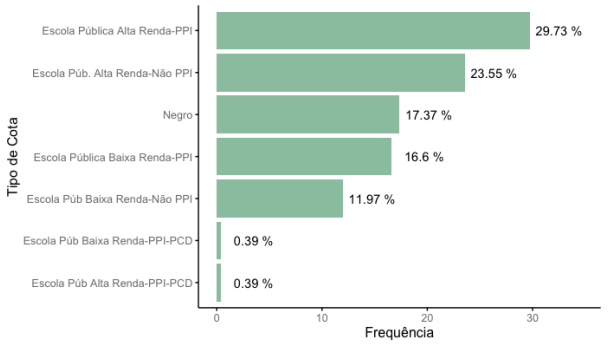
Ao analisar a origem educacional dos discentes, a Figura 8b, destaca-se os alunos são majoritariamente provenientes de escolas privadas (62%), comparados ao quantitativo de alunos oriundos de escolas públicas (38%).

A Figura 8c traz informações sobre os diferentes formas de ingresso na universidade, onde destacam-se as provas de seleção, como Vestibular, Programa de Avaliação Seriada (PAS) e Exame Nacional do Ensino Médio (ENEM), que foram a forma de entrada de 87.18% dos alunos.

Ainda tratando a informação sobre o ingresso dos alunos na Universidade, foram trazidos dados sobre ingressos por meio do sistema de cotas, bem como detalhes sobre este processo.



(a)



(b)

Figura 9: Distribuição dos alunos segundo Sistema de Cotas (a) e Tipo de Cota (b). Bacharelado em Ciência da Computação-UnB, 2012-2019.

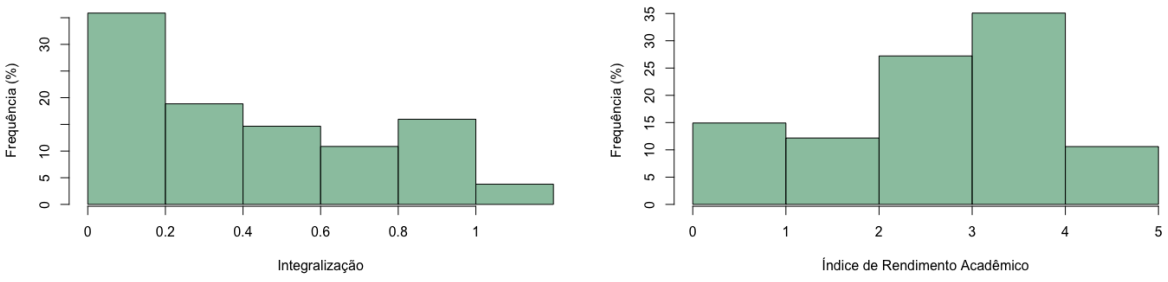
A Figura 9a indica que 66% dos alunos não ingressaram por meio do sistema de cotas, enquanto 34% sim. Registra-se que os tipos de cotas mais frequentes são aquelas destinadas a alunos provenientes de escolas públicas e de alta renda, alcançando um total de 53,28%. A terceira mais frequente é a cota destinada para estudantes negros (17,37%).

Para uma melhor compreensão dos tipos de cotas, é importante esclarecer algumas definições relacionadas a cada uma delas. Alunos classificados como de alta renda são aqueles cuja renda bruta per capita ultrapassa 1,5 salários mínimos, enquanto os de baixa renda têm renda bruta inferior a esse valor. Os alunos PPI são aqueles que se autodeclaram pretos, pardos ou indígenas. Além disso, existem cotas específicas para pessoas com deficiência (PCD) e uma cota destinada a candidatos que se autodeclaram negros.

4.1.3 Vida Acadêmica

No intuito de compreender a vivência acadêmica dos alunos do Bacharelado em Ciência da Computação da UnB no período selecionado, o estudo buscou verificar mais elementos sobre o rendimento dos alunos durante a vida acadêmica, uma vez que pode estar diretamente vinculado aos fatores de evasão.

Para ter uma visão geral dos estudantes, foram consideradas duas variáveis que representam o percurso e o desempenho acadêmico: a integralização e o IRA (Índice de Rendimento Acadêmico).



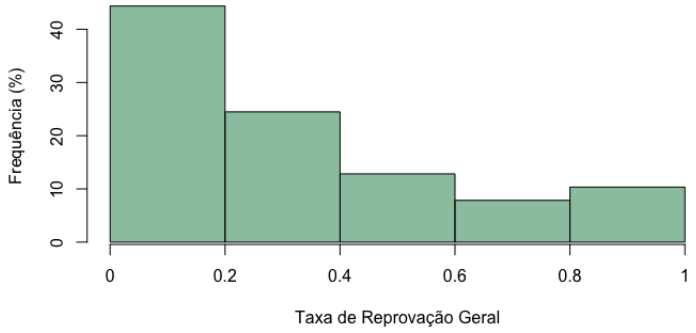
(a) (b)

Figura 10: Distribuição dos alunos segundo Integralização (a) e IRA (b).
Bacharelado em Ciência da Computação-UnB, 2012-2019.

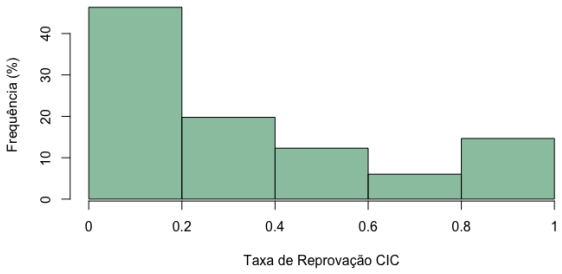
A integralização reflete a proporção do curso que o aluno completou até sua saída, sendo o cálculo detalhado no Tópico 3.2.1. Observando a Figura 10, nota-se que a maioria dos alunos cursou entre 0 e 20% do curso, sendo que uma pequena proporção ultrapassou os créditos mínimos exigidos para a formatura.

No que diz respeito ao IRA, as notas de rendimento dos alunos concentram-se entre 2 e 4. É importante ressaltar que a frequência de IRA entre 0 e 1 é alta, o que é um fator preocupante.

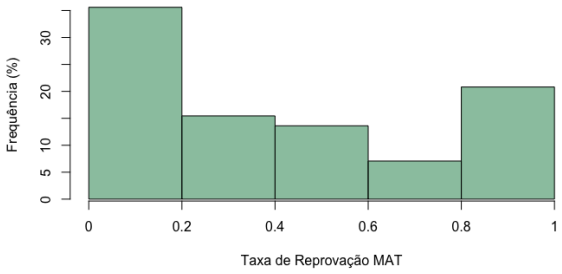
Visando uma compreensão mais abrangente do histórico acadêmico dos alunos, foram analisadas diversas variáveis, como taxas de reprovação, abandono em disciplinas e trancamentos.



(a)



(b)



(c)

Figura 11: Distribuição dos alunos segundo Taxa de Reprovação - Geral (a), Ciência da Computação (b) e Matemática (c). Bacharelado em Ciência da Computação-UnB, 2012-2019.

Ao analisar a taxa de reprovação geral apresentada na Figura 11a, percebe-se que as maiores frequências estão concentradas entre 0 e 20%. No entanto, chama a atenção uma frequência alta entre 0,8 e 1, indicando que há alunos que enfrentaram reprovação em mais de 80% dos créditos cursados.

Por meio de uma análise prévia, fundamentada no currículo do curso de Bacharelado em Ciência da Computação, identificou-se que os períodos inciais do curso são bastante voltados para disciplinas da Ciência da Computação e da Matemática. Diante disso, também foram analisadas as taxas de reprovação nas disciplinas de cada um desses

cursos (Figuras 11b e 11c). Nota-se que ambas exibem padrões semelhantes à taxa de reprovação geral, contudo, nas disciplinas de Matemática, as reprovações entre 80% e 100% destacam-se de maneira expressiva.

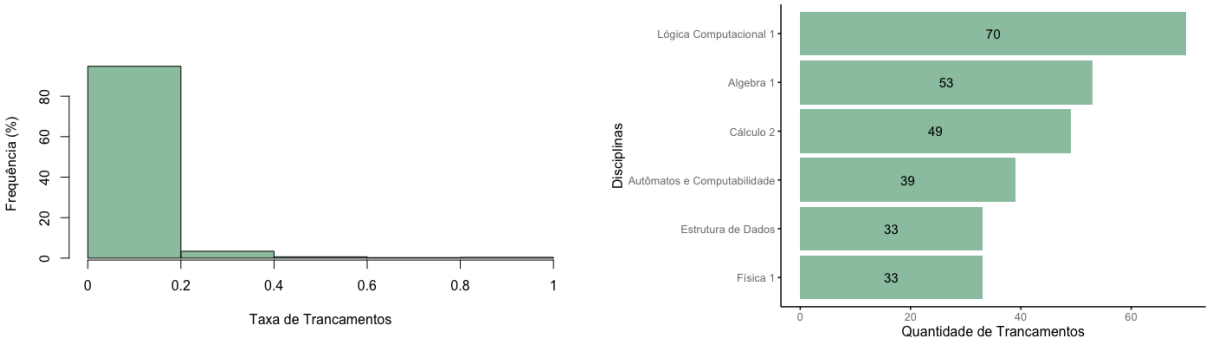
Além da avaliação das taxas de reprovação, também foram examinados os dados de ausência de rendimento (representada pela menção SR), uma vez que essa categoria sinaliza um fator de abandono das disciplinas.

(a) Taxa de menções SR (b) Disciplinas com maior atribuição de menção SR

Figura 12: Distribuição dos alunos segundo Menções SR. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Ao observar a Figura 12a, é notável que a taxa de SR se concentra predominantemente entre 0 e 20%. No contexto da Figura 12b, observa-se que três das cinco disciplinas com os maiores índices de abandono pertencem ao curso de Ciência da Computação. Importante ressaltar que todas as disciplinas apresentadas na Figura 12b referem-se aos dois primeiros semestres do curso.

Outra modalidade de abandono de disciplinas é representada pelo trancamento, uma prática observada com considerável frequência.

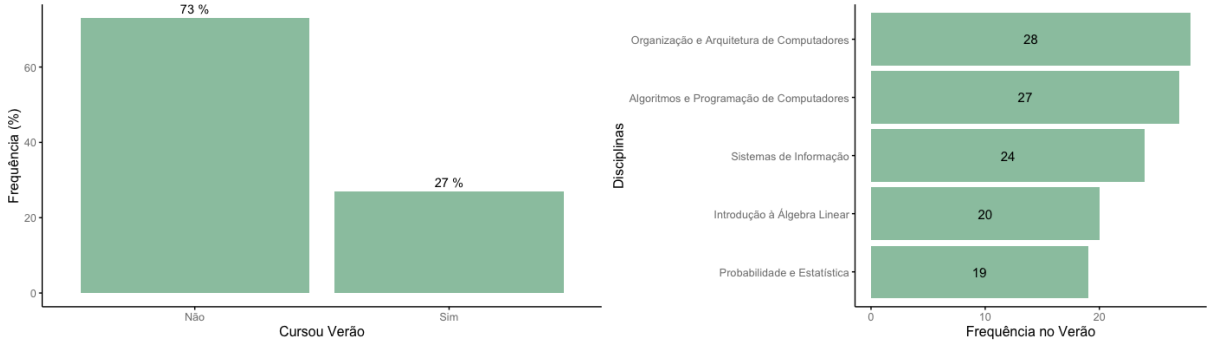


(b) Taxa de Trancamento (c) Disciplinas mais trancadas

Figura 13: Distribuição dos alunos segundo Trancamento. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Ao analisar a Figura 13a, é evidente que a taxa de trancamento concentra-se predominantemente entre os valores de 0 e 0,2. Já observando a Figura 13b, nota-se que a maioria das disciplinas que os alunos trancam são dos cursos de Ciência da Computação e Matemática.

Outra característica relevante é verificar se o aluno já cursou alguma matéria durante um semestre de verão. A UnB disponibiliza disciplinas específicas nesse período para atender às necessidades acadêmicas dos alunos, proporcionando-lhes a oportunidade de recuperar disciplinas não concluídas ou reprovadas.



(b) Cursou verão (c) Disciplinas mais cursadas no verão

Figura 14: Distribuição dos alunos segundo Cursou Verão. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Por meio da Figura 14, é possível constatar que 73% dos alunos optaram por não cursar semestres de verão, enquanto 27% escolheram cursá-los. Quanto às disciplinas mais cursadas durante esse período, predominam aquelas associadas ao curso de Ciência da Computação.

Outro elemento interessante a ser registrado é a quantidade de ingressos dos alunos no curso. Ainda que a massiva maioria dos alunos tenha ingressado formalmente apenas uma vez, há registros de discentes que foram reintegrados até 7 vezes no mesmo curso.

Tabela 8: Distribuição dos alunos segundo Quantidade de Ingressos. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Ingressos	Frequência
1	735
2	24
3	4
7	1

4.1.4 Saída do curso

A variável "Evasão", detalhada na Seção 3.2, foi criada para mensurar a proporção de alunos que abandonam o curso de Bacharelado em Ciência da Computação.

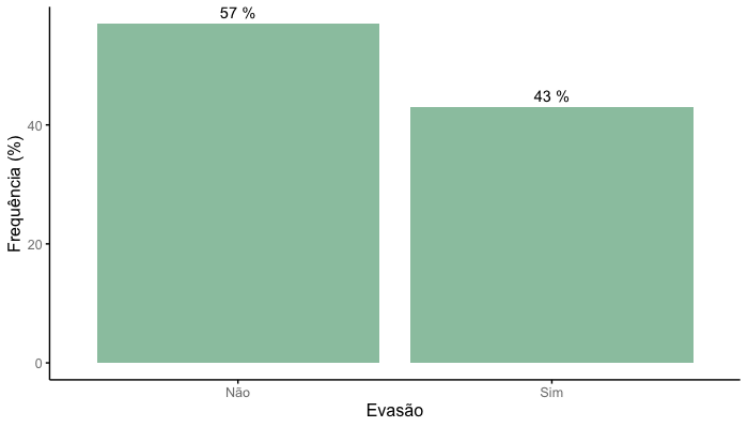


Figura 15: Distribuição dos alunos segundo evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019

Com base na Figura 15, verifica-se que 57% dos alunos não evadiram, indicando que estavam ativos até o semestre 2019/2 ou concluíram o curso. Por outro lado, os alunos que evadiram representam 43% da amostra. Para uma compreensão mais aprofundada da evasão, foram detalhadas as formas de saída na Tabela 9.

Tabela 9: Distribuição dos alunos segundo Formas de saída do curso de Bacharelado em Ciência da Computação-UnB, 2012-2019

Formas de Saída	Frequência	Percentual
Alunos ativos	367	48,04%
Formatura	68	8,9%
Desligamento - Não cumpriu condição	138	18,06%
Reprovar 3x a mesma disciplina obrigatória	27	3,53%
Desligamento por força de intercâmbio	5	0,65%
Desligamento - Força de Convênio	1	0,13%
Desligamento - Abandono	64	8,38%
Desligamento Voluntário	23	3,01%
Novo vestibular	59	7,72%
Mudança de Curso	6	0,79%
Mudança de Turno	1	0,13%
Transferência	4	0,52%
Falecimento	1	0,13%

Quanto às formas consideradas como evasão, 18,06% foram desligados por não atenderem a condições acadêmicas, 8,38% abandonaram o curso, e 3,01% optaram por desligamento voluntário. Além disso, 7,72% ingressaram por meio de novo vestibular, enquanto 3,53% dos alunos reprovaram três vezes a mesma disciplina obrigatória.

4.2 Análise Bivariada

De forma a abordar os objetivos deste trabalho, que consiste em avaliar os fatores que influenciam na evasão ou não dos estudantes, a análise bivariada se destaca como uma visão inicial para compreender o comportamento da evasão em relação a diferentes características dos alunos.

4.2.1 Evasão

A Tabela 10 apresenta informações sobre a evasão e características individuais e socioeconômica dos alunos, juntamente com os resultados do Teste Qui-Quadrado de Pearson, incluindo a estatística correspondente e seu respectivo p-valor. Estes valores são apresentados para verificar a associação entre essas características e a evasão.

Tabela 10: Análise bivariada por evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Variável		Evasão		Estatística	P-valor
		Sim	Não		
Gênero	Feminino	39%	61%	0,3656	0,5454
	Masculino	44%	56%		
Tipo de Escola	Particular	43%	57%	0,0810	0,7759
	Pública	44%	56%		
Forma de Ingresso	Vestibular	46%	54%	26,454	$< 0,0001$
	PAS	35%	65%		
	ENEM	37%	63%		
	Transferência	52%	48%		
	Portador de Diploma	68%	32%		
	Convênios e Outros	83%	17%		
Sistema de Cotas	Não	44%	56%	0,6034	0,4373
	Sim	41%	59%		
Tipo de Cota	Negro	56%	44%	11,354	0,1239
	Escola Pública Alta Renda-PPI	54%	46%		
	Escola Púb. Alta Renda-Não PPI	30%	70%		
	Escola Pública Baixa Renda-PPI	37%	63%		
	Escola Púb Baixa Renda-Não PPI	39%	61%		
Local de Residência	Baixa renda	52%	48%	2,092	0,5535
	Média baixa renda	38%	62%		
	Média alta renda	44%	56%		
	Alta renda	44%	56%		
Cursou verão	Não	51%	49%	46,633	$< 0,0001$
	Sim	23%	77%		
Currículo	Antigo	51%	49%	39,967	$< 0,0001$
	Novo	26%	74%		

No que diz respeito ao gênero, nota-se que 39% das mulheres e 44% dos homens saíram do curso, resultando em taxas de permanência de 61% e 56%, respectivamente. Em relação à modalidade de ingresso, os alunos admitidos por convênios e outras formas apresentaram a maior taxa de evasão (83%), embora seja crucial destacar que a incidência desses casos seja reduzida, envolvendo apenas 12 alunos. Por outro lado, os ingressantes pelo Programa de Avaliação Seriada (PAS) registraram a menor taxa de evasão (35%).

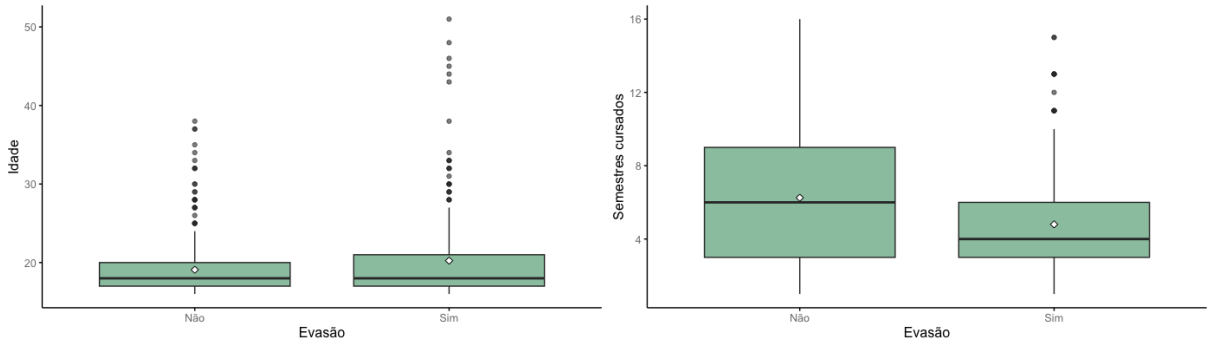
Quanto ao sistema de cotas, foram registrados índices de 44% para os não beneficiários de cotas e 41% para os beneficiários. Dentro das categorias de cotas, os estudantes inseridos no grupo "Escola Pública Alta Renda - Não PPI" apresentaram a menor taxa de evasão (30%), enquanto aqueles que participaram do regime de cota para pessoas negras registraram a taxa mais elevada (56%).

Ao analisar a localidade de residência, observa-se que alunos de baixa renda apresentaram uma taxa de evasão mais elevada (52%), enquanto aqueles de renda média alta e alta demonstraram taxas equivalentes (44%). A frequência de alunos cursando aulas de verão e o tipo de currículo também exerceram influência na evasão, com taxas mais baixas para os alunos que realizaram matérias de verão (23%) e para os adeptos do currículo novo (26%).

Após a análise descritiva bivariada das variáveis, o próximo passo essencial é identificar possíveis fatores explicativos para o modelo em estudo. Inicialmente, cada variável foi submetida a um teste com o objetivo de avaliar sua associação e impacto na evasão, que é o foco central desta pesquisa. Para essas variáveis qualitativas, a avaliação da associação foi conduzida por meio do teste Qui-Quadrado de Pearson.

Os resultados do teste na Tabela 10 revelam que, ao considerar um nível de significância de 5%, não foi observada uma associação significativa entre a evasão e as variáveis gênero, escola, sistema de cotas e local de residência. A variável "Tipo de Cota" apresentou um p-valor próximo a 10%, mas não atingiu um patamar suficientemente baixo para ser considerada estatisticamente significativa. Embora não tenha sido identificada uma associação forte, tanto essas variáveis quanto as demais serão incorporadas nos modelos a serem testados, com exceção da variável "Tipo de Cota", devido à significativa quantidade de valores ausentes.

Já para analisar as variáveis quantitativas, foi utilizada uma abordagem para compreender o comportamento em relação à variável evasão, a qual é feita por meio de gráficos *boxplot*. A Figura 16 exibe a representação gráfica das variáveis idade ao ingressar e semestres cursados em relação à variável evasão.



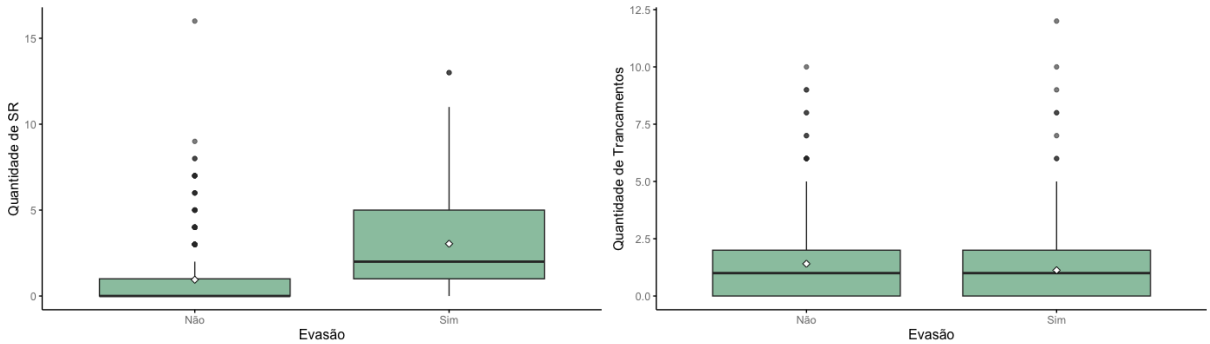
(a) (b)

Figura 16: Distribuição dos alunos segundo Idade ao ingressar (a) e Semestres cursados (b) em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Na análise da idade em relação à evasão (Figura 16a), observa-se que o comportamento dos alunos é bastante semelhante, com medianas iguais (18 anos). No entanto, é possível notar que, para os alunos que evadem, a idade é um pouco maior até o percentil 75%, e seu valor máximo atinge 51 anos.

Em relação à 16b, observa-se que a distribuição dos semestres cursados por alunos que evadiram concentra-se em valores mais baixos, apontando para uma saída do curso antes da formatura. A mediana desse grupo é de 4 semestres, com 75% dessa população não ultrapassando 6 semestres. Em contraste, para os não evadidos, a distribuição é mais ampla, indicando a presença de alunos ativos, mas muitos ultrapassam 9 semestres, que é o tempo padrão ou esperado para a conclusão do curso.

Tendo em vista uma visão abrangente sobre o desempenho acadêmico em relação à evasão, as Figuras 17 e 18 apresentam a representação gráfica de algumas variáveis pertinentes.

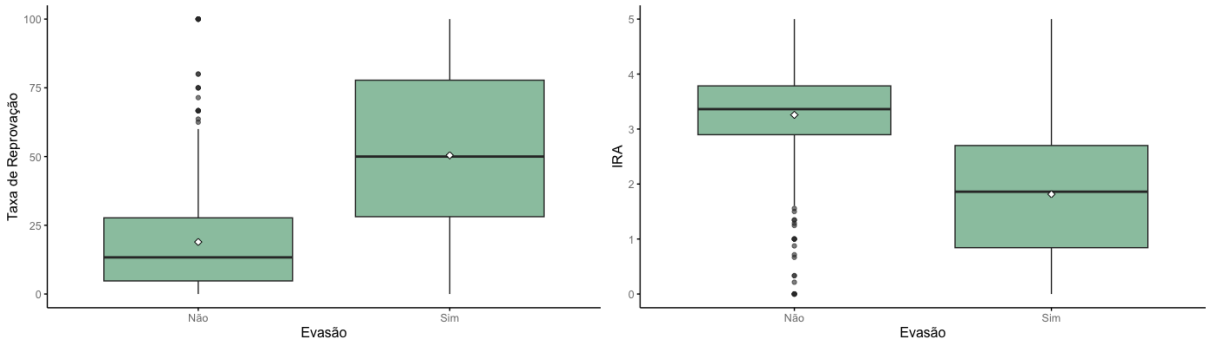


(a) (b)

Figura 17: Distribuição dos alunos segundo Quantidade de Menções SR (a) e Quantidade de Trancamentos (b)em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Analisando a evasão em relação às variáveis "Quantidade de menções SR" e "Número de trancamentos", nota-se que, no que diz respeito à quantidade de menções SR, os alunos que evadiram apresentam uma frequência consideravelmente maior em comparação aos alunos que não evadiram. Quanto ao número de trancamentos, observa-se um comportamento bastante semelhante entre os alunos, com medianas iguais (1 trancamento), embora os alunos que evadiram apresentem um número máximo de trancamentos maior.

Na análise das variáveis "Taxa de Reprovação" e "IRA", destaca-se um notável contraste no comportamento entre alunos que evadiram e aqueles que não evadiram.



(a) (b)

Figura 18: Distribuição dos alunos segundo Taxa de Reprovação (a) e IRA (b) em relação à Evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019.

No contexto da taxa de reprovação, a mediana para os não evadidos situa-se em torno de 0,13, em contraste com os evadidos, cuja mediana atinge 0,50. Já em relação ao Índice de Rendimento Acadêmico (IRA), observa-se uma relação inversa com a taxa de reprovação. Os alunos não evadidos exibem um IRA superior em comparação aos evadidos, uma tendência esperada.

Identificam-se como *outliers* os valores relacionados aos não evadidos que possuem IRAs baixos ou taxa de reprovação alta, decorrentes de múltiplas reprovações, mas que não evadiram. Ao analisar o *boxplot* da Figura 18b, percebe-se que discrepância da evasão na variável "IRA" parece ser significativa, destacando-se que a mediana para não evadidos é de 3,36, enquanto para evadidos é de 1,86.

Após a análise descritiva dessas variáveis quantitativas, também foi realizado um teste de associação conduzido por meio de uma regressão logística simples.

Tabela 11: Associação das variáveis com evasão. Bacharelado em Ciência da Computação-UnB, 2012-2019

Variável	Estatística do teste	P-valor
Idade	3,484	0,0005
Semestres cursados	-5,462	$< 0{,}0001$
Quantidade de trancamentos	-2,202	0,0277
Quantidade de SR	10,156	$< 0{,}0001$
Taxa de reprovação	12,46	$< 0{,}0001$
Taxa de reprovação CIC	12,53	$< 0{,}0001$
Taxa de reprovação MAT	12,30	$< 0{,}0001$
IRA	-13,26	$< 0{,}0001$

Observa-se que, ao considerar um nível de significância de 5%, todas as variáveis demonstram associação com a evasão. No entanto, ao adotar um critério mais rigoroso de 10%, a variável "Quantidade de trancamentos" é rejeitada.

4.2.2 Correlação entre variáveis

Explorar a correlação entre variáveis é essencial para contornar os desafios decorrentes da multicolinearidade nos modelos de regressão. A existência de relações significativas entre as variáveis pode comprometer a construção do modelo. Portanto, nesta seção, será apresentado uma análise de correlação entre algumas variáveis do banco de dados.

- **IRA e Taxa de Reprovação**

  Ao analisar as variáveis "IRA" (Índice de Rendimento Acadêmico) e "Taxa de Reprovação", ambas medidas quantitativas do desempenho do aluno, torna-se crucial verificar a correlação entre elas, dado que ambas refletem conceitos semelhantes.

  Para essa análise de correlação, utilizou-se o coeficiente de correlação de Pearson (ρ), o qual apresentou um valor de -0,8512. Este resultado indica uma correlação linear entre as variáveis "IRA" e "Taxa de Reprovação", sendo esta uma correlação forte e negativa, sugerindo uma relação inversamente proporcional entre os dois indicadores acadêmicos. A representação visual dessa análise está ilustrada na Figura abaixo.

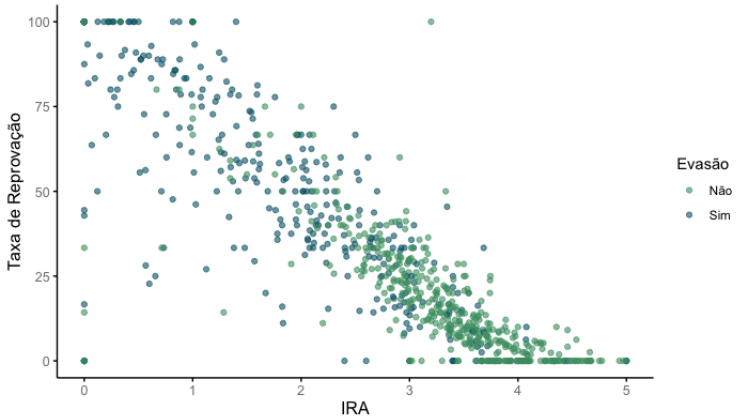
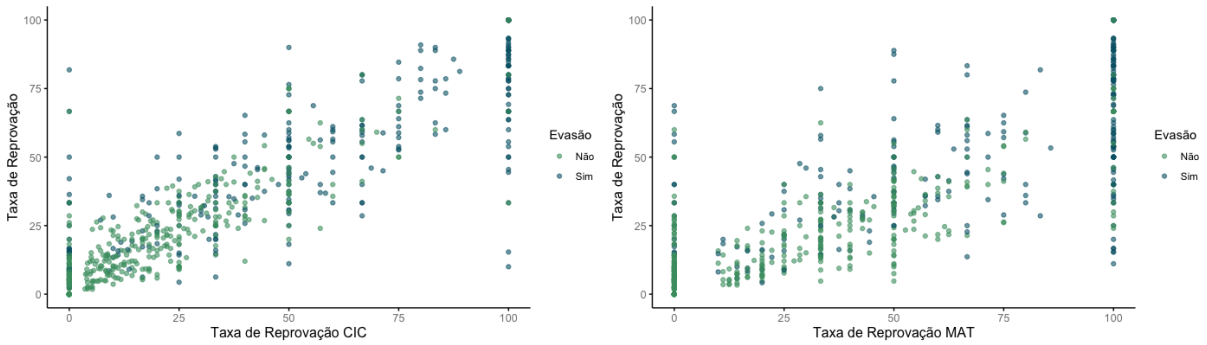


Figura 19: Correlação entre IRA e Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019

- **Taxas de Reprovação (Geral, CIC e MAT)**

  Após a constatação da correlação entre o Índice de Rendimento Acadêmico (IRA) e a Taxa de Reprovação, procedeu-se à verificação da correlação entre as taxas de reprovação geral, da Ciência da Computação e da Matemática, uma vez que a Figura 11 evidenciou comportamentos semelhantes entre essas variáveis, sendo todas elas medidas quantitativas do desempenho acadêmico.

  Da mesma forma que na análise anterior, empregou-se o coeficiente de correlação de Pearson (ρ) para avaliar essas relações. Os resultados revelaram correlações fortes e positivas, sendo o coeficiente de correlação entre a Taxa Geral e a Taxa CIC igual a 0,9046, e entre a Taxa Geral e a Taxa MAT igual a 0,8228. A representação visual dessa análise está ilustrada na Figura abaixo.



(a) (b)

Figura 20: Correlação entre a Taxas de Reprovação Geral com as Taxas de Reprovação CIC (a) e MAT (b). Bacharelado em Ciência da Computação-UnB, 2012-2019

- **Sistema de Cotas e Tipo de Escola**

  Ao analisar as variáveis "Sistema de Cotas" e "Tipo de Escola" surge a suspeita de uma possível correlação, dado que muitos sistemas de cotas são direcionados a alunos de escolas públicas. Para testar essa associação, foi empregado o teste qui-quadrado de Pearson, utilizando a tabela de contingência (Tabela 12) e com base nas hipóteses abaixo.

Tabela 12: Distribuição dos alunos segundo Sistema de Cotas e Tipo de Escola. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Variável		Escola	
		Particular	Pública
Sistema de Cotas	Não	91,49%	25,51%
	Sim	8,51%	74,49%

$$\begin{cases} H_0 : \text{Não existe associação entre as variáveis} \\ H_1 : \text{Existe associação entre as variáveis} \end{cases}$$

Ao realizar o teste, obteve-se um p-valor menor que 0,0001, indicando que há fundamentos para rejeitar a hipótese nula. Portanto, há evidências estatísticas para concluir que as variáveis "Sistema de Cotas" e "Tipo de Escola" estão relacionadas.

4.3 Modelagem

Com base na seção anterior, procedeu-se o desenvolvimento de dois modelos independentes para analisar o banco de dados. Um desses modelos foi direcionado à variável Índice de Rendimento Acadêmico (IRA), enquanto o segundo foi elaborado considerando a Taxa de Reprovação. Essa abordagem específica foi escolhida devido à evidente correlação e relevância dessas variáveis no contexto da pesquisa.

Conforme destacado na Seção 4.2.2, as variáveis de taxa de reprovação (geral, CIC e MAT) apresentam uma forte correlação e, por consequência, também podem estar relacionadas ao IRA. Diante desse cenário, torna-se inviável incluí-las simultaneamente nos modelos. Assim, foi necessário optar por apenas uma dessas variáveis para representar o modelo de Taxa de Reprovação. Após testar os modelos com cada uma das variáveis, constatou-se que a escolha mais apropriada para o modelo utilizado foi a variável Taxa de Reprovação MAT.

É importante ressaltar que devido à baixa quantidade de observações, a variável "Forma de Ingresso" passou por um agrupamento de categorias. Nesse contexto, as in-

formações referentes a "Portador de Diploma" e "Convênios e Outros" foram consolidadas como uma única categoria denominada "Outros". No que diz respeito à variável "Local de Residência", as categorias foram agrupadas da seguinte maneira: "Baixa renda" e "Média-baixa renda" foram combinadas em "Baixa", enquanto "Média-alta renda" e "Alta renda" foram designadas como "Alta".

Para iniciar o processo de modelagem, o banco de dados foi dividido em duas amostras distintas: uma destinada à construção dos modelos e outra para fins de validação. Essa divisão tem como objetivo primordial avaliar a capacidade de generalização dos modelos, permitindo a comparação das estimativas resultantes.

Tabela 13: Teste inicial com a base de construção para todas variáveis do modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019

Variável	Estimativa	Desvio Padrão	Estatística	P-valor
Intercepto	4,7711	1,1508	4,1460	$< 0,0001$
IRA	-1,1582	0,1817	-6,3753	$< 0,0001$
Gênero				
Masculino	-0,2970	0,3990	-0,7444	0,4567
Feminino				
Escola				
Pública	0,2322	0,3572	0,6502	0,5156
Particular				
Forma de Ingresso				
Vestibular	0,0189	0,4513	0,0419	0,9666
PAS	-0,9265	0,5040	-1,8384	0,0660
ENEM	-0,6474	0,5309	-1,2194	0,2227
Outros				
Sistema de Cotas				
Sim	-0,2125	0,3799	-0,5593	0,5759
Não				
Local de Residência				
Baixa	-0,4175	0,3721	-1,1220	0,2618
Alta				
Cursou verão				
Sim	-1,2965	0,3217	-4,0306	$< 0,0001$
Não				
Currículo				
Novo	-2,1390	0,3603	-5,9371	$< 0,0001$
Antigo				
Idade	0,0029	0,0352	0,0816	0,9350
Semestres cursados	-0,1006	0,0457	-2,2015	0,0277
Quantidade de trancamentos	-0,1389	0,0888	-1,5654	0,1175
Quantidade de SR	0,1547	0,0752	2,0572	0,0397

Tabela 14: Teste inicial com a base de construção para todas variáveis do modelo com Taxa de Reprovação MAT. Bacharelado em Ciência da Computação-UnB, 2012-2019

Variável	Estimativa	Desvio Padrão	Estatística	P-valor
Intercepto	1,1836	1,1660	1,0150	0,3101
Taxa de Reprovação MAT	0,0295	0,0046	6,4067	$< 0{,}0001$
Gênero				
Masculino	-0,3410	0,4310	-0,7913	0,4288
Feminino				
Escola				
Pública	0,1527	0,4096	0,3729	0,7092
Particular				
Forma de Ingresso				
Vestibular	-0,1364	0,5076	-0,2687	0,7882
PAS	-1,3285	0,5549	-2,3941	0,0167
ENEM	-1,3300	0,5714	-2,3277	0,0199
Outros				
Sistema de Cotas				
Sim	-0,3557	0,4341	-0,8194	0,4125
Não				
Local de Residência				
Baixa	-0,3253	0,4025	-0,8081	0,4190
Alta				
Cursou verão				
Sim	-1,4552	0,3400	-4,2797	$< 0{,}0001$
Não				
Currículo				
Novo	-2,2474	0,3956	-5,6812	$< 0{,}0001$
Antigo				
Idade	0,0023	0,0394	0,0585	0,9533
Semestres cursados	-0,2112	0,0469	-4,5019	$< 0{,}0001$
Quantidade de trancamentos	-0,0537	0,0937	-0,5729	0,5667
Quantidade de SR	0,2915	0,0712	4,0936	$< 0{,}0001$

Ao analisar os resultados alcançados nos testes iniciais (Tabelas 10 e 11), percebe-se que, em ambos os modelos, certas variáveis que demonstraram significância isoladamente não mantiveram essa relevância ao serem integradas ao modelo junto às demais. Por outro lado, em certos casos, ocorreu a situação oposta, em que variáveis inicialmente não significantes de forma isolada, passaram a ter importância ao serem consideradas no contexto do modelo.

- **Modelo com a variável IRA**

Para obter um modelo mais simplificado para a variável IRA, recorreu-se ao método de seleção *stepwise*, conforme descrito na Seção 2.4.1. Contudo, percebeu-se que somente esse método não foi suficiente para a seleção apropriada das variáveis, dada a presença de correlações com a variável IRA. Diante desse cenário, as variáveis "Forma de Ingresso" e "Quantidade de SR" foram excluídas do modelo. Além disso, a variável "Semestres cursados" também foi removida, pois, na amostra de construção o p-valor estava muito próximo do limite de 5%, levantando dúvidas sobre sua inclusão no modelo. Ao examinar a amostra de validação, constatou-se um p-valor próximo a 0,8, fornecendo mais indícios de que a variável deveria ser removida.

Adicionalmente, foram realizadas tentativas para incluir interações entre o IRA e a quantidade de semestres cursados, bem como entre o IRA e o sistema de cotas. No entanto, observou-se que essas interações acabaram comprometendo a estabilidade do modelo e, consequentemente, não foram incorporadas.

Após essas etapas, o modelo reduzido inclui as seguintes variáveis: IRA, cursou verão e currículo.

Tabela 15: Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019

Parâmetro	Estimativa-Construção	Estimativa-Validação	Estimativa-Geral
Intercepto	4,24	3,79	4,10
IRA	-1,42	-1,31	-1,38
Cursou Verão - Sim	-1,41	-1,11	-1,33
Currículo - Novo	-1,73	-1,20	-1,54

Com base nas estimativas da Tabela 15, observa-se que as estimativas dos parâmetros nos três conjuntos de dados (construção, validação e geral) são relativamente parecidas. Diante disso, é possível afirmar que o modelo proposto é válido. O modelo final para a variável IRA está apresentado na Tabela 16, com um AIC de 676,52.

Tabela 16: Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com IRA. Bacharelado em Ciência da Computação-UnB, 2012-2019

Variável	Estimativa	Desvio Padrão	Estatística	P-valor
Intercepto	4,1014	0,3340	12,2810	$< 0{,}0001$
IRA	-1,3822	0,1064	-12,9940	$< 0{,}0001$
Cursou verão				
Sim	-1,3256	0,2310	-5,7373	$< 0{,}0001$
Não				
Currículo				
Novo	-1,5362	0,2374	-6,4699	$< 0{,}0001$
Antigo				

- **Modelo com a variável Taxa de Reprovação MAT**

No modelo relacionado à Taxa de Reprovação MAT, foram aplicados os mesmos procedimentos do empregados no modelo anterior. A variável "Quantidade de SR" foi excluída devido à correlação identificada com a Taxa de Reprovação MAT. Além disso, tentativas de incluir interações entre a Taxa de Reprovação MAT e a quantidade de semestres cursados, bem como entre a Taxa de Reprovação MAT e o sistema de cotas, foram descartadas devido ao comprometimento da estabilidade do modelo.

A Tabela 17 apresenta as estimativas dos parâmetros para a amostra de construção, validação e o conjunto de dados completo.

Tabela 17: Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019

Parâmetro	Estimativa-Construção	Estimativa-Validação	Estimativa-Geral
Intercepto	0,89	-0,19	0,47
Taxa de Reprovação MAT	0,04	0,03	0,03
Forma de Ingresso - Vestibular	-0,23	0,79	0,17
Forma de Ingresso - PAS	-1,58	-0,06	-0,96
Forma de Ingresso - ENEM	-1,46	-0,82	-1,16
Cursou Verão - Sim	-1,29	-1,12	-1,24
Currículo - Novo	-2,35	-1,55	-1,98
Semestres cursados	-0,18	-0,11	-0,15

Com base nas estimativas da Tabela 17, observa-se que as estimativas dos parâmetros nos três conjuntos de dados (construção, validação e geral) são relativamente parecidas. Diante disso, é possível afirmar que o modelo proposto é relativamente estável. O modelo final para a variável Taxa de Reprovação está apresentado na Tabela 18, com um AIC de 653,98.

Tabela 18: Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Variável	Estimativa	Desvio Padrão	Estatística	P-valor
Intercepto	0,4728	0,3539	1,3358	0,1816
Taxa de Reprovação MAT	0,0324	0,0030	10,8161	$< 0,0001$
Forma de Ingresso				
Vestibular	0,1673	0,3447	0,4855	0,6273
PAS	-0,9578	0,3424	-2,7970	0,0052
ENEM	-1,1558	0,3889	-2,9718	0,0030
Outros				
Cursou verão				
Sim	-1,2379	0,2348	-5,2715	$< 0,0001$
Não				
Currículo				
Novo	-1,9841	0,2667	-7,4385	$< 0,0001$
Antigo				
Semestres cursados	-0,1548	0,0316	-4,8990	$< 0,0001$

4.4 Interpretação dos Parâmetros

A interpretação dos parâmetros estimados do modelo de regressão logística pode ser conduzida por meio da Razão de Chances (*Odds Ratio*). Esse método possibilita uma compreensão mais aprofundada das relações entre as variáveis explicativas e a variável de resposta.

Os valores estimados da Razão de Chances para cada parâmetro e seus intervalos de confiança correspondentes nos modelos IRA e Taxa de Reprovação MAT foram apresentados nas Tabelas 19 e 20, respectivamente.

4.4.1 Modelo com IRA

Tabela 19: Razão de chance e IC de 95% para o modelo IRA.
Bacharelado em Ciência da Computação-UnB, 2012-2019.

Variável Explicativa	Razão de Chances	IC 95%	
		LI	LS
IRA	0,251	0,204	0,309
Verão - Sim	0,266	0,169	0,418
Currículo - Novo	0,215	0,135	0,418

Dentre os resultados obtidos, destaca-se que a variável IRA indica que a chance de evasão reduz 74,9% para cada aumento unitário no IRA. Em outras palavras, à medida que o IRA do aluno aumenta, a chance de evasão diminui. Além disso, a realização de disciplinas durante o verão está associada a uma redução de 73,4% na chance de evasão, enquanto a reestruturação do currículo sugere um impacto positivo, uma vez que houve uma redução de 78,5% na chance de evasão em comparação com o currículo antigo. Importante ressaltar que essas associações se mantêm válidas quando as demais variáveis são mantidas constantes.

4.4.2 Modelo com Taxa de reprovação MAT

Tabela 20: Razão de chance e IC de 95% para o modelo Taxa de Reprovação. Bacharelado em Ciência da Computação-UnB, 2012-2019.

Variável Explicativa	Razão de Chances	IC 95%	
		LI	LS
Taxa de Reprovação MAT	1,033	1,027	1,039
Forma de Ingresso - Vestibular	1,182	0,602	2,323
Forma de Ingresso - PAS	0,384	0,196	0,751
Forma de Ingresso - ENEM	0,315	0,147	0,675
Cursou Verão - Sim	0,290	0,183	0,459
Currículo - Novo	0,138	0,082	0,232
Semestres cursados	0,857	0,805	0,911

Dentre os resultados obtidos, é importante destacar que todas as relações apresentadas se mantêm válidas quando as demais variáveis são mantidas constantes. Dessa forma, observa-se que alunos admitidos pelo Vestibular apresentaram uma chance 18,2% maior de evasão em relação à categoria "Outros". Em contrapartida, aqueles provenientes do PAS e ENEM mostraram uma redução de na chance de evasão, tendo, respectivamente, 61,6% e 68,5% menos chance de evasão em comparação com a mesma categoria.

A realização de disciplinas no verão demonstrou um forte impacto, com uma chance 71% menor de evasão para alunos que cursaram disciplinas nos semestres de verão. Por outro lado, a reestruturação do currículo também impactou positivamente, uma vez que o aluno estar cursando o currículo novo está associada a uma chance 86,2% menor de evasão em comparação com o currículo antigo.

Adicionalmente, observa-se que o número de semestres cursados está associado a uma redução na chance de evasão de 14,3% a cada semestre a mais cursado. Essa constatação sugere que, para cada aumento unitário na quantidade de semestres cursados, a chance de evasão diminui. Em outras palavras, à medida que um aluno cursa mais semestres, a chance de evasão diminui.

4.5 Teste de ajuste e diagnóstico dos modelos

Após a elaboração e ajuste de um modelo estatístico é fundamental avaliar a adequação desse ajuste para analisar quão bem ele se adapta ao conjunto de dados em questão. Para isso, utilizou-se os testes de Hosmer-Lemeshow e os resíduos Pearson e Deviance, conforme abordados na Seção 2.5.

4.5.1 Modelo com IRA

- **Teste de adequabilidade**

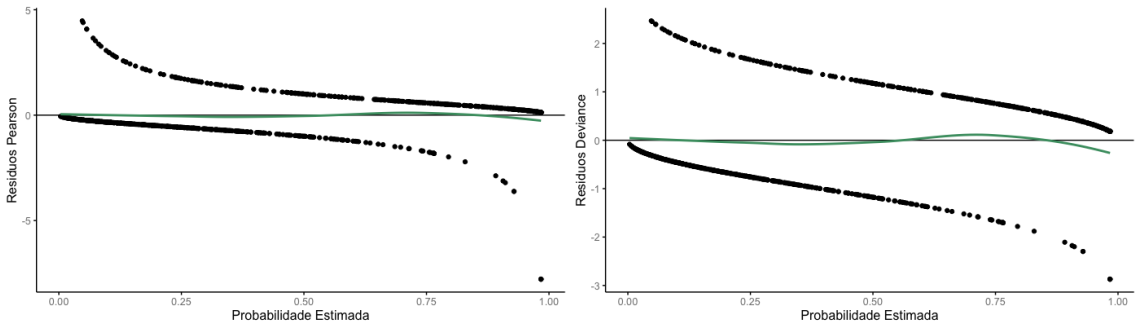
  Ao analisar o resultado do teste de Hosmer-Lemeshow, observou-se que, a um nível de significância de 5%, não há evidências suficientes para rejeitar a hipótese de que o modelo está adequadamente ajustado aos dados.

Tabela 21: Teste de adequabilidade para o modelo IRA.

Estatística do teste	Graus de Liberdade	P-valor
6,5394	8	0,587

- **Resíduos**

  As Figuras abaixo ilustram os resíduos de Pearson e *Deviance* em relação às probabilidades estimadas pelo modelo de regressão logística. Pode-se observar que ambos os gráficos possuem um comportamento semelhante. A premissa central é que, se o modelo proposto estiver correto, a suavização de *Lowess* deverá resultar em uma linha praticamente horizontal, com um intercepto próximo de zero. Portanto, ao analisar ambas as suavizações, que exibem uma inclinação próxima de zero, não há evidências que sustentem a conclusão de que o modelo seja inadequado.



(a) Resíduos de Pearson (b) Resíduos *Deviance*

Figura 21: Resíduos para o modelo IRA.
Bacharelado em Ciência da Computação-UnB, 2012-2019.

- **Distância de Cook**

  A distância de Cook é empregada para identificar possíveis observações influentes e avaliar o impacto delas nas estimativas dos parâmetros. A Figura abaixo exibe as distâncias de Cook para cada observação no modelo geral.

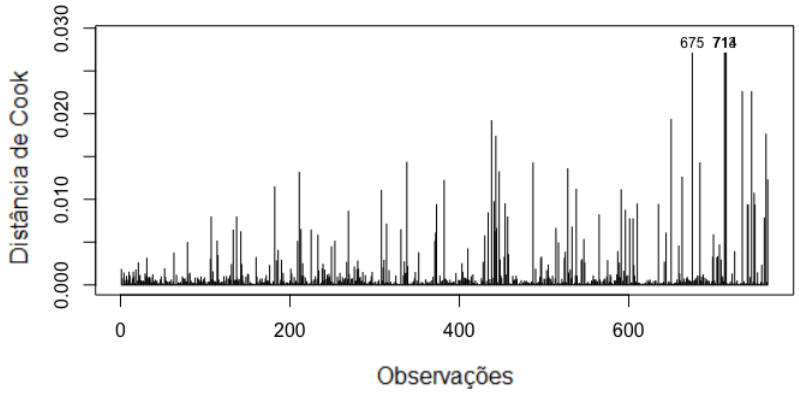


Figura 22: Distância de Cook para o modelo IRA.
Bacharelado em Ciência da Computação-UnB, 2012-2019

A Figura 22 revela a presença de algumas observações que são consideradas influentes, o que sugere a necessidade de uma investigação mais aprofundada. É importante destacar que como as observações 713 e 714 ficam muito próximas, o gráfico sobrepõe o rótulo dos valores.

As observações 675 e 714 correspondem a alunos que ainda estão ativos na Universidade de Brasília (UnB), cursaram apenas dois semestres e receberam menções SR em todas as disciplinas que cursaram. Quanto à observação 713, ela indica um aluno ativo que cursou dois semestres, no entanto, os únicos créditos registrados são de créditos concedidos, o que significa que o aluno ainda não cursou nenhuma disciplina na UnB.

- **Qualidade do ajuste - Curva ROC**

  A fim de avaliar a qualidade do ajuste, a curva ROC (Receiver Operating Characteristic) se destaca como uma ferramenta para fornecer uma visão clara da capacidade do modelo em distinguir duas categorias. Neste caso específico, a curva ROC é aplicada para analisar a evasão e não evasão dos alunos no curso de Bacharelado em Ciência da Computação.

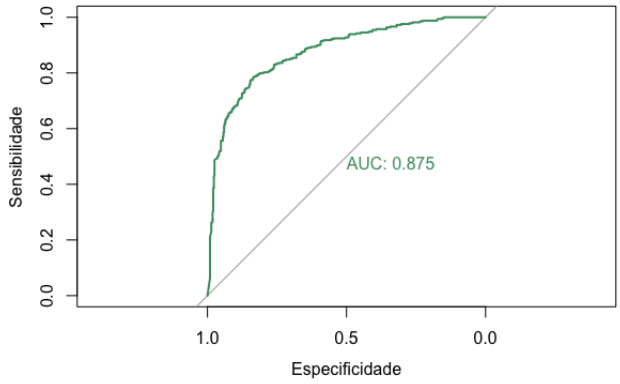


Figura 23: Curva ROC para o modelo IRA.
Bacharelado em Ciência da Computação-UnB, 2012-2019

Conforme destacado anteriormente na Seção 2.8, a área sob a curva ROC (AUC) é um indicador da capacidade do modelo em realizar discriminação, sendo que valores mais altos sugerem uma capacidade de classificação mais robusta. No caso do modelo com o Índice de Rendimento Acadêmico (IRA), a AUC atingiu 0,875, uma pontuação considerada excelente.

4.5.2 Modelo com Taxa de Reprovação MAT

- **Testes de adequabilidade e resíduos**

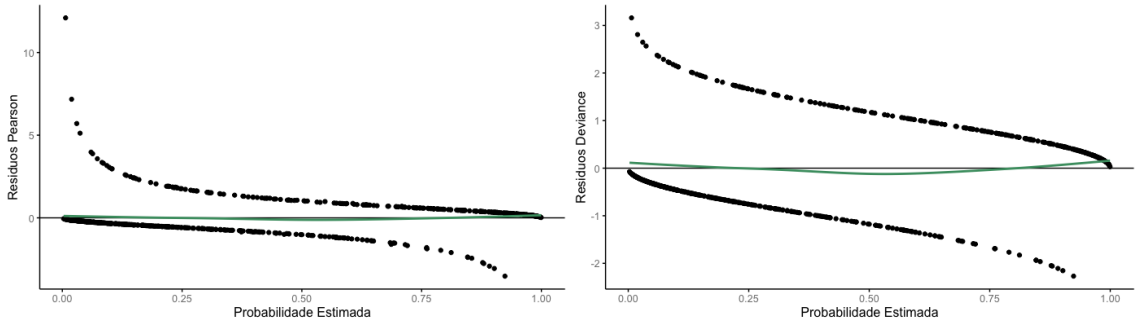
  Ao analisar o resultado do teste de Hosmer-Lemeshow, observou-se que, a um nível de significância de 5%, não há evidências suficientes para rejeitar a hipótese de que o modelo está adequadamente ajustado aos dados.

Tabela 22: Teste de adequabilidade para o modelo IRA.

Estatística do teste	Graus de Liberdade	P-valor
7,0845	8	0,5275

- **Resíduos**

  As Figuras abaixo ilustram os resíduos de Pearson e *Deviance* em relação às probabilidades estimadas pelo modelo de regressão logística. É evidente que, mesmo apresentando resultados diferentes no teste de adequação, ambos os gráficos exibem padrões semelhantes. Além disso, nota-se que a suavização de *Lowess* para ambos os gráficos aproxima-se de zero, indicando que o modelo está bem ajustado.



(a) Resíduos de Pearson (b) Resíduos *Deviance*

Figura 24: Resíduos modelo Taxa de Reprovação.
Bacharelado em Ciência da Computação-UnB, 2012-2019.

- **Distância de Cook**

  Da mesma forma realizada para o modelo anterior, a distância de Cook foi utilizada para identificar potenciais pontos influentes no modelo de taxa de reprovação. A Figura abaixo ilustra as distâncias de Cook para cada observação neste modelo específico.

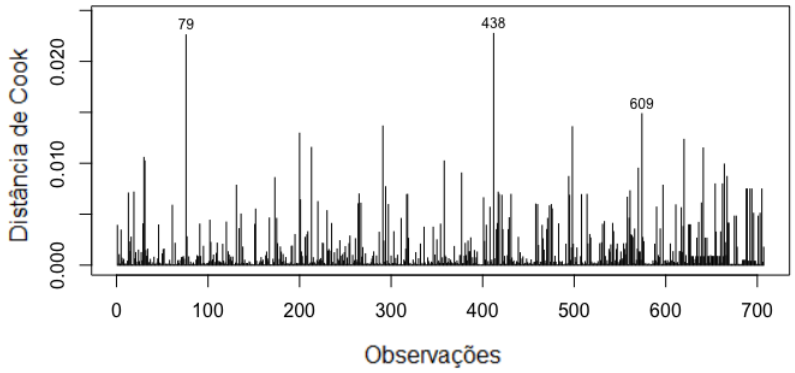


Figura 25: Distância de Cook para o modelo Taxa de Reprovação.
Bacharelado em Ciência da Computação-UnB, 2012-2019

Para o modelo da taxa de reprovação, também é necessário investigar alguns pontos específicos. A observação 438 refere-se a um aluno que evadiu durante um semestre de verão após cursar dois semestres. Este aluno obteve duas menções SR e apresentou uma taxa de reprovação em disciplinas da Matemática igual a 16,67%. Por outro lado, a observação 609 pertence a um aluno ativo no curso, que cursou dois semestres, mas não reprovou em nenhuma disciplina da Matemática.

No caso da observação 79, destaca-se os dados de um aluno que evadiu após ter cursado 15 semestres, com 7 menções SR e cursou algum de verão.

- **Qualidade do ajuste - Curva ROC**

  Assim como para o modelo anterior, utilizou-se a curva ROC para avaliar a qualidade do ajuste para o modelo.

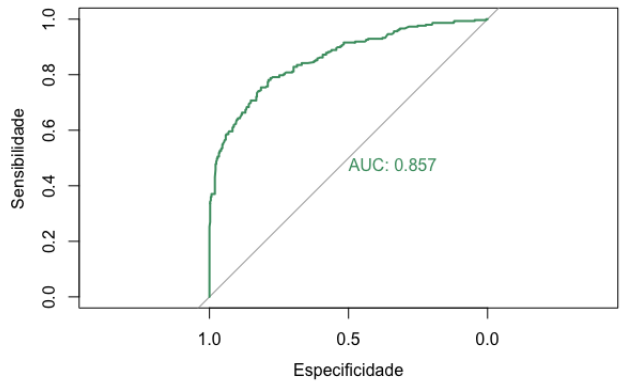


Figura 26: Curva ROC para o modelo Taxa de Reprovação.
Bacharelado em Ciência da Computação-UnB, 2012-2019

Diante a curva ROC acima foi obtido um AUC (área abaixo da curva ROC) de 0.857, ou seja, o modelo também fornece uma habilidade de classificação excelente.

5 Conclusão

Considerando que o objetivo principal do estudo é identificar as características dos estudantes associadas à evasão acadêmica e traçar o perfil dos alunos que abandonaram o curso, pôde-se observar que há evidências interessantes na presente análise.

Ao analisar o perfil dos alunos, observou-se que o curso de Bacharelado em Ciência da Computação na UnB é composto majoritariamente por alunos do gênero masculino (90%). A mediana da idade dos alunos ao ingressarem no curso é de 18 anos, sendo que 75% deles têm até 20 anos. Quanto à renda familiar, nota-se que os alunos em sua maioria residem em áreas de alta renda no Distrito Federal (81%).

Para fazer uma melhor análise da evasão no curso de Ciência da Computação da UnB, buscou-se entender a vida acadêmica desse aluno, desde o momento em que ele ingressou na Universidade, até sua trajetória de sucesso ou insucesso no curso, passando pelo seu rendimento acadêmico, desistências e outros fatores.

Foram identificados desafios enfrentados nos primeiros três semestres, especialmente em disciplinas de cálculo e computação básica. Ainda sobre as disciplinas em que alunos apresentam dificuldades, com a implementação do currículo novo, a não obrigatoriedade das disciplinas de física pode ter influenciado positivamente na redução da evasão de alunos uma vez que o índice de reprovação dessa disciplina é tradicionalmente muito alto.

Com base na saída de alunos do curso, constatou-se que, da totalidade da amostra, 57% evadiram, enquanto 43% concluíram ou permaneceram ativos até o segundo semestre de 2019. Verificou-se também que os alunos que evadiram não ultrapassaram 06 semestres cursados, enquanto a maioria dos que não evadiram estudou em média 09 semestres, que é o tempo padrão para a conclusão do curso.

Na elaboração do modelo de regressão logística, foi decidido criar dois modelos distintos. Um desses modelos inclui a variável IRA (Índice de Rendimento Acadêmico) entre as variáveis explicativas, enquanto o outro leva em consideração a variável Taxa de Reprovação MAT. Essa abordagem foi adotada devido à elevada correlação entre essas variáveis explicativas. É importante destacar que o modelo de Taxa de Reprovação MAT inclui dois parâmetros adicionais além dos parâmetros do modelo IRA (Forma de Ingresso e Semestres cursados), pois apenas nesse contexto essas variáveis demonstraram significância estatística.

Em ambos os modelos, dois fatores se destacaram: alunos que cursaram disciplinas no verão e aqueles que seguem o currículo novo têm chances significativamente menores de evasão em comparação com aqueles que não cursaram no verão e seguem o

currículo antigo. Na análise do modelo que considera a variável IRA, a chance de evasão é 73,4% menor para alunos que optaram pelo verão e 78,5% menor para os matriculados no currículo novo. No caso do modelo de Taxa de Reprovação MAT, essas reduções foram de 71% e 86,2%, respectivamente. Pelos resultados verificados com a variável dos cursos de verão, pode-se inferir que a estratégia da oferta dessas disciplina ajuda sobremaneira os alunos a resgatarem seu rendimento e trajetória acadêmica e/ou adiantar sua formação.

Além disso, ao considerar os parâmetros adicionais do modelo de Taxa de Reprovação MAT, observa-se que alunos admitidos pelo Vestibular apresentaram uma chance 18,2% maior de evasão em comparação com a categoria "Outros". Por outro lado, aqueles do PAS e ENEM mostraram notáveis reduções na chance de evasão, com diminuições de 61,6% e 68,5%, respectivamente.

Ambos os modelos demonstraram um ajuste satisfatório, com uma especificidade superior a 80%. Diante da similaridade nos resultados entre os modelos, recomenda-se a utilização do modelo IRA, dada sua simplicidade e a praticidade na obtenção das informações associadas ao Índice de Rendimento Acadêmico, as quais são prontamente calculadas pela Universidade de Brasília.

Sugere-se para futuros estudos:

- Aprofundar análises com o foco na diferença da evasão com relação ao gênero, a fim de avaliar a possível existência de desafios ou obstáculos na entrada de mulheres no curso de Ciência da Computação. Tal investigação se torna diante da notável diferença no número de estudantes ingressantes entre os gêneros;
- Analisar de forma comparativa os resultados obtidos no curso de Ciência da Computação da UnB e outros cursos de Ciência da Computação de outros estados da federação de forma a buscar mais informações sobre amostras de alunos de baixa renda (uma vez que a maioria da população desta pesquisa era de alta renda);
- Analisar mais profundamente a composição curricular das matérias iniciais do Bacharelado da Ciência da Computação, de forma a compreender as demandas iniciais e equilibrar o grau de dificuldade das disciplinas vis-a-vis o preparo dos alunos ingressantes.

Os resultados desta pesquisa são apenas uma contribuição aos estudos sobre a formação de alunos no ensino superior das ciências exatas. O aprofundamento nessa área é essencial para aprimorar os resultados da graduação e reduzir fatores que levam à evasão. Ainda, a UnB pode utilizar essas reflexões para desenvolver estratégias de redução da evasão e melhoria do desempenho dos alunos, resultando, a longo prazo, em um aumento de profissionais qualificados para áreas cruciais para o crescimento do país.

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis, 2nd edition*. [S.l.]: John Wiley and Sons Inc, New York, 2019.

ANDIFES; ABRUEM; SESU/MEC. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. comissão especial de estudo sobre a evasão nas universidades públicas brasileiras. 1996.

CNI. *Para ser mais competitivo, Brasil precisa investir e capacitar para a indústria 4.0*. 2021. ⟨https://noticias.portaldaindustria.com.br/noticias/educacao/para-ser-mais-competitivo-brasil-precisa-investir-e-capacitar-para-a-industria-40/⟩. Acesso em: 29/04/2023.

DIEESE. *Pesquisa de Emprego e Desemprego*. 2018. ⟨https://www.dieese.org.br/analiseped/2018/201804pedbsb.html⟩. Acesso em: 02/09/2023.

EDUCAÇÃO, D. da. *Evasão bate recordes no ensino superior*. 2022. ⟨https://desafiosdaeducacao.com.br/evasao-bate-recordes-no-ensino-superior/⟩. Acesso em: 28/04/2023.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. Cad. Pesqui, v. 37, p. 641–659, 2007.

GARCIA, L. M. L. d. S.; GOMES, R. S. Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica. Revista Educar Mais, v. 6, p. 940–941, 2022.

HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. John Wiley and Sons Inc, New York, 2019.

NETER, J. et al. *Applied Linear Statistical Models, 5th edition*. [S.l.]: McGraw-Hill/Irwin, 2004.

R. *The R Project for Statistical Computing*. ⟨https://www.r-project.org/⟩.

UNB. *Parque Científico e Tecnológico da UnB (PCTec/UnB)*. 2023. ⟨https://www.pctec.unb.br/sobre⟩. Acesso em: 28/04/2023.