

Universidade de Brasília
Departamento de Estatística

Estudo da Evasão nos Cursos de Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília

Vítor Hott Amorim

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

Brasília
2025

Vítor Hott Amorim

Estudo da Evasão nos Cursos de Engenharia da Computação e Engenharia Mecatrônica da Universidade de Brasília

Orientador(a): Prof(a). Leandro Tavares Correia

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

Brasília
2025

Agradecimentos

Primeiramente, gostaria de agradecer aos meus pais e à minha irmã, que sempre estiveram ao meu lado, apoiando e dando incentivo não apenas durante a graduação, mas em todas as etapas da minha vida.

Também gostaria de agradeçer ao meu orientador, o Professor Leandro Tavares. Sua orientação, paciência e apoio foram essenciais durante todo o processo e fundamentais para a realização deste Trabalho de Conclusão de Curso.

Por fim, expresso meus agradecimentos ao Departamento de Estatística, a todos os professores e à Universidade de Brasília. Esses anos de graduação foram muito importantes para o meu desenvolvimento acadêmico e também para o meu crescimento pessoal.

Resumo

Este estudo tem como objetivo verificar os fatores que levam os estudantes de Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília a evadirem de seus cursos. Foi elaborado um modelo de regressão logística para cada curso, além de um modelo multinível para avaliar os cursos de forma conjunta. Foram considerados somente estudantes ingressantes entre o primeiro semestre de 2015 e o primeiro semestre de 2023.

A Engenharia Mecatrônica, que apresentou maior evasão, teve como fatores significativos o Índice de Rendimento Acadêmico (IRA), números trancamentos e ensino remoto. Para Engenharia de Computação, além dos fatores citados, foram relevantes a distância de deslocamento, raça/cor, forma de ingresso e tipo de escola. Avaliando conjuntamente os dois cursos, também foi significativa a variável carga horária obrigatória de aula do curso. De modo geral, o desempenho acadêmico mais alto reduz a chance de evasão, assim como o fato de o estudante ter cursado disciplinas durante o ensino remoto, devido às flexibilizações em relação à reprovação e critérios de desligamento.

Palavras-chaves: Evasão; Regressão Logística; Modelo Multinível; Engenharia de Computação; Engenharia Mecatrônica.

Lista de Tabelas

1 Formas de saída do os cursos de Engenharia de Computação e Engenharia Mecatrônica - UnB; 2015-2023. 25

2 Distribuição dos estudantes segundo gênero e raça/cor. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 29

3 Distribuição dos estudantes segundo informações de entrada na universidade. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 31

4 Distribuição dos estudantes segundo tipo de cota declarada. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 32

5 Distribuição dos estudantes segundo características acadêmicas. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 32

6 Tabela de medidas resumo da distribuição do número de menções SR. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 . . . 33

7 Tabela de medidas resumo da distribuição do número de trancamentos. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 . 33

8 Distribuição dos estudantes segundo cocorrência de evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 34

9 Percentual de evasão segundo gênero e raça/cor. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 35

10 Percentual de evasão segundo informações de entrada na universidade e curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 36

11 Percentual de evasão segundo características acadêmicas. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 37

12 Tabela de medidas resumo da distribuição do número de menções SR com relação à evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 38

13 Tabela de medidas resumo da distribuição do número de trancamentos com relação à evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 38

14 Associação das variáveis com Evasão 39

15 Modelo com IRA do *Stepwise* para o curso de Engenharia de Computação com os dados de treino 41

16 Modelo com Taxa de Reprovação do *Stepwise* para o curso de Engenharia de Computação com os dados de treino 42

17 Modelo final para o curso de Engenharia de Computação 43

18 Razão de chances e IC 95% - Modelo final Engenharia de Computação . . 43

19 Teste de adequação de ajuste - Modelo Engenharia de Computação 44

20 Matriz de Confusão - Modelo Engenharia de Computação 46

21 Modelo com IRA do *Stepwise* para o curso de Engenharia Mecatrônica com os dados de treino 47

22 Modelo sem IRA do *Stepwise* para o curso de Engenharia Mecatrônica com os dados de treino 47

23 Modelo final para o curso de Engenharia Mecatrônica 48

24 Razão de chances e IC 95% - Modelo final Engenharia Mecatrônica 48

25 Teste de adequação de ajuste - Modelo Engenharia Mecatrônica 48

26 Matriz de Confusão - Modelo Engenharia Mecatrônica 50

27 Modelo sem variáveis explicativas - Modelo Nulo 51

28 Modelo com variáveis explicativas do 1º nível 52

29 Modelo com variáveis explicativas do 1º nível e do 2º nível 53

30 Razão de chances e IC 95% - Modelo Multinível 54

31 Matriz de Confusão - Modelo Multinível 56

Lista de Figuras

1 Distribuição das variáveis Idade ao ingressar e Distância de Deslocamento (km), por curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 30

2 Distribuição das variáveis IRA e Taxa de Reprovação, por curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 33

3 Distribuição das variáveis Idade ao ingresssar e Distância de Deslocamento (km), por curso e evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 36

4 Distribuição das variáveis IRA e Taxa de Reprovação, por curso e evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023 . 38

5 Gráficos de análise para os resíduos do Modelo de Engenharia de Computação 45

6 Curva ROC - Modelo de Engenharia de Computação 46

7 Gráficos de análise para os resíduos do Modelo de Engenharia Mecatrônica 49

8 Curva ROC - Modelo de Engenharia Mecatrônica 50

9 Gráficos de análise para os resíduos do Modelo Multinível 55

10 Curva ROC - Modelo Multinível 55

11 Gráficos de análise para os resíduos do Modelo de Engenharia de Computação com a Taxa de Reprovação 60

12 Gráficos de análise para os resíduos do Modelo de Engenharia de Mecatrônica sem IRA 61

Sumário

1 Introdução 8

2 Objetivos 10

2.1 Objetivo Geral 10

2.2 Objetivos Específicos 10

3 Referencial Teórico 11

3.1 Regressão Logística 11

3.1.1 Estimação dos Parâmetros do Modelo 11

3.1.2 Interpretação dos Parâmetros 12

3.1.3 Intervalo de Confiança para os Parâmetros do Modelo 13

3.1.4 Testes de Significância do Modelo 14

3.1.5 Seleção do Modelo 15

3.1.6 Diagnóstico do Modelo 16

3.1.7 Qualidade do Ajuste 17

3.2 Regressão Linear Multinível 18

3.2.1 Modelo Multinível 18

3.2.2 Modelo de Regressão Básico de Dois Níveis 18

3.2.3 Estimação dos Parâmetros 19

3.2.4 Estratégia de Análise 20

3.2.5 Regressão Logística Multinível 22

4 Metodologia 24

4.1 Banco de dados 24

4.1.1 Variáveis 24

4.1.2 Criação de variáveis 25

4.1.3 Variável dos cursos de graduação 27

4.2 Procedimento para Modelagem 27

5 Resultados 29

5.1 Análise Descritiva 29

5.1.1 Análise Conjunta 34

5.1.2 Análise de associação entre as variáveis explicativas e a evasão . . . 39

5.2 Associação entre as variáveis explicativas 40

5.3 Modelagem Estatística 41

5.3.1 Modelo do curso de Engenharia de Computação 41

5.3.2 Modelo logístico do curso de Engenharia Mecatrônica 46

5.3.3 Modelo logístico multinível 50

6 Conclusão 57

Referências 59

Apêndice 60

A Diagnóstico do Modelo com Taxa de Reprovação para o curso de Engenharia de Computação 60

B Diagnóstico do Modelo sem IRA para o curso de Engenharia Mecatrônica 61

1 Introdução

A evasão nos cursos de graduação é um problema importante que atinge as instituições de ensino superior no Brasil. Primeiramente, é necessário compreender o conceito de evasão, que pode ser entendido como a saída definitiva do curso de origem sem conclusão, ou a diferença entre ingressantes e concluintes, após uma geração completa (ANDIFES; ABRUEM et al., 1996).

De acordo com Filho et al. (2007), a evasão de estudantes dos seus cursos são desperdícios sociais, acadêmicos e econômicos. No setor público, os recursos investidos em professores, funcionários, equipamentos e espaço público não geram o devido retorno quando a taxa de evasão é elevada. Portanto, além de gerar consequências negativas aos estudantes, a evasão também gera prejuízos às instituições de ensino.

Segundo o Mapa do Ensino Superior do Brasil do Instituto Semesp (2025), a taxa de evasão no ensino superior em 2023 foi de 26,4% para os cursos presenciais. Nas instituições de ensino públicas, a taxa foi de 23,1%, o que representa uma diminuição em relação ao ano de 2022, em que a taxa era 23,4%, mas representa um aumento em comparação aos anos anteriores (2019 a 2021).

Diversos fatores podem ser a causa da evasão no ensino superior. Segundo Silva, Cabral e Pacheco (2020), os principais motivos que levam um estaudante a evadir são: falta de clareza pela escolha do curso; pouco envolvimento com o curso; baixo desempenho acadêmico; imaturidade do estudante; falta de assessoramento aos alunos calouros; mudança de cidade; entre outros fatores.

Diante desse cenário, é necessário que as instituições de ensino público como a Universidade de Brasília (UnB) compreendam o fenômeno da evasão e as suas causas, a fim de se obter ferramentas para amenizar este problema. Ao compreender quais fatores são responsáveis pelo abandono da graduação, é possível elaborar estratégias e ações eficazes para a diminuição deste fenômeno e, assim, reduzir o gasto de recursos públicos sem retorno e formar mais profissionais qualificados que possam contribuir para a sociedade.

No Brasil, percebe-se que há uma alta evasão nos cursos dos campos de ciências exatas, ciências naturais e engenharias, da acordo com Saccaro, França e Jacinto (2019). Nesse sentido, estudos sobre esse assunto se mostram bastante importantes para ajudar a elaborar políticas públicas que auxiliem na permanência dos estudantes em seus cursos de graduação.

Este trabalho tem o objetivo de estudar a evasão de estudantes dos cursos de

Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília (UnB), entre o primeiro semestre de 2018 e o primeiro semestre de 2023. Neste estudo, será realizada a modelagem via regressão logística para cada curso, devido ao fato da variável reposta (evasão) ser qualitativa binária. Além disso, será utilizado também um modelo multinível de dois níveis, a fim de verificar qual a melhor abordagem. A modelagem permitirá identificar quais fatores interferem na evasão e verificar o perfil dos estudantes evadidos nos cursos analisados.

Na Seção 2, são apresentados o Objetivo Geral do trabalho e os Objetivos Específicos. Em seguida, na Seção 3, há o Referencial Teórico contendo informações acerca da regressão logística, dos modelos multinível, das formas de avaliar a qualidade do ajuste e dos métodos de seleção de variáveis. A Seção 4 contém as informações sobre as variáveis do banco de dados e o procedimento utilizado para a modelagem. Na Seção 5, são apresentados os resultados das análises descritivas, dos modelos de regressão logística para cada curso e do modelo multinível. Por fim, há a conclusão do trabalho com uma síntese dos principais resultados e pontos discutidos ao longo do estudo.

2 Objetivos

2.1 Objetivo Geral

O objetivo deste projeto é estudar a evasão universitária no período do primeiro semestre de 2015 ao primeiro semestre de 2023 entre os estudantes do Bacharelado em Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília, com o intuito de verificar quais variáveis contribuem para a desistência dos cursos de graduação.

2.2 Objetivos Específicos

- Avaliar as características dos alunos dos cursos de Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília;
- Identificar e mensurar os fatores associados à evasão por meio da técnica de regressão logística e regressão logística multinível;
- Verificar se o ensino remoto durante a pandemia teve impacto na evasão;
- Comparar a evasão entre ambos os cursos de graduação.

3 Referencial Teórico

Para mais detalhes, as informações desta seção podem ser obtidas em Agresti (2007), Hosmer e Lemeshow (2000) e Hox, Moerbeek e Schoot (2017).

3.1 Regressão Logística

O objetivo dos modelos de regressão é analisar a relação entre uma variável de interesse, chamada de variável resposta ou dependente, e um conjunto de variáveis explicativas ou independentes. Utiliza-se o modelo de regressão logística quando a variável resposta é binária. A relação entre a variável resposta e as variáveis explicativas é modelada através de uma função de ligação e, nesse caso, utiliza-se a função logito.

Na regressão logística, a variável resposta Y tem distribuição Bernoulli, em que $P(Y = 1) = \pi$ é a probabilidade de sucesso e $P(Y = 0) = 1 - \pi$ é a probabilidade de fracasso. A estrutura do modelo logístico é expressa da seguinte forma:

$$P(Y = 1|X) = \pi(x) = \frac{exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}, \tag{3.1.1}$$

em que $\pi(x)$ é a probabilidade condicional de sucesso da variável resposta dadas as variáveis explicativas, p é o número de variáveis explicativas (X_i) incluídas no modelo e os β_i's são os parâmetros do modelo.

Ao aplicar a transformação logito na equação 3.1.1, observa-se que:

$$\text{logito}(\pi(x)) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

3.1.1 Estimação dos Parâmetros do Modelo

O método de estimação dos parâmetros do modelo de regressão logística é o de máxima verossimilhança. De forma geral, esse método obtém valores para os parâmetros desconhecidos que maximizam a probabilidade de se ter o conjunto de dados observados. A função de máxima verossimilhança é a seguinte:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1 - \pi_i)^{1-Y_i}.$$

Aplicando o logaritmo nessa expressão, tem-se o log da verossimilhança:

$$\ln L(\beta) = \ln \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} = \sum_{i=1}^{n} \left[Y_i \ln \left(\frac{\pi_i}{1-\pi_i} \right) \right] + \sum_{i=1}^{n} \ln(1-\pi_i),$$

Como $1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_i)]^{-1}$, a função de log-verossimilhança fica da seguinte forma:

$$l(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i) - \sum_{i=1}^{n} \ln\left[1 + \exp(\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i)\right].$$

Para encontrar as estimativas de máxima verossimilhança, deriva-se a função de log-verossimilhança em relação aos parâmetros $\beta_0, \beta_1, \ldots, \beta_p$. Como não há fórmula fechada para encontra essas estimativas, utiliza-se métodos numéricos, como o Newton-Raphson.

3.1.2 Interpretação dos Parâmetros

A interpretação das estimativas dos parâmetros no modelo de regressão logística é feita em termos de *odds* (chances). Por meio da função $logito(\pi(x))$ tem-se que

$$\text{logito}(\pi(x)) = \frac{\pi(x)}{1-\pi(x)} = \text{odds} = e^{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p}. \tag{3.1.2}$$

Ao tomar dois valores distintos de uma variável explicativa x_j e $x_j + 1$ mantendo as demais variáveis constantes, e considerando $odds_1 = \pi_1/(1-\pi_1)$ e $odds_2 = \pi_2/(1-\pi_2)$, nota-se que

$$\ln(odds_2) - \ln(odds_1) = \ln\left(\frac{odds_2}{odds_1}\right) = \beta_1. \tag{3.1.3}$$

Ao aplicar o exponencial de ambos os lados da equação 3.1.3, obtém-se a *odds ratio* (OR) ou razão de chances, que é denotada por

$$OR = \frac{odds_2}{odds_1} = e^{\beta_1}.$$

Para variáveis quantitativas, a chance estimada de sucesso (*odds*) é multiplicada por e^{β_i} para cada unidade a mais da variável explicativa X_i correspondente. Para variáveis qualitativas, a razão de chances é a razão das chances do sucesso ocorrer em um grupo quando comparado ao outro.

3.1.3 Intervalo de Confiança para os Parâmetros do Modelo

O Estimador de Máxima Verossimilhança dos parâmetros tem a propriedade de normalidade assintótica. Ou seja, para amostras grandes, tem-se que

$$\hat{\beta}_j \approx N(\beta_j, \widehat{\mathrm{Var}}(\hat{\beta}_j)),$$

em que $\widehat{\mathrm{Var}}(\hat{\beta}) = \widehat{SE}(\hat{\beta}_j)^2 = (\mathbf{X}'\mathbf{V}\mathbf{X})^{-1}$ e em que:

$$\mathbf{X} = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

e a matriz $\mathbf{V}$ é:

$$\mathbf{V} = \begin{bmatrix} \hat{\pi}(X_1)(1-\hat{\pi}(X_1)) & 0 & \cdots & 0 \\ 0 & \hat{\pi}(X_2)(1-\hat{\pi}(X_2)) & \cdots & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & \hat{\pi}(X_n)(1-\hat{\pi}(X_n)) \end{bmatrix}.$$

Portanto, o intervalo de confiança de 100(1-α)% para um coeficiente β_i é definido como

$$\hat{\beta}_i \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}(\hat{\beta}_i),$$

em que $z_{1-\alpha/2}$ é o quantil 100(1-α)% da distribuição Normal Padrão e $\widehat{\mathrm{SE}}(\hat{\beta}_i)$ a estimativa do erro padrao do parâmetro β_i.

O intervalo de confiança para a Razão de Chances é definido como

$$exp(\hat{\beta}_i \pm z_{1-\alpha/2}\, \widehat{\mathrm{SE}}(\hat{\beta}_i)).$$

3.1.4 Testes de Significância do Modelo

Depois de ajustar um modelo particular de regressão logística, inicia-se o processo de avaliar a significância das variáveis do modelo, ou seja, verificar se as variáveis explicativas estão relacionadas com a variável resposta. Para isso, utiliza-se principalmente os testes da Razão de Verossimilhança e o de Wald.

Para o teste da Razão de Verossimilhança, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_1 = \ldots = \beta_p = 0 \\ H_1 : \beta_i \neq 0 \text{ para algum i} = 1,..,\text{p}. \end{cases}$$

A estatística do teste denotada por G^2 é dada por

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

sendo,

- $L(R)$ a verossimilhança para o modelo reduzido, quando a hipótese nula é verdadeira;
- $L(F)$ a verossimilhança para o modelo completo;
- p o número de parâmetros do modelo.

Para n grande, a estatística G^2 segue uma distribuicao Qui-quadrado com $p - q$ graus de liberdade, sendo p o número de parâmetros do modelo completo e q o número de parâmetros do modelo reduzido.

Para o teste de Wald, tem-se as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_j = 0 \\ H_1 : \beta_i \neq 0. \end{cases}$$

Sendo

$$Z = \frac{\hat{\beta}_i}{\widehat{\text{SE}}(\hat{\beta}_i))} \sim N(0, 1),$$

a estatística do Teste de Wald é

$$Z^2 = \left(\frac{\hat{\beta}_i}{\widehat{\text{SE}}(\hat{\beta}_i))} \right)^2 \sim \chi^2 \text{ com 1 g.l.}$$

3.1.5 Seleção do Modelo

É necessário na regressão logística selecionar um modelo que seja de fácil interpretação e que seja complexo o suficiente para ajustar os dados. Portanto, são utilizados critérios como Critério de Informação Akaike (AIC) e o Critério de Informação Bayesiano (BIC) para selecionar o melhor modelo para o problema. Esses critérios fornecem o quão próximo os valores ajustados do modelo se aproximam dos verdadeiros valores esperados e aplicam penalidades para modelos que têm maior número de parâmetros. Os critérios são definidos por:

$$AIC = -2l(\beta) + 2p \text{ e}$$

$$BIC = -2l(\beta) + p\ln(n),$$

em que $l(\beta)$ é a log-verossimilhança e p é o número de parâmetros.

Também existem métodos de seleção automático e que buscam por um modelo parcimonioso. Esses métodos, como o *Forward*, o *Forward* e o *Stepwise*, serão descritos a seguir.

- *Forward*

  Começa com uma quantidade mínima de variáveis e adiciona variáveis sequencialmente, até que novas adições não melhorem significativamente o ajuste.

- *Backward*

  É o procedimento inverso ao do *Forward*. Começa com todas as variáveis explicativas e remove sequencialmente essas variáveis, até que o critério de seleção estabelecido seja atingido.

- *Stepwise*

  É uma combinação dos métodos *Forward* e *Backward*. As variáveis são adicionadas e removidas, para identificar quais devem permanecer ou não no modelo, até que o critério de seleção estabelecido seja atingido.

3.1.6 Diagnóstico do Modelo

Na regressão logística, é necessário realizar testes de adequabilidade do ajuste e análise de resíduos para verificar se o modelo é apropriado ou não para o conjunto de observações.

- **Teste Qui-Quadrado de Pearson**

  O teste de Qui-Quadrado de Pearson determina se o modelo de regressão logística se ajusta aos dados. As hipóteses do teste são:

  $$\begin{cases} H_0 : \text{O modelo de regressão logística ajusta-se aos dados} \\ H_1 : \text{O modelo de regressão logística não se ajusta aos dados.} \end{cases}$$

  A estatística do teste é definida por

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

  Sob a hipótese nula, χ^2 segue aproximadamente uma distribuição Qui-Quadrado com $c - p$ graus de liberdade, em que c é o número do conjunto de valores distintos das variáveis explicativas e p é o número de parâmetros do modelo.

- **Resíduos de Pearson**

  Os resíduos de Pearson são a diferença de um valor observado e o seu valor estimado, divido pelo erro padrao estimado de Y_i. A estatística Qui-Quadrado em função dos resíduos de Pearson é a seguinte:

  $$\chi^2 = \sum_{i=1}^{n} r(y_i, \hat{\pi}_i)^2,$$

  em que

  $$e_i = \frac{y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}, \quad i = 1, \ldots, n.$$

  Sob a hipótese nula, segue uma distribuição Qui-Quadrado com $n - (p + 1)$ graus de liberdade.

- **Resíduos *Deviance***

Esses resíduos permitem avaliar a adequação do modelo aos dados observados. A estatística baseada nos residuos *deviance* é definida pela equação:

$$Deviance = \sum_{j=1}^{n} d(y_j, \hat{\pi}_j)^2,$$

sendo,

$$d(y_i, \hat{\pi}_i) = \pm \left\{ 2 \left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i) \log\left(\frac{1 - y_i}{1 - \hat{\pi}_i}\right) \right] \right\}^{1/2}.$$

Sob a hipótese nula, a estatística *Deviance* possui $n - (p + 1)$ graus de liberdade.

3.1.7 Qualidade do Ajuste

- **Curva ROC**

  A curva ROC (Receiver Operating Characteristic) é um gráfico utilizado na análise do desempenho de um sistema binário. A curva mostra a probabilidade de detectar verdadeiros positivos (sensibilidade) e falsos positivos (1 - especificidade) para o conjunto de todos os pontos de corte possíveis. O eixo das ordenadas representa a sensibilidade e o eixo das abcissas representa 1 - especificidade.

  A área embaixo da curva (AUC) vai de 0 a 1 e fornece uma medida para a habilidade que o modelo tem em discriminar em dois conjuntos quais experimentaram o evento de interesse e quais não. Na medida em que o valor da AUC se aproxima de 1, indica melhora na capacidade de discriminação do modelo.

  Como regra geral:

  - Se UC = 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 AUC $\geq$ 0,9: a discriminação é excepcional.

- **Matriz de Confusão**

  A matriz de confusão é uma tabela que mostra as frequências para cada classe predita e observada do modelo. É composta por:

  - Verdadeiro positivo (VP): previsão classificada como positiva dado que ela realmente é positiva;

- Falso positivo (FP): previsão classificada como positiva dado que ela negativa;
- Verdadeiro neagtivo (VN): previsão classificada como negativa dado que ela realmente é negativa;
- Falso negativo (FN): previsão classificada como negativa dado que ela é positiva;

A sensibilidade, a especificidade e a acurácia são métricas usadas para avaliar a qualidade do modelo. São dadas pelas seguintes equações:

$$\text{Sensibilidade} = \frac{VP}{VP + FN},$$

$$\text{Especificidade} = \frac{VN}{VN + FP} \text{ e}$$

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}.$$

3.2 Regressão Linear Multinível

3.2.1 Modelo Multinível

Modelos multinível lidam com populações que possuem uma estrutura hierárquica. Uma amostra desses tipos de populações pode ser descrita como uma amostra extraída em mais de um estágio em que, primeiro, toma-se uma amostra do nível mais alto e, em seguida, toma-se amostras de sub-unidades das unidades avaliadas.

Neste trabalho, a regressão logística multinível permite analisar, simultaneamente, as influências das variáveis em diferentes níveis hierárquicos nos dados. Nesse caso, o primeiro nível representa os estudantes de graduação e o segundo nível representa os cursos de graduação (Engenharia da Computação e Engenharia Mecatrônica).

3.2.2 Modelo de Regressão Bãsico de Dois Níveis

O modelo de regressão básica de dois níveis, com duas variáveis explicativas, pode ser descrito como:

$$Y_{ij} = \beta_{0j} + \beta_{1j}X_{1ij} + \beta_{2j}X_{2ij} + \varepsilon_{ij},$$

em que,

- Y_{ij} é a variável resposta;
- β_{0j} é o intercepto;
- β_{1j} e β_{2j} são os coeficientes de regressão;
- ε_i é o erro aleatório da regressão.

O primeiro nível é indexado por i $(i = 1, \ldots, n_j)$ e o segundo nível é indexado por j $(j = 1, \ldots, J)$.

Ao introduzir variáveis explicativas do segundo nível (Z_j), tem-se que

$$\begin{aligned}\beta_{0j} &= \gamma_{00} + \gamma_{01} Z_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} Z_j + u_{1j} \\ \beta_{2j} &= \gamma_{20} + \gamma_{21} Z_j + u_{2j}\end{aligned}$$

sendo β_{0j} o intercepto e β_{1j} e β_{2j} os coeficientes de regressão. Os termos u_{0j}, u_{1j} e u_{2j} são termos de erro residual (aleatórios) no 2^{o} nível, tendo média zero e sendo independentes dos erros residuais ε_{ij} no primeiro nível.

Portanto, o modelo com duas variáveis explicativas pode ser escrito como uma única equação de regressão multinível completa:

$$Y_{ij} = \gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_j + \gamma_{11} X_{1ij} Z_j + \gamma_{21} X_{2ij} Z_j + u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij}.$$

O segmento $[\gamma_{00} + \gamma_{10} X_{1ij} + \gamma_{20} X_{2ij} + \gamma_{01} Z_j + \gamma_{11} X_{1ij} Z_j + \gamma_{21} X_{2ij} Z_j]$ contém os coeficientes fixos do modelo e o segmento $[u_{1j} X_{1ij} + u_{2j} X_{2ij} + u_{0j} + e_{ij}]$ contém os termos randômicos do modelo.

3.2.3 Estimação dos Parâmetros

A estimação de parâmetros no modelo de regressão multinível, tanto dos coeficientes de regressão quanto dos componentes de variância, é feita principalmente pelo método da Máxima Verossimilhança (MV). Há duas abordagens dentro do método de

máxima verossimilhança: a Máxima Verossimilhança Completa e a Máxima Verossimilhança Restrita. A Máxima Verossimilhança Completa estima conjuntamente os coeficientes de regressão e os componentes de variância. Na Máxima Verossimilhança Restrita, a estimação é feita sequencialmente, sendo que os componentes de variância são estimados primeiro, eliminando a influência dos efeitos fixos e corrigindo o viés.

No contexto deste trabalho, foi utilizada a Máxima Verossimilhança Completa, pois facilita a seleção do modelo, permitindo a comparação direta de modelos pelos testes da razão de verossimilhança.

3.2.4 Estratégia de Análise

O modelo de regressão multinível geral de dois níveis é descrito como:

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} X_{pij} Z_{qj} + u_{pj} X_{pij} + u_{0j},$$

em que

- p é o número de variáveis explicativas X_{ij} incluídas no primeiro nível;
- q é o número de variáveis explicativas Z_j incluídas no segundo nível;
- γ_{00} é o intercepto global;
- γ_{p0} são os coeficientes fixos para o primeiro nível;
- γ_{0q} são os coeficientes fixos para o segundo nível;
- γ_{pq} são os coeficientes fixos para as interações entre o primeiro e segundo nível;
- u_{pj} são os resíduos do segundo nível associados às variáveis explicativas do primeiro nível;
- u_{0j} é o erro aleatório no nível do grupo j; e_{ij} é o erro aleatório no nível do indivíduo i no grupo j;

Segundo Hox, Moerbeek e Schoot (2017), é recomendável seguir um procedimento de seleção do modelo multinível em cinco passos, como será descrito a seguir:

Passo 1

Analisar o modelo nulo, sem variáveis explicativas e apenas com o intercepto. O modelo nulo é definido como:

$$Y_{ij} = \gamma_{00} + u_{0j} + e_{ij},$$

em que γ_{00} é o intercepto global, ou simplesmente a média geral de todos os grupos, u_{0j} é o erro aleatório no nível do grupo j e e_{ij} é o erro aleatório no nível do indivíduo i no grupo j. O modelo nulo é útil pois dá a estimativa do Coeficiente de Correlação Intraclasse (ICC):

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_e}$$

em que $\sigma^2_{u_0}$ é a variância entre os grupos u_{0j}, σ^2_e é a variância dentro dos grupos e ρ varia entre 0 e 1. Um ICC próximo de 0 indica que quase toda a variação ocorre dentro dos grupos e um ICC próximo a 1 indica que quase toda a variação ocorre entre os grupos. Se o coeficiente for alto, é apropriado usar o modelo multinível

Para o modelo de regressão logística multinível, a variância dentro dos grupos (σ^2_e) utilizada é $\pi^2/3 \approx 3{,}29$, onde $\pi \approx 3{,}14$ (FORBES et al., 2011). Dessa forma, tem-se:

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + 3{,}29}.$$

Passo 2

Analisar o modelo somente com variáveis explicativas do primeiro nível. Isso significa que os componentes de variância dos coeficientes são fixados em zero.

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + u_{0j} + e_{ij}$$

onde X_{pij} são os valores das p variáveis explicativas do primeiro nível.

Passo 3

Adicionar variáveis explicativas do 2º nível.

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{0j} + e_{ij}$$

em que Z_{qj} representa os valores das q variáveis explicativas do nível 2.

Os modelos no Passo 2 e neste passo são chamados de "modelos de componente de variância", pois decompõem a variância do intercepto em diferentes componentes hierárquicos.

Passo 4

Verificar se algum dos coeficientes de alguma variável explicativa no primeiro nível tem um componente de variância entre os grupos significativamente diferente de zero.

$$Y_{ij} = \gamma_{00} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + u_{pj} X_{pij} + u_{0j} + e_{ij}$$

em que u_{pj} são os os resíduos do nível maior dos coeficientes das variáveis explicativas do nível menor, X_{pij}.

Passo 5

Adicionar interações entre as variáveis explicativas do 2º nível e as variáveis explicativas do 1º nível que obtiveram uma variabilidade significante, verificadas no Passo 4

$$Y_{ij} = \gamma_{p0} + \gamma_{p0} X_{pij} + \gamma_{0q} Z_{qj} + \gamma_{pq} X_{pij} Z_{qj} + u_{pj} X_{pij} + u_{0j}.$$

3.2.5 Regressão Logística Multinível

Para a modelagem logística multinível, a estrutura do modelo e do preditor linear é a mesma da regressão logítica padrão vista anteriormente, porém aplicada ao contexto multinível.

Seja Y_{ij} a variável reposta com distribuição Bernoulli(π_{ij}). A forma do modelo de regressão logística multinível pode ser descrita da seguinte forma, como visto na equação 3.1.1:

$$\pi_{ij} = \frac{exp(\beta_{0j} + \beta_{pj}X_{ij})}{1 + exp(\beta_{0j} + \beta_{pj}X_{ij})},$$

em que o preditor linear, de acordo com a equação 3.1.2, é dado por

$$\eta_{ij} = logito(\pi_{ij}) = \ln\left(\frac{\pi_{ij}}{1-\pi_{ij}}\right) = \beta_{0j} + \beta_{pj}X_{ij},$$

sendo que

- $\beta_{0j} = \gamma_{00} + \gamma_{0q}Z_{qj} + u_{0j}$;
- $\beta_{pj} = \gamma_{p0} + \gamma_{pq}Z_{qj} + u_{pj}$,

com p representando o número de variáveis do primeiro nível e q o número de variáveis do segundo nível.

4 Metodologia

4.1 Banco de dados

Este estudo utiliza dados extraídos do Sistema de Informações Acadêmicas de Graduação (SIGRA) e do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) da Universidade de Brasília, os quais contêm informações sociodemográficas e acadêmicas dos estudantes dos cursos de Engenharia de Computação e Engenharia Mecatrônica.

4.1.1 Variáveis

As bases de dados possuem as seguintes variáveis:

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

4.1.2 Criação de variáveis

- Evasão

  O objetivo deste trabalho é avaliar a evasão acadêmica nos cursos de Engenharia de Computação e Engenharia Mecatrônica da Universidade de Brasília. A variável "Forma de Saída do Curso" foi utilizada como base para a construção dessa variável.

  A evasão é variável dicotômica e recebe dois valores: "Sim" quando houve evasão do curso e "Não" quando não houve evasão do curso. A Tabela 1 mostra a classificação das formas de saída do curso de acordo com o conceito de evasão.

Tabela 1: Formas de saída do os cursos de Engenharia de Computação e Engenharia Mecatrônica - UnB; 2015-2023.

Formas de saída	Evasão
Ativo	Não
Formatura	
Integralização de discente	
Novo vestibular	Sim
Mudança de curso	
Desligamento - Abandono	
Desligamento - Rendimento	
Desligamento - Voluntário	
Desligamento - Intercâmbio	
Desligamento - Não cumpriu condição	
Efetivação de novo cadastro	
Reprovou 3 vezes na mesma disc obrigatória	
Solicitação espontânea	
Falecimento	

- Idade ao ingressar

  Foi criada com base na variável "Data de Nascimento" e refere-se à idade em anos do estudante quando ingressou no curso.

- IRA

  Essa variável foi construída com base no Índice de Rendimento Acadêrmico do estudante no último semestre que cursou alguma disciplina. O objetivo é verificar de que forma o desempenho do estudante influencia na sua permanência no curso.

- Trancamentos

  Essa variável foi criada a partir do número de vezes que um estudante solicitou o trancamento de alguma disciplina. Para isso, foram consideradas as categorias TJ

e TR da variável "Menção de Disciplina".

1. TR: trancamento parcial de matrícula excepcional e justificado;
2. TJ: trancamento parcial de matrícula de concessão automático.

- Variável SR

  A variável SR registra o número de vezes em que o estudante recebeu a menção SR. Essa menção é atribuída quando há o abandono de uma disciplina, ou seja, quando o número de faltas ultrapassa 25% do total de aulas da disciplina. Foi obtida pela variável "Menção da Disciplina".

- Taxa de Reprovação

  Variável criada para avaliar a proporção de reprovações dentre o total de créditos de matérias obrigatórias cursados. Para construí-la, foram utilizadas as variáveis "Menção da Diciplina" e "Código da Disciplina". A taxa de reprovação é definida pela seguinte equação:

$$\text{Taxa de Reprovação} = \frac{\text{Quantidade de créditos obrigatórios reprovados}}{\text{Quantidade de créditos obrigatórios cursados}} \quad (4.1.1)$$

  A taxa de reprovação varia de 0 a 1, porque a quantidade de disciplinas reprovadas fazem parte do total de disciplinas cursadas.

- Cursou Verão

  A variável "cursou verão" foi criada com o objetivo de identificar se o aluno cursou alguma disciplina durante um semestre de verão. Para construí-la, foi utilizada a variável "período cursou diciplina", para verificar quais alunos cursaram disciplinas durante o verão. A variável "cursou verão" é dicotômica e assume os valores "Sim" e "Não".

- Distância até a Universidade

  A variável "Distância até a Universidade" informa a distância, em metros, entre a residência do estudante e a Universidade de Brasília (UnB). Para construí-la, foi utlizada a variável CEP, que foi convertida para o formato de coordenadas geográficas (latitude e longitude) com o objetivo de identificar a localização aproximada da residência do estudante e o Instituto Central de Ciências da UnB, que é o ponto de referência da universidade.

- Ensino remoto

  Também foi criada uma variável para identificar se o estudante cursou ao menos um semestre no remoto, devido à pandemia de Covid-19. Para a sua criação, foi utilizada a variável "Período que Cursou a Disciplina". Os semestres de ensino remoto foram 2020.1, 2020.2, 2021.1 e 2021.2,

4.1.3 Variável dos cursos de graduação

Para o nível dos cursos, será criada a variável Carga Horária (CH) obrigatória de aula. Vale ressaltar que foram selecionados somente componentes curriculares do tipo "disciplina" para a análise, excluindo Estágio, Trabalho de Conclusão de Curso e outros tipos de atividades. Para Engenharia de Computação, a CH obrigatória de aula foi de 2040h, enquanto a de Engenharia Mecatrônica foi 2850h.

Foi possível obter os dados da carga horária das disciplinas dos cursos através da Consulta de Componente Curricular, na parte de Consultas Gerais do SIGAA.

O curso de Engenharia de Computação tem o prazo mínimo para conclusão de 10 semestres, com 2460h de carga horária obrigatória total (aula e orientação acadêmica/ profissional). O curso de Engenharia Mecatrônica tem prazo mínimo para conclusão de 9 semestres, com 3270h de carga horária obrigatória total (aula e orientação acadêmica/ profissional).

4.2 Procedimento para Modelagem

Para a modelagem da evasão pelas variáveis, primeiramente serão feitos dois modelos de regressão logística: um para o curso de Engenharia da Computação e outro para Engenharia Mecatrônica. Foi ajustado, para cada curso, um modelo inicial com as variáveis explicativas apresentadas na análise descritiva e de associação. Para a seleção das covariáveis, foi utilizado o método de seleção *stepwise*. Foram realizados testes para avaliar o ajuste e a relevância estatística do modelo, assim como a inclusão de interações entre as variáveis, com o intuito de verificar se melhoravam o desempenho e ajuste. Depois, será feito um modelo de regressão logística multinível, em que os dois cursos serão analisados em conjunto e com variáveis de dois níveis: 1^{o} nível (estudante) e 2^{o} nível (curso).

O objetivo é verificar, por meio dos testes de ajuste e diagnósticos, qual das formas de modelagem é a mais adequada para o estudo da evasão e identificar os fatores que impactam de forma significativa no fenômeno da evasão.

Os dados foram divididos em treinamento e validação. O banco de dados de treinamento contém 70% dos dados, enquanto o banco de validação é composto pelos outros 30% restantes.

5 Resultados

5.1 Análise Descritiva

É importante identificar o perfil dos estudantes e as suas características sociodemográficas e acadèmicas. Para isso, será realizada uma análise descritiva das variáveis em estudo. Essa análise permitirá uma compreensão mais geral do perfil dos estudantes dos cursos de Engenharia de Computação e Engenharia Mecatrônica que ingressaram entre 2015 e 2023 na Universidade de Brasília.

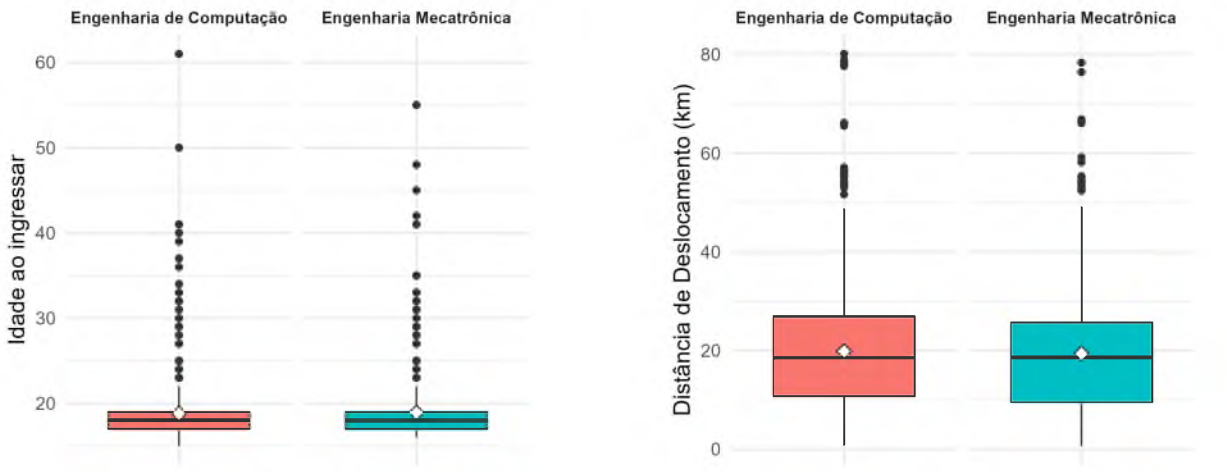
Tabela 2: Distribuição dos estudantes segundo gênero e raça/cor. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Engenharia de Computação	**Engenharia Mecatrônica**
Gênero		
Feminino	11,66% (82)	17,65% (78)
Masculino	88,34% (621)	82,35% (364)
raça/cor		
Indígena	0,14% (1)	0,23% (1)
Amarela	1,00% (7)	0,45% (2)
Branca	37,98% (267)	43,89% (194)
Parda	27,60% (194)	33,03% (146)
Preta	6,12% (43)	7,47% (33)
Não Informado	27,17% (191)	14,93% (66)

Nota-se que, em ambos os cursos, há predominância dos alunos do gènero masculino. Em Engenharia de Computação, homens representam 88,34% do total de alunos, enquanto mulheres apenas 11,66%. A diferença é menor na Engenharia Mecatrônica, embora ainda seja considerável, com 82,35% dos estudantes do gènero masculino e 17,65% do gênero feminino.

Em relação à raça/cor, os brancos representam o maior percentual, com 37,98% em Engenharia de Computação e 43,89% em Engenharia Mecatrônica. Em seguida, os pardos representam 27,60% (Engenharia de Computação) e 33,03% (Engenharia Mecatrônica). Já o percentual de pessoas autodeclaradas pretas ficou próximo de 6% e 7%. Amarelos e indígenas tiveram representatividade menor ou igual a 1% nos dois cursos. Há também estudantes da categoria que não informaram a raça/cor, com percentuais de 27,17% (Engenharia de Computação) e 14,93% (Engenharia Mecatrônica) .

Figura 1: Distribuição das variáveis Idade ao ingressar e Distância de Deslocamento (km), por curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023



Observa-se que a distribuição da idade foi similar nos dois cursos, com a média em torno de 19 anos. Há presença de um número considerável de *outliers*, com destaque para um estudante com mais de 60 anos em Engenharia de Computação.

Quanto à distância de deslocamento do local de residência dos estudantes até a Universidade de Brasília (UnB), não houve diferença significativa entre os cursos. A média e a mediana das distâncias ficam por volta de 20km, com alguns *outliers* entre 50km e 80km. Para a criação dessa variável, os estudantes com CEPs localizados fora do Distrito Federeal ou da região do Entorno tiveram as distâncias consideradas como valor nulo.

Tabela 3: Distribuição dos estudantes segundo informações de entrada na universidade. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	**Eng. Mecatrônica**
Tipo de Escola		
Escola Particular	32,86% (231)	42,31% (187)
Escola Pública	67,00% (471)	57,47% (254)
Não Informado	0,14% (1)	0,23% (1)
Forma de Ingresso na UnB		
Vestibular	24,61% (173)	28,05% (124)
PAS	38,98% (274)	38,69% (171)
ENEM	11,95% (84)	6,56% (29)
SISU	3,56% (25)	5,66% (25)
Outras formas	21,82% (147)	21,04% (93)
Utilizou cotas		
Sim	45,66% (321)	46,83% (207)
Não	54,34% (382)	53,17% (235)

Em relação ao tipo de escola de origem, estudantes oriundos de escola pública foram maioria. Em Engenharia de Computação, 67% vieram de escola pública enquanto 32,86% vieram de escola particular. Em Engenharia Mecatrônica, o percentual é mais equilibrado, com 57,47% oriundos de escola pública e 42,31% de escola particular.

Ao analisar a forma de ingresso na universidade, a principal forma de ingressar foi o Programa de Avaliação Seriada (PAS), com percentual entre 38% e 39% em ambos os cursos. Em seguida, o Vestibular, com 24,61% (Engenharia de Computação) e 28,05% (Engenharia Mecatrônica). Em terceiro lugar, observa-se outras formas de ingresso (transferências, portador de diploma de curso superior, etc.). Por fim, o ENEM e o SISU aparecem como os menores percentuais de ingresso nos cursos.

Sobre a utilização do sistema de cotas, os alunos que ingressaram por cotas foram 45,66% em Engenharia de Computação e 46,83% em Engenharia Mecatrônica.

Tabela 4: Distribuição dos estudantes segundo tipo de cota declarada. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	Eng. Mecatrônica
Tipo de Cota		
Indígena	0,14% (1)	0,45% (2)
Negro	3,98% (28)	2,71% (12)
Pública Alta Renda - PPI	14,51% (102)	8,14% (36)
Pública Baixa Renda - PPI	9,39% (66)	8,82% (39)
Pública Alta Renda - Não PPI	11,24% (79)	8,14% (36)
Pública Baixa Renda - Não PPI	6,40% (45)	6,33% (28)
Não informado	-	12,22% (54)
Universal	54,34% (382)	53,17% (235)

Nota: Sinal utilizado: – Dado numérico igual a zero não resultante de arredondamento.

Observa-se pela Tabela 4 que, entre os estudantes que utilizaram o sistema de cotas, o tipo de cota mais declarado foi o de Escola Pública Alta Renda - PPI para engenharia de computação e Escola Pública Baixa Renda - PPI para engenharia mecatrônica. Vale ressaltar que 12,22% dos alunos de engenharia mecatrônica que utilizaram cotas não informaram o tipo. Em engenharia de computação, todos declararam o tipo de cota.

Tabela 5: Distribuição dos estudantes segundo características acadêmicas. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	Eng. Mecatrônica
Cursou Verão		
Sim	10,53% (74)	25,11% (111)
Não	89,47% (629)	74,89% (331)
Ensino Remoto		
Sim	57,04% (401)	66,97% (296)
Não	42,96% (302)	33,03% (146)

No que se refere à participação em disciplinas de verão, apenas 10,53% dos estudantes cursaram em Engenharia de Computação. Na Engenharia Mecatrônica o percentual foi maior, com participação de 25,11% dos alunos nessas disciplinas.

Quanto à participação dos estudantes em períodos de ensino remoto, observa-se que a maioria cursou ao menos um semestre de ensino remoto devido à pandemia de Covid-19. Em Engenharia de Computação foram 57,04% e em Engenharia Mecatrônica

foram 66,97%.

Figura 2: Distribuição das variáveis IRA e Taxa de Reprovação, por curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

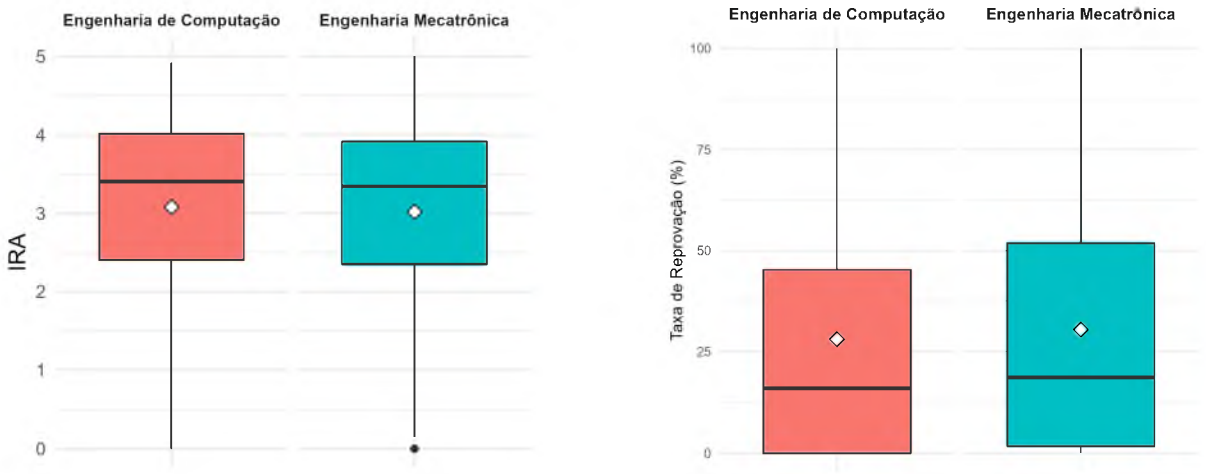


Tabela 6: Tabela de medidas resumo da distribuição do número de menções SR. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

Medidas Resumo	Eng. de Computação	Eng. Mecatrônica
Mínimo	0,00	0,00
1º Quartil	0,00	0,00
Mediana	0,00	0,00
Média	0,76	0,87
3º Quartil	1,00	1,00
Máximo	16,00	20,00

Tabela 7: Tabela de medidas resumo da distribuição do número de trancamentos. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

Medidas Resumo	Eng. de Computação	Eng. Mecatrônica
Mínimo	0,00	0,00
1º Quartil	0,00	0,00
Mediana	0,00	0,00
Média	0,19	0,23
3º Quartil	0,00	0,00
Máximo	5,00	6,00

Em relação às variăveis sobre desempenho acadêmico, tem-se que os estudantes

de Engenharia de Computação apresentaram, em média, IRA maior e taxa de reprovação menor, quando comparado aos estudantes de Engenharia Mecatrônica. O IRA médio dos cursos fica próximo de 3 e a média da taxa de reprovação é por volta de 30%. Este resultado indica que, em média, os estudantes reprovam em aproximadamente um terço dos créditos obrigatórios cursados.

O número de trancamentos e menções SR (abandono de disciplina) apresentou praticamente a mesma distribuição entre os dois cursos, com a concentração dos valores em zero. O número de trancamentos obteve média próxima a zero e o número médio de menções SR ficou pouco abaixo de 1. Foram observados alguns valores extremos, como 6 trancamentos em Engenharia de Computação e 20 menções SR.

Tabela 8: Distribuição dos estudantes segundo cocorrência de evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	**Eng. Mecatrônica**
Evasão		
Sim	27,17% (191)	33,26% (147)
Não	72,83% (512)	66,74% (295)

A evasão é a principal variável de interesse deste estudo. Ao analisar a distribuição dos estudantes segundo a ocorrência desse fenômeno, percebe-se que a evasão de alunos em Engenharia Mecatrônica (33,26%) foi maior que no curso de Engenharia de Computação (27,17%).

5.1.1 Análise Conjunta

Nessa etapa da análise, será examinada a relação entre a variável resposta do estudo (evasão) e as demais variáveis explicativas. O intuito é verificar as associações e características que podem levar à evasão dos estudantes em cada curso e, dessa forma, fornecer uma visão geral dos fatores que serão importantes na etapa da modelagem.

Tabela 9: Percentual de evasão segundo gênero e raça/cor. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

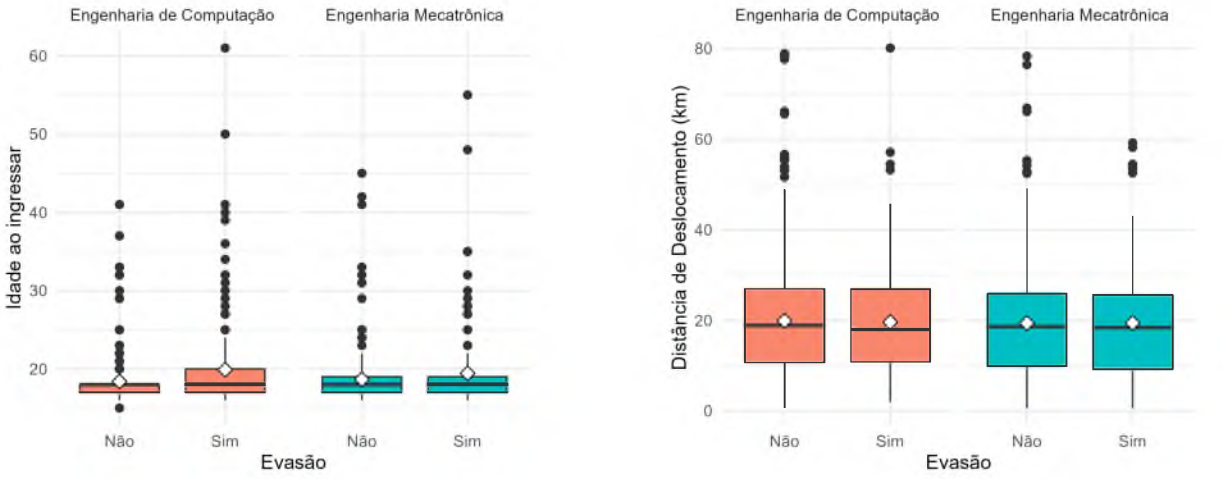
	Engenharia de Computação	**Engenharia Mecatrônica**
Gênero		
Feminino	24,39% (20)	34,62% (27)
Masculino	27,54% (171)	32,97% (120)
raça/cor		
Indígena	-	-
Amarela	-	-
Branca	26,22% (70)	29,90% (58)
Parda	29,90% (58)	34,93% (51)
Preta	32,56% (14)	54,55% (18)
Não Informado	26,65% (49)	30,30% (20)

Nota: Sinal utilizado: – Dado numérico igual a zero não resultante de arredondamento.

Em relação ao gênero dos estudantes, percebe-se que as porcentagens foram bem parecidas entre cada gênero e apresentaram resultados diferentes em cada curso. Na Engenharia de Computação, 27,54% dos homens evadiram e a evasão de mulheres foi de 24,39%. Na Engenharia Mecatrônica, em que o gênero feminino apresentou maior evasão, os valores foram bem próximos: 34,62% para mulheres e 32,97% para homens.

Referente à raça/cor, em ambos os cursos não houve evasão de estudantes autodeclarados indígenas ou amarelos. Em Engenharia de Computação, os resultados para as demais raças/cores foram os seguintes: 26,22% para os brancos, 29,90% para os pardos e 32,56% para os pretos. Em Engenharia Mecatrônica, os valores foram maiores para cada raça/cor: 29,90% (branca), 34,93% (parda) e 54,55% (preta). Portanto, observa-se que mais da metade dos estudantes autodeclarados pretos saíram do curso de Engenharia Mecatrônica.

Figura 3: Distribuição das variáveis Idade ao ingresssar e Distância de Deslocamento (km). por curso e evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023



Em relação à idade ao ingressar no curso, a média de idade dos estudantes que evadiram aparenta ser um pouco maior nos dois cursos. Não há indicativo de uma relação forte entre a evasão e a idade dos alunos.

Ao analisar a distância de deslocamento, é possível notar que não há diferença entre os estudantes que evadiram e os que permaneceram no curso. Portanto, a distância que os estudantes precisam se deslocar de seus domicílios até a Universidade de Brasília aparenta não influenciar na permanência ou abandono dos cursos.

Tabela 10: Percentual de evasão segundo informações de entrada na universidade e curso. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	**Eng. Mecatrônica**
Tipo de Escola		
Escola Particular	31,17% (72)	29,95% (56)
Escola Pública	25,05% (118)	35,43% (90)
Não Informado	100,00% (1)	100,00% (1)
Forma de Ingresso na UnB		
Vestibular	33,53% (58)	41,13% (51)
PAS	22,99% (63)	26,32% (45)
ENEM	9,52% (8)	13,79% (4)
SISU	88% (22)	88% (22)
Outras formas	27,21% (40)	26,88% (25)
Utilizou cotas		
Sim	29,91% (96)	38,65% (80)
Não	24,87% (95)	28,51% (67)

Na Tabela 10, observa-se o percentual de evasão em cada curso de acordo com as variáveis sobre a entrada dos estudantes na universidade. Estudantes que vieram de escola particular apresentaram evasão maior em Egenharia de Computação (31,17%), enquanto em Engenharia Mecatrônica alunos de escola pública tiveram maior evasão (35,43%). No que se refere a forma de ingresso, a distribuição da evasão foi bem parecida entre os cursos, com destaque para os estudantes que entraram pelo SISU, que apresentaram evasão de 88% em cada curso. As demais formas de ingresso tiveram porcentagens de evasão menores que 42%, a menor delas de ingressantes pelo ENEM.

Sobre o uso de cotas para ingresso, nota-se que alunos cotistas tiveram percentual de evasão maior nos cursos, principalmente em Engenharia Mecânica (38,65%).

Tabela 11: Percentual de evasão segundo características acadêmicas. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Eng. de Computação	**Eng. Mecatrônica**
Cursou Verão		
Sim	29,73% (22)	21,62% (24)
Não	26,87% (169)	37,16% (123)
Ensino Remoto		
Sim	7,73% (31)	12,50% (37)
Não	52,98% (160)	75,34% (111)

Observa-se da Tabela 11 que, no curso de Engenharia de Computação, alunos que cursaram disciplina durante um semestre de verão tiveram taxa de evasão praticamente iguais aos que não cursaram (29,73% e 26,87%, respectivamente). Na Engenharia Mecatrônica o resultado foi diferente: 37,16% dos estudantes que não cursaram verão evadiram, enquanto 21,62% que cursaram disciplina de verão evadiram.

Ao analisar se os estudantes cursaram ao menos uma disciplina no período de ensino remoto (pandemia de Covid-19), percebe-se que a evasão foi consideravelmente maior entre os que não estavam durante o período de ensino remoto. Para Engenharia de Computação, 52,98% dos alunos que não cursaram disciplina durante a pandemia evadiram e, para Engenharia Mecatrônica, esse valor foi de 75,34%.

Figura 4: Distribuição das variáveis IRA e Taxa de Reprovação, por curso e evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

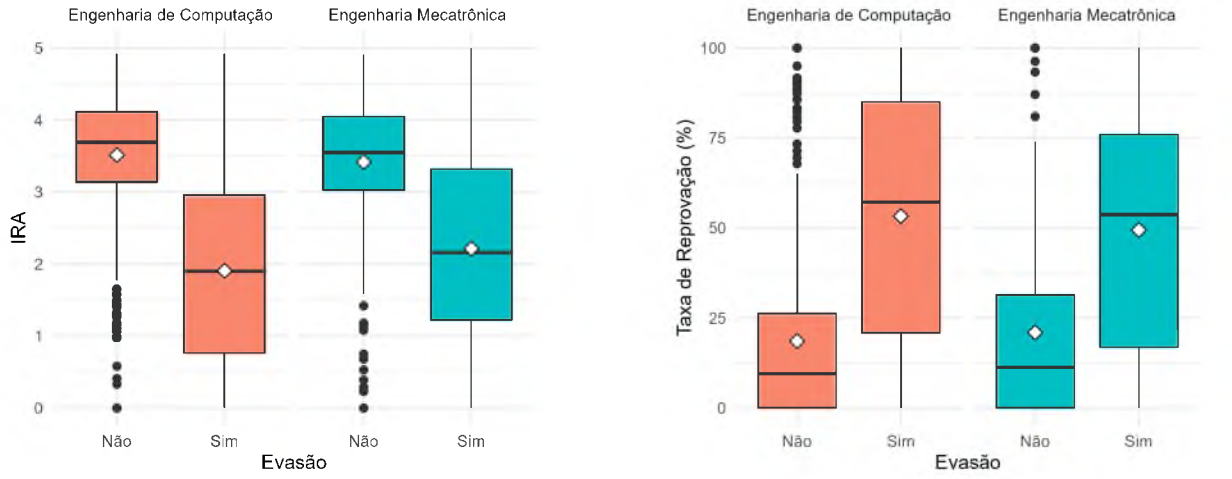


Tabela 12: Tabela de medidas resumo da distribuição do número de menções SR com relação à evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Evasão			
Medidas resumo	**Eng. de Computação**		**Eng. Mecatrônica**	
	Não	**Sim**	**Não**	**Sim**
Mínimo	0,00	0,00	0,00	0,00
1º Quartil	0,00	0,00	0,00	0,00
Mediana	0,00	1,00	0,00	1,00
Média	0,51	1,86	0,52	1,26
3º Quartil	0,00	3,00	1,00	2,00
Máximo	13,00	20,00	9,00	16,00

Tabela 13: Tabela de medidas resumo da distribuição do número de trancamentos com relação à evasão. Engenharia de Computação e Engenharia Mecatrônica - UnB, 2015-2023

	Evasão			
Medidas resumo	**Eng. de Computação**		**Eng. Mecatrônica**	
	Não	**Sim**	**Não**	**Sim**
Mínimo	0,00	0,00	0,00	0,00
1º Quartil	0,00	0,00	0,00	0,00
Mediana	0,00	0,00	0,00	0,00
Média	0,03	0,76	0,02	0,54
3º Quartil	0,00	1,00	0,00	1,00
Máximo	6,00	6,00	2,00	5,00

De acordo com a Figura 4, oberva-se que há diferença no IRA entre os estudantes que abandonaram os cursos e os que pernaeceram, com valores de IRA menores entre os alunos que evadiram. Na taxa de reprovação, a evasão parece estar associada à maiores taxas em ambos os cursos. Além disso, os alunos que evadiram obtiveram maior número de trancamentos e de menções SR ao longo da graduação. Logo, baixos desempenhos acadêmicos, números altos de trancamentos e o abandono de disciplina aparentam estar diretamente relacionados à evasão.

5.1.2 Análise de associação entre as variáveis explicativas e a evasão

A Tabela 14 apresenta os resultados dos testes de associação realizados entre as variáveis explicativas e a variável resposta (evasão), com os valores da estatísticas de teste e os p-valores. Para as variáveis qualitativas, foi realizado o teste qui-quadrado de Pearson. Para variáveis quantitativas, foram empregados modelos de regressão logística simples. Esta análise é importante para verificar quais as possíveis variáveis que serão incluídas na modelagem.

Tabela 14: Associação das variáveis com Evasão

Associação com Evasão	**Engenharia de Computação**		**Engenharia Mecatrônica**	
	Estatística	p-valor	Estatística	p-valor
IRA	-12,58	<0,001	-8,87	<0,001
Gênero	0,22	0,639	0,02	0,882
Raça/Cor	1,18	0,277	2,83	0,093
Idade ao ingressar	4,15	<0,001	1,78	0,075
Forma de Ingresso	65,92	<0,001	47,58	<0,001
Distância	-0,23	0,820	0,05	0,963
Tipo Escola	2,64	0,105	1,23	0,268
Cotas	1,99	0,158	4,65	0,031
Esino Remoto	175,97	<0,001	171,13	<0,001
Cursou Verão	0,15	0,700	8,36	0,004
Taxa de reprovação	11,48	<0,001	8,116	<0,001
Menções SR	7,24	<0,001	4,46	<0,001
Trancamentos	6,23	<0,001	4,76	<0,001

Pelos resultados dos testes, as variáveis que possuem associação estatisticamente significativa ($p < 0{,}05$) nos dois cursos foram: Índice de Rendimento Acadêmico (IRA),

forma de ingresso, cursar ao menos uma disciplina no ensino remoto, taxa de reprovação, número de menções SR e número de trancamentos. A idade ao ingressar teve associação com a variável resposta somente em Engenharia de Computação. O uso de cotas para ingresso e cursar disciplina de verão só estivaram associadas à evasão em Engenharia Mecatrônica.

5.2 Associação entre as variáveis explicativas

É importante que seja avaliada a relação entre as variáveis explicativas antes da realização da modelagem. A multicolinearidade, que é a correlação entre as variáveis explicativas, pode inflacionar o erro padrão dos estimatores e prejudicar o ajuste do modelo.

Para analisar a correlação entre as variáveis quantitativas, foi utilizado o coeficiente de correlação de Spearman. Ao realizar essa análise, verificou-se que a variável IRA (Índice de Rendimento Acadêmico) tem forte correlação negativa com a taxa de reprovação, com $\rho = -0,830$ em Engenharia de Computação e $\rho = -0,876$ em Engenharia Mecatrônica. Além disso, o número de menções SR tem correlação moderada com as variáveis IRA e taxa de reprovação.

Para analisar a associação entre as variáveis qualitativas, foi utilizado o teste qui-quadrado de Pearson. Para os dois cursos, houve associação entre a forma de ingresso no curso e ensino remoto, com $p < 0,001$.

5.3 Modelagem Estatística

Nesta etapa, o objetivo é identificar os principais fatores que estão associados à evasão nos cursos de Engenharia de Computação e Engenharia Mecatrônica, através de modelos de regressão logística.

Para evitar problemas devido a baixas frequências em algumas categorias, a variável raça/cor precisou ser reconfigurada em dois grupos: um grupo com as cores Branca e Amarela; outro grupo com as cores Preta, Parda e Indígena.

Nas análises anteriores, foi identificada uma correlação forte entre as variáveis IRA e taxa de reprovação. Com o objetivo de evitar problemas de multicolinearidade, foram ajustados modelos diferentes com apenas uma dessas duas variáveis.

5.3.1 Modelo do curso de Engenharia de Computação

Foram ajustados dois modelos iniciais: um com a variável IRA e outro com a Taxa de Reprovação.

Tabela 15: Modelo com IRA do *Stepwise* para o curso de Engenharia de Computação com os dados de treino

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	2,7461	1,4265	1,925	0,0542
IRA	-1,2788	0,1928	-6,633	<0,0001
Distância deslocamento	-0,0414	0,0184	-2,251	0,0244
Trancamentos	0,6438	0,2792	2,306	0,0211
Raça/Cor (Parda/Preta/Indígena)	-0,7529	0,4437	-1,697	0,0897
Ingresso (Outras Formas)	2,5369	1,2619	2,010	0,0444
Ingresso (PAS)	2,3193	1,1820	1,962	0,0498
Ingresso (SISU)	4,3686	1,6277	2,684	0,0073
Ingresso (Vestibular)	2,0937	1,2199	1,716	0,0861
Tipo escola (Pública)	-0,9083	0,4513	-2,012	0,0442
Ensino remoto (Sim)	-2,6119	0,4709	-5,547	<0,0001

Tabela 16: Modelo com Taxa de Reprovação do *Stepwise* para o curso de Engenharia de Computação com os dados de treino

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	-3,3454	1,6970	-1,971	0,0487
Taxa de reprovação	3,1798	0,7100	4,478	<0,0001
Idade de ingresso	0,0802	0,0545	1,472	0,1412
Distância deslocamento	-0,0385	0,0176	-2,181	0,0292
Menções SR	0,1853	0,0928	1,997	0,0458
Trancamentos	0,7363	0,2664	2,763	0,0057
Raça/Cor (Parda/Preta/Indígena)	-0,7684	0,4342	-1,770	0,0768
Ingresso (Outras Formas)	1,7841	1,2730	1,402	0,1611
Ingresso (PAS)	2,1878	1,2286	1,781	0,0750
Ingresso (SISU)	4,1469	1,6592	2,499	0,0124
Ingresso (Vestibular)	1,8666	1,2522	1,491	0,1361
Tipo escola (Pública)	-1,1034	0,4362	-2,530	0,0114
Ensino remoto (Sim)	-2,9281	0,5178	-5,655	<0,0001

Observa-se que ambos os modelos apresentaram variáveis em comum: Trancamentos, Raça, Forma de ingresso, Tipo de escola, Ensino Remoto e Distância de deslocamento. O modelo com Taxa de Reprovação apresentou variáveis diferentes do modelo com IRA: a própria Taxa de Reprovação, o número de Menções SR e a Idade de ingresso. O AIC do modelo com IRA foi de 194,98 e o do modelo com Taxa de Reprovação foi 209,99.

Com o objetivo de verificar qual modelo se ajusta melhor aos dados, foi realizada a inclusão e retirada de variáveis, assim como a inclusão de interações. Nenhum desses processos resultou em um modelo melhor que os modelos ajustados pelo método *stepwise*. Portanto, ao levar em conta os critérios de seleção como o *AIC*, a adequação do ajuste e a análise dos resíduos, o melhor modelo obtido foi o que possui o IRA. Além disso, o IRA é uma variável mais completa e informativa que a Taxa de Reprovação, pois leva em conta as notas de cada menção e atribui pesos às disciplinas.

Para visualizar os gráficos dos resíduos e do envelope do modelo com a Taxa de Reprovação, consulte a Figura 11, disponível no Apêndice A deste trabalho.

Tabela 17: Modelo final para o curso de Engenharia de Computação

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	2,7461	1,4265	1,925	0,0542
IRA	-1,2788	0,1928	-6,633	<0,0001
Distância deslocamento	-0,0414	0,0184	-2,251	0,0244
Trancamentos	0,6438	0,2792	2,306	0,0211
Raça/Cor (Parda/Preta/Indígena)	-0,7529	0,4437	-1,697	0,0897
Ingresso (Outras Formas)	2,5369	1,2619	2,010	0,0444
Ingresso (PAS)	2,3193	1,1820	1,962	0,0498
Ingresso (SISU)	4,3686	1,6277	2,684	0,0073
Ingresso (Vestibular)	2,0937	1,2199	1,716	0,0861
Tipo escola (Pública)	-0,9083	0,4513	-2,012	0,0442
Ensino remoto (Sim)	-2,6119	0,4709	-5,547	<0,0001

O modelo final apresenta as variáveis IRA, Distância de deslocamento, número de Trancamentos, Raça, Forma de ingresso, Tipo de escola e Ensino remoto. A Raça/Cor (Parda/Preta/Indígena) e a Forma de ingresso (Vestibular) são os únicos coeficientes que apresentaram p-valor maior que o nível de significância de 5%, mas ainda abaixo de 10%. O AIC do modelo é 194,98.

Tabela 18: Razão de chances e IC 95% - Modelo final Engenharia de Computação

Variável Explicativa	**Razão de Chances**	**IC 95%**
IRA	0,278	[0,186; 0,397]
Distância deslocamento	0,959	[0,924; 0,994]
Trancamentos	1,904	[1,188; 3,628]
Raça/Cor (Parda/Preta/Indígena)	0,471	[0,191; 1,103]
Ingresso (Outras Formas)	12,640	[1,489; 301,232]
Ingresso (PAS)	10,168	[1,478; 220,463]
Ingresso (SISU)	78,932	[4,911; 3818,326]
Ingresso (Vestibular)	8,115	[1,048; 182,789]
Tipo escola (Pública)	0,403	[0,163; 0,966]
Ensino remoto (Sim)	0,073	[0,027; 0,176]

Ao analisar a tabela da razão de chances e IC de 95% de confiança, percebe-se que o IRA reduz em 72,20% a chance de evasão a cada ponto de acréscimo nesse índice. Portanto, o rendimento acadêmico alto favorece a permanência dos estudantes no curso. O aumento da distância de deslocamento reduz a chance de evasão, embora a

redução seja pequena, de aproximadamente 4% a cada quilômetro a mais. Já o número de trancamentos aumenta a chance de abandono do curso, com 90,40% de aumento a cada trancamento adicional.

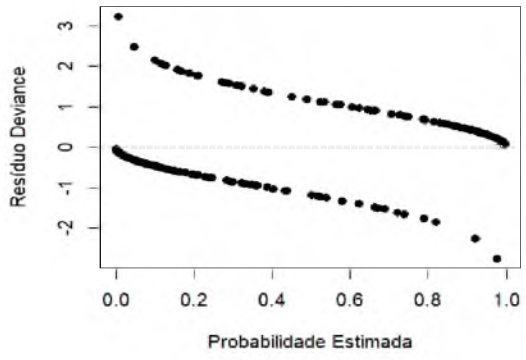
Pardos, pretos e indígenas apresentam 52,90% menos chance de evadir do que os brancos e amarelos. Ao analisar a forma de ingresso na UnB, verifica-se que todas as formas de ingresso aumentam a chance de evasão quando comparadas á categoria de referência (ENEM). O destaque fica para o ingresso pelo SISU, que aumenta em 78,932 vezes a chance de evasão em relação ao ingresso pelo ENEM. O SISU, além de ser um sistema nacional que permite estudantes de outros estados ingressarem, permite também o ingresso pela segunda opção de curso. Tanto o fator geográfico quanto a falta de clareza na escolha do curso podem estar associados ao aumento da propensão a evasão por essa forma de entrada.

Alunos de escola pública têm menor chance de evadir do que os alunos de escola particular, com diminuição de 59,70% na chance. Os alunos de engenharia de computação que cursaram ao menos uma disciplina no ensino remoto também apresentam menor chances de evadir, com 92,70% a menos de chance que os estudantes que não cursaram nesse período. Isso provavelmente se deve às flexibilizações ocorridas no período pandêmico, como a opção de retirada automática de disciplinas até o último dia letivo, além de interromper os critérios de desligamentos.

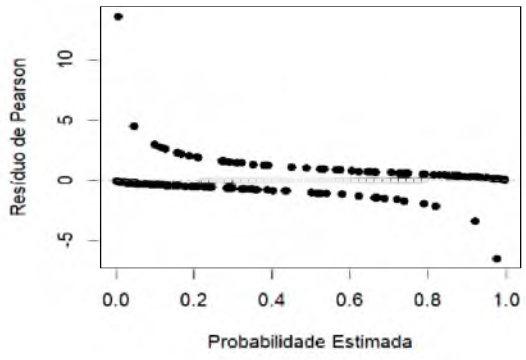
Tabela 19: Teste de adequação de ajuste - Modelo Engenharia de Computação

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	7,2236	0,5127

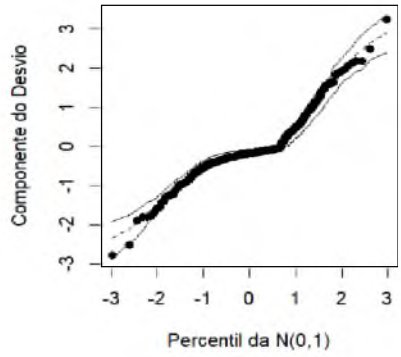
O teste de Hosmer Lemeshow teve estatística de 7,2236 e p-valor de 0,5127. O p-valor foi maior que o nível de significância de 5% e não se rejeita a hipótese nula de que o modelo teve bom ajuste aos dados. Logo, o modelo final se adequou bem aos dados.



(a) Resíduos *Deviance*



(b) Resíduos de *Pearson*



(c) Gráfico de envelope para os resíduos

Figura 5: Gráficos de análise para os resíduos do Modelo de Engenharia de Computação

O gráfico dos resíduos *deviance* demonstra uma distribuição simétrica em torno de zero. O gráfico dos resíduos de *Pearson* apresenta distribuição sem tendência e próxima de zero. Esses gráficos, em conjunto com o envelope para os resíduos, indicam que não há grandes desvios e que o modelo está bem ajustado aos dados.

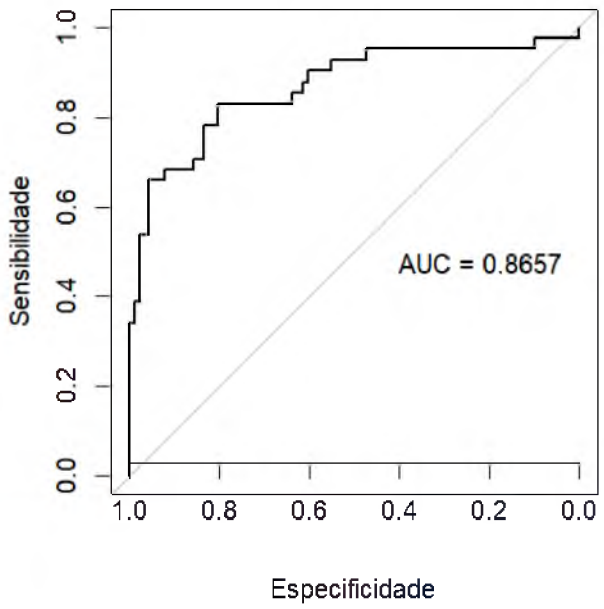


Figura 6: Curva ROC - Modelo de Engenharia de Computação

A curva ROC é uma ferramenta utilizada para verificar a capacidade preditiva de um modelo de classificação. Neste caso, em que o modelo foi utilizado para previsão sobre os dados de validação, a curva ROC demonstrou mostra boa capacidade preditiva do modelo com AUC = 0,8657. O ponto de corte utilizado é de 0,5.

Tabela 20: Matriz de Confusão - Modelo Engenharia de Computação

	Observado	
	$Y = 1$	$Y = 0$
Previsto $\hat{Y} = 1$	24	4
Previsto $\hat{Y} = 0$	17	87

A matriz de confusão mostra que a acurácia do modelo foi de 84,09%, sensibilidade de 58,54% e especificidade de 95,60%. Portanto, o modelo apresenta uma excelente capacidade de identificar estudantes que não evadiram, mas uma capacidade menor de identificar os que evadiram.

5.3.2 Modelo logístico do curso de Engenharia Mecatrônica

Para o curso de Engenharia Mecatrônica também foram construídos dois modelos iniciais: um com a variável IRA e outro com a Taxa de Reprovação.

Tabela 21: Modelo com IRA do *Stepwise* para o curso de Engenharia Mecatrônica com os dados de treino

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	2,0892	0,5621	3,7170	0,0002
IRA	-0,4618	0,1750	-2,6380	0,0083
Trancamentos	0,8847	0,6264	1,4130	0,1578
Remotos (Sim)	-2,6454	0,4064	-6,5100	<0,0001

Tabela 22: Modelo sem IRA do *Stepwise* para o curso de Engenharia Mecatrônica com os dados de treino

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	0,7733	0,3102	2,4930	0,0127
Mencões SR	0,2103	0,1187	1,7720	0,0764
Trancamentos	0,9051	0,6577	1,3760	0,1688
Remotos (Sim)	-3,0311	0,3949	-7,6750	<0,0001

Novamente, foi realizado o *stepwise* em dois modelos diferentes, um com IRA e outro com a Taxa de Evasão. Percebe-se que o número de variáveis selecionadas foi menor no curso de Engenharia Mecatrônica do que em Engenharia de Computação.

No modelo sem IRA, verifica-se que as variáveis selecionadas foram o número de Menções SR, o número de Trancamentos e Ensino remoto. A Taxa de Reprovação não foi incluída. O AIC para esse modelo foi de 192,97. Já no modelo com IRA, além do próprio índice, também foram incluídos o número de Trancamentos e o Ensino remoto. O AIC foi de 188,77.

Para verificar qual modelo se ajusta melhor aos dados, foi realizada a inclusão e retirada de variáveis, assim como a inclusão de interações. Novamente, nenhum desses procedimentos resultou em um modelo melhor que os modelos ajustados pelo método *stepwise*. Portanto, ao levar em conta os critérios de seleção como o *AIC*, a adequação do ajuste e a análise dos resíduos, o melhor modelo obtido foi o que possui o IRA.

Para visualizar os gráficos dos resíduos e do envelope do modelo sem o IRA, consulte a Figura 12, disponível no Apêndice B deste trabalho.

Tabela 23: Modelo final para o curso de Engenharia Mecatrônica

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	2,0892	0,5621	3,7170	0,0002
IRA	-0,4618	0,1750	-2,6380	0,0083
Trancamentos	0,8847	0,6264	1,4130	0,1578
Remotos (Sim)	-2,6454	0,4064	-6,5100	<0,0001

O modelo final apresenta as variáveis IRA, número de Trancamentos e Ensino remoto. A variável Número de trancamentos apresentou p-valor de 0,1578, maior que o nível de significância de 5%. O p-valor das demais variáveis foi abaixo de 5%. O AIC para esse modelo é 188,77.

Tabela 24: Razão de chances e IC 95% - Modelo final Engenharia Mecatrônica

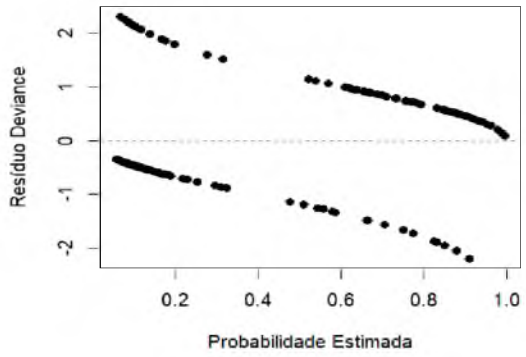
Variável Explicativa	**Razão de Chances**	**IC 95%**
IRA	0,630	[0,444; 0,886]
Trancamentos	2,422	[0,990; 13,374]
Remotos (Sim)	0,071	[0,031; 0,155]

Pela tabela de razão de chances e IC 95%, nota-se que o IRA reduz em 37% a chance de evasão a cada aumento de unidade deste índice. Ou seja, para Engenharia Mecatrônica um bom rendimento acadêmico também significa maior chance de permanência no curso. Já o número de trancamentos aumenta em 2,422 vezes a chance de evasão a cada trancamento adicional. Por fim, os estudantes que cursaram ao menos uma disciplina no ensino remoto apresentam 92,90% menos chance de evadirem quando comparados aos estudantes que não cursaram nesse período. Mais uma vez, as flexibilizações durante a pandemia podem ter influenciado na permanência dos estudantes no curso.

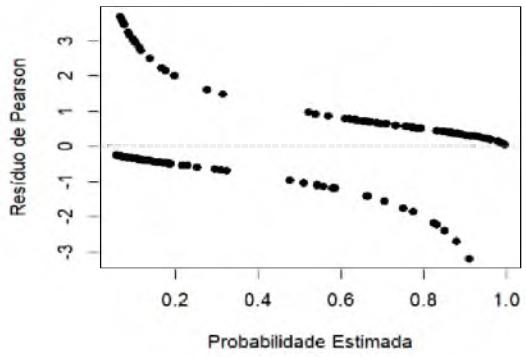
Tabela 25: Teste de adequação de ajuste - Modelo Engenharia Mecatrônica

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	4,2441	0,8345

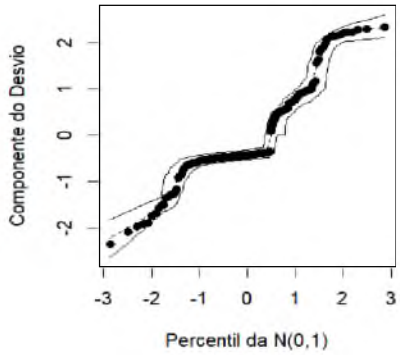
A estatística do teste de Hosmer e Lemeshow foi 4,2441 e o p-valor foi de 0,8345. Como o p-valor foi maior que o nível de significância de 5%, não se rejeita a hipótese nula de que o modelo teve bom ajuste aos dados. Portanto, o modelo final apresentou boa adequação aos dados.



(a) Resíduos *Deviance*



(b) Resíduos de *Pearson*



(c) Gráfico de envelope para os resíduos

Figura 7: Gráficos de análise para os resíduos do Modelo de Engenharia Mecatrônica

O gráfico dos resíduos *deviance* demonstra uma distribuição simétrica em torno de zero. O gráfico dos resíduos de *Pearson* apresenta distribuição sem tendência e próxima de zero. Esses gráficos, em conjunto com o envelope para os resíduos, indicam que não há grandes desvios e que o modelo está bem ajustado aos dados.

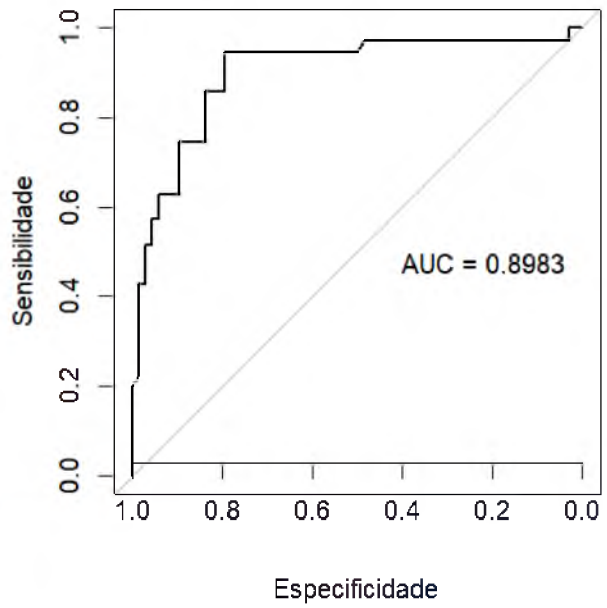


Figura 8: Curva ROC - Modelo de Engenharia Mecatrônica

A curva ROC do modelo ajustado tem AUC = 0,8983, o que indica uma boa capacidade preditiva. Logo, o modelo se mostrou eficaz para discrimar os dados de evasão. O ponto de corte utilizado é de 0,5.

Tabela 26: Matriz de Confusão - Modelo Engenharia Mecatrônica

	Observado	
	$Y = 1$	$Y = 0$
Previsto $\hat{Y} = 1$	26	11
Previsto $\hat{Y} = 0$	9	57

Observa-se pela matriz de confusão que o modelo teve bom desempenho preditivo. A acurácia foi de 80,58%, com sensibilidade de 74,29% e especificidade de 83,82%. Mais uma vez, a sensibilidade apresentou valor menor que as demais medidas, mas o modelo foi eficaz na classificação correta dos casos de evasão.

5.3.3 Modelo logístico multinível

Para a construção do modelo multinível, foi necessário fazer a junção das bases de dados dos dois cursos. Juntou-se as duas bases de treinamento para que haja uma base de treinamento completa dos dois cursos. Para a base de validação, seguiu-se o mesmo procedimento.

Primeiramente, foi construído o modelo nulo, sem variáveis explciativas, para obter o Coeficiente de Correlação Intraclasse. Em seguida, foi construído um modelo com as variáveis explicativas do 1º nível (estudante). Por fim, foi feito um modelo com variáveis do 1º nível e do 2º nível (curso).

Tabela 27: Modelo sem variáveis explicativas - Modelo Nulo

Parâmetro	**Estimativa**	**Erro padrão**	**p-valor**
Efeito Fixo			
Intercepto	-0,8800	0,1334	<0,0001
Efeito Aleatório			
Variância entre os cursos	0,0229	0,1512	-
Correlação intraclasse (ICC)	0,0069	-	-

A estimativa da variância entre os cursos, ou do intercepto, foi de 0,0229, com desvio padrão de 0,1512. O modelo nulo também fornece a estimativa do Coeficiente de Correlação Intraclasse (ICC), que nesse caso foi 0,0069. Logo, 0,69% da variação do modelo está relacionada à variabilidade entre os cursos. Esse valor é consideravelmente baixo, indicando que não há necessidade da utilização de uma estrutura multinível. Apesar dessa indicação, o modelo multinível será ajustado, porque inclui os dois cursos na modelagem, aumenta o tamanho da amostra e possibilita a inclusão da variável do nível do curso, que é a Carga Horária Obrigatória de Aula.

Tabela 28: Modelo com variáveis explicativas do 1º nível

Parâmetro	**Estimativa**	**Erro padrão**	**p-valor**
Efeito Fixo			
Intercepto	2,0316	1,0339	0,0495
IRA	-0,9320	0,1257	<0,0001
Ensino Remoto (Sim)	-2,2598	0,2963	<0,0001
Raça/Cor (Parda/Preta/Indígena)	-0,1511	0,2918	0,6046
Trancamentos	0,7466	0,2568	0,0036
Distância deslocamento	-0,0177	0,0112	0,1115
Ingresso (Outras Formas)	1,5565	0,8690	0,0733
Ingresso (PAS)	1,6340	0,8189	0,0460
Ingresso (SISU)	3,0492	1,1409	0,0075
Ingresso (Vestibular)	1,6197	0,8443	0,0550
Tipo escola (Pública)	-0,8117	0,3062	0,0080
Efeito Aleatório			
Variância entre os cursos	0,1637	0,4047	-

A Tabela 28 apresenta o modelo com as variáveis explicativas do 1º nível, o nível dos estudantes. A seleção dessas variáveis foi feita com a inclusão de uma variável por vez, observando o Critério de Informação de Akaike (AIC) e realizando o teste da razão de verossimilhança. O AIC do modelo é 393,5. A variância entre os cursos aumentou em relação ao modelo nulo, com valor de 0,1637.

Tabela 29: Modelo com variáveis explicativas do 1º nível e do 2º nível

Parâmetro	**Estimativa**	**Erro padrão**	**p-valor**
Efeito Fixo			
Intercepto	-0,6451	1,2339	0,6011
IRA	-0,9329	0,1264	<0,0001
Ensino Remoto (Sim)	-2,2957	0,2995	<0,0001
Raça/Cor (Parda/Preta/Indígena)	-0,1569	0,2929	0,5922
Trancamentos	0,7547	0,2571	0,0033
Distância deslocamento	-0,0184	0,0112	0,1010
Ingresso (Outras Formas)	1,5059	0,8671	0,0824
Ingresso (PAS)	1,5892	0,8170	0,0517
Ingresso (SISU)	2,9877	1,1395	0,0087
Ingresso (Vestibular)	1,5536	0,8421	0,0650
Tipo escola (Pública)	-0,7903	0,3063	0,0099
CH Obrigatória de Aula	0,0011	0,0004	0,0024
Efeito Aleatório			
Variância entre os cursos	$1{,}22 \times 10^{-16}$	$1{,}11 \times 10^{-8}$	-

O modelo observado na Tabela 29, além de conter as variáveis do 1º nível, acrescenta a variável explicativa do 2º nível. Esta variável é a Carga Horária Obrigatória de Aula dos cursos, correspondente a 2040h para Engenharia de Computação e 2850h para Engenharia Mecatrônica. A estimativa e o desvio padrão da Variância entre os cursos obtiveram valores menores que 0,0001. O AIC foi de 390, indicando melhor ajuste em relação ao modelo que considera apenas as variáveis do 1º nível.

Tabela 30: Razão de chances e IC 95% - Modelo Multinível

Variável Explicativa	**Razão de Chances**	**IC 95%**
Intercepto	0,525	[0,047; 5,890]
IRA	0,393	[0,307; 0,504]
Ensino Remoto (Sim)	0,101	[0,056; 0,181]
Raça/Cor (Parda/Preta/Indígena)	0,855	[0,481; 1,518]
Trancamentos	2,126	[1,285; 3,521]
Distância deslocamento	0,982	[0,960; 1,004]
Ingresso (Outras Formas)	4,509	[0,824; 24,664]
Ingresso (PAS)	4,899	[0,988; 24,298]
Ingresso (SISU)	19,846	[2,126; 185,125]
Ingresso (Vestibular)	4,730	[0,908; 24,631]
Tipo escola (Pública)	0,454	[0,249; 0,827]
CH Obrigatória de Aula	1,001	[1,000; 1,002]

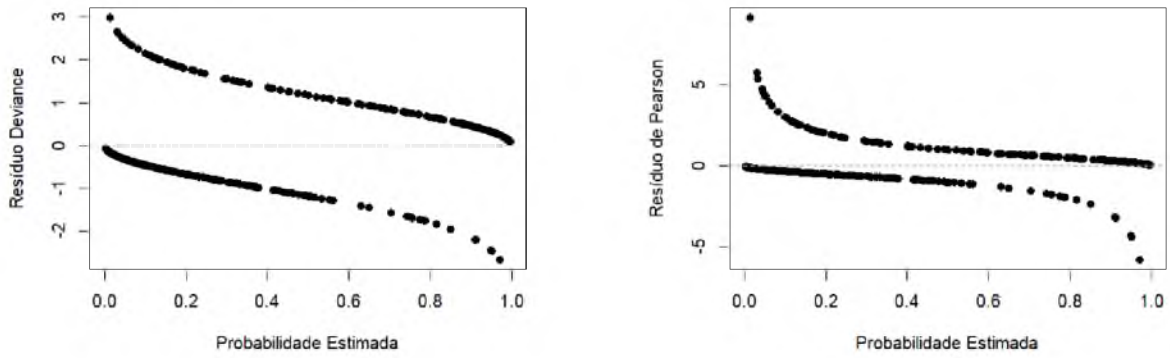
No modelo multinível ajustado, percebe-se que o IRA reduz em 60,30% a chance de evasão a cada ponto adicional deste índice. Estudantes que cursaram ao menos uma disciplina no ensino remoto têm 89,90% de redução na chance de evasão em comparação com os estudantes que não cursaram neste período. A raça dos alunos também foi um fator de redução da evasão, com estudantes das cores Parda/Preta/Indígena apresentando 14,50% menos chance comparados aos alunos de cor Branca/Amarela. No entanto, é importante destacar que a raça/cor não foi uma variável significativa no modelo, com p-valor de aproximadamente 0,6.

Uma variável que se mostrou relevante somente no modelo multinível foi a Distância de deslocamento, que nesse caso reduz em 1,80% a chance de os alunos evadirem a cada quilômetro adicional. Vale ressaltar que o p-valor da estimativa dessa variável foi maior que o nível de significância de 5%, porém próximo de 10%.

Ao analisar a razão de chances das formas de ingresso, nota-se que o ingresso pelo SISU aumentou em quase 20 vezes a chance de evasão em relação ao ingresso pelo ENEM. As demais formas de ingresso têm chances de evasão entre 4,5 e 5 vezes maiores que o ENEM. Em relação ao tipo de escola, alunos que vieram de escola pública têm 54,60% menos chance de evadir em comparação aos alunos de escola particular. Observa-se também que a chance de evasão aumenta em 2,126 vezes a cada trancamento realizado pelo estudante. Assim, quanto maior o número de trancamentos, maior a evasão.

A Carga Horária de aula obrigatória, variável do nível do curso, também está

relacionada a uma maior evasão. Há um aumento de 0,1% na chance de evasão a cada hora de aula acrescentada.



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

Figura 9: Gráficos de análise para os resíduos do Modelo Multinível

Os gráficos da Figura 9, referentes aos resíduos *Deviance* e de *Pearson*, exibem distribuições relativamente simétricas em torno de zero e sem grandes desvios.

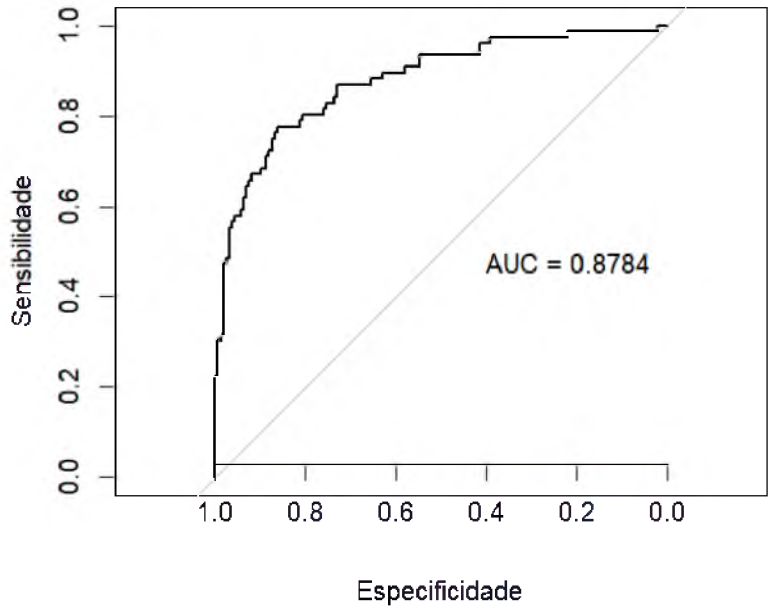


Figura 10: Curva ROC - Modelo Multinível

A Figura 10 apresenta a curva ROC do modelo multinível com as variáveis explicativas dos dois níveis. A área sob a curva (AUC) foi de 0,8784, indicando que o modelo

foi eficaz para discriminar os dados de evasão. O ponto de corte utilizado é de 0,5.

Tabela 31: Matriz de Confusão - Modelo Multinível

	Observado	
	$Y = 1$	$Y = 0$
Previsto $\hat{Y} = 1$	50	13
Previsto $\hat{Y} = 0$	26	146

A matriz de confusão revela uma acurácia de 83,40%, sensibilidade de 65,79% e especificidade de 91,82%. Esses resultados são um indicativo de que o modelo possui boa capacidade para identificar estudantes que evadiram e uma excelente capacidade para identificar estudantes que não evadiram.

6 Conclusão

Este estudo teve como objetivo investigar os principais fatores associados à evasão nos cursos de Engenharia de Computação e Engenharia Mecatrônica, através de análises descritivas e modelos de regressão logística, incluindo a abordagem multinível que avalia conjuntamente os dois cursos.

Primeiramente, verificou-se as características sociodemográficas dos estudantes, indicando que a maior parte deles é do gênero masculino, com raça/cor predominante branca e parda, idade de ingresso em torno de 18 e 19 anos e distância média de deslocamento por volta de 20km. Quanto às variáveis acadêmicas, os principais fatores são o Índice de Rendimento Acadêmico (IRA) no último semestre, o número de trancamentos e ter cursado disciplina durante o ensino remoto.

Em relação à modelagem, foram ajustados modelos de regressão logística separadamente para cada curso, com seleção de variáveis via processo *stepwise*. Também foi feita a inclusão e exclusão de variáveis individualmente, com o objetivo de verificar se havia melhoria no ajuste ou interação entre as variáveis. Além disso, foi feito um modelo multinível de dois níveis, para verificar se essa abordagem é adequada ao conjunto de dados e se melhora a análise e interpretação do modelo.

No curso de Engenharia de Computação, o modelo selecionado possui IRA, distância de deslocamento, número de trancamentos, raça, forma de ingresso, tipo de escola e ensino remoto. Verifica-se que estudantes com menor IRA e maior número de trancamentos têm propensão maior à evasão. O fato de ter estudado ao menos uma disciplina durante o ensino remoto reduz a chance de evasão, dadas as flexibilizações que ocorreram nesse período. Destaca-se também que todas as formas de ingresso apresentaram maior chance de evasão quando comparadas ao ingresso pelo ENEM, principalmente ao ingressar pelo SISU. O modelo obteve bom desempenho preditivo, com AUC de 0,8657 e acurácia de 84,09%.

No curso de Engenharia Mecatrônica, somente as variáveis IRA, trancamentos e ensino remoto foram incluídas no modelo final. Mais uma vez, um rendimento acadêmico alto e ter cursado pelo menos uma disciplina no ensino remoto contribuem para menor chance de evasão. Um número alto de trancamentos, por sua vez, aumenta a chance de os estudantes evadirem do curso. O modelo teve boa capacidade preditiva, com AUC de 0,8983 e acurácia de 80,58%.

Por fim, ajustou-se um modelo multinível de dois níveis, com 1º nível (aluno) e 2º

nível (curso). Vale ressaltar que o Coeficiente de Correlação Intraclasse teve um valor consideravelmente baixo ($\rho = 0{,}0069$), indicando que a abordagem multinível não se mostra a mais necessária para o conjunto de dados. As variáveis explicativas do 1º nível selecionadas foram: IRA, ensino remoto, raça, trancamentos, distância de deslocamento, forma de ingresso e tipo de escola, todas com interpretações similares aos modelos anteriores. A variável de 2º nível utilizada foi a carga horária de aula obrigatória, que demonstrou um pequeno aumento na chance de evasão a cada hora adicional. O modelo apresentou um bom ajuste e capacidade preditiva parecida com a dos modelos de regressão ajustados anteriormente para cada curso.

Observa-se que o modelo de Engenharia Mecatrônica possui as variáveis IRA, número de trancamentos e se o aluno cursou ao menos uma disciplina durante o ensino remoto. O modelo de Engenharia de Computação possui, além dessas três variáveis, a distância de deslocamento, a raça/cor, a forma de ingresso e o tipo de escola. Já o modelo multinível possui, além das variáveis do Modelo de Engenharia de Computação, a carga horária obrigatória de aula. A interpretação das variáveis foi semelhante entre todos os modelos.

Para estudos futuros, sugere-se a inclusão de mais cursos de graduação na análise, que consigam capturar melhor a estrutura hierárquica dos dados. Outras sugestões são a inclusão de variáveis socioeconômicas e a utilização de outras técnicas modernas de classificação, como as florestas aleatórias, redes neurais e *Support Vector Machine* (SVM).

Referências

AGRESTI, A. *An introduction to categorical data analysis*. [S.l.]: John Wiley & Sons, 2007.

ANDIFES, A.; ABRUEM, A. et al. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas: resumo do relatório apresentado a andifes, abruem e sesu/mec pela comissão especial. *Avaliação: Revista da Avaliação da Educação Superior*, v. 1, n. 2, 1996.

FILHO, R. L. L. S. et al. A evasão no ensino superior brasileiro. *Cadernos de pesquisa*, SciELO Brasil, v. 37, p. 641–659, 2007.

FORBES, C. et al. *Statistical distributions*. [S.l.]: John Wiley & Sons, 2011.

HOSMER, D. W.; LEMESHOW, S. *Applied logistic regression*. [S.l.]: Wiley New York, 2000.

HOX, J.; MOERBEEK, M.; SCHOOT, R. Van de. *Multilevel analysis: Techniques and applications*. [S.l.]: Routledge, 2017.

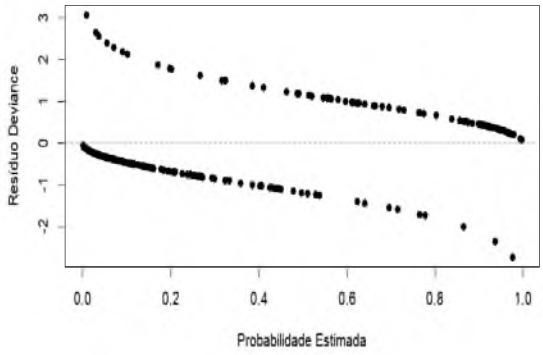
SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. d. A. Fatores associados à evasão no ensino superior brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de ciência, matemática e computação e de engenharia, produção e construção em instituições públicas e privadas. *Estudos Econômicos (São Paulo)*, SciELO Brasil, v. 49, n. 2, p. 337–373, 2019.

SEMESP. *Evasão- Dados Brasil - 15^{o}nbsp; mapa do ensino superior*. 2025. Acesso em: 12 Mai. 2025. Disponível em: ⟨https://www.semesp.org.br/mapa/edicao-15/brasil/⟩.

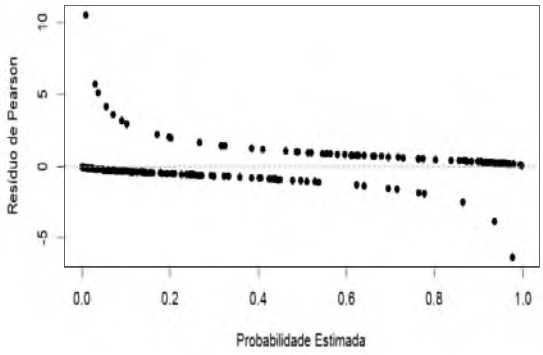
SILVA, F. C. da; CABRAL, T. L. de O.; PACHECO, A. S. V. Evasão ou permanência? modelos preditivos para a gestão do ensino superior. *Education Policy Analysis Archives*, v. 28, p. 149–149, 2020.

Apêndice

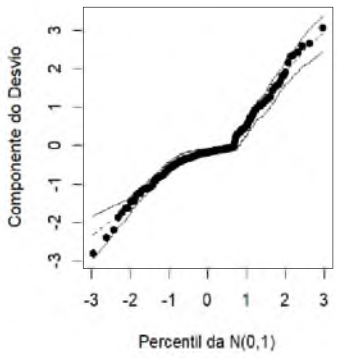
A Diagnóstico do Modelo com Taxa de Reprovação para o curso de Engenharia de Computação



(a) Resíduos *Deviance*



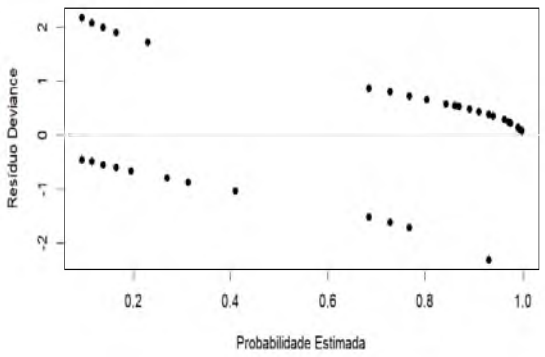
(b) Resíduos de *Pearson*



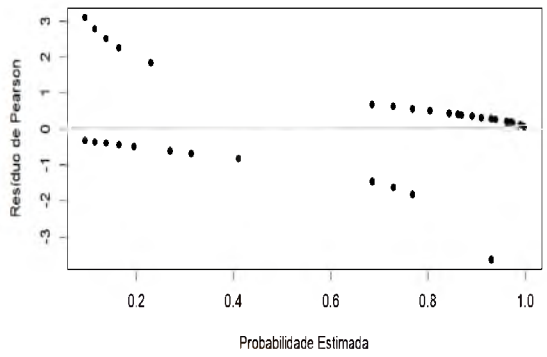
(c) Gráfico de envelope para os resíduos

Figura 11: Gráficos de análise para os resíduos do Modelo de Engenharia de Computação com a Taxa de Reprovação

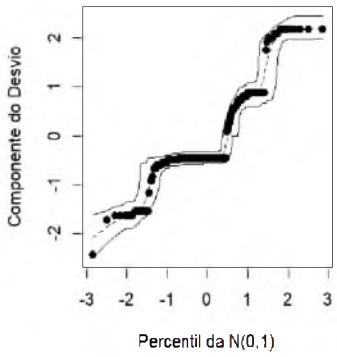
B Diagnóstico do Modelo sem IRA para o curso de Engenharia Mecatrônica



(a) Resíduos *Deviance*



(b) Resíduos de *Pearson*



(c) Gráfico de envelope para os resíduos

Figura 12: Gráficos de análise para os resíduos do Modelo de Engenharia de Mecatrônica sem IRA