Universidade de Brasília
Departamento de Estatística

Estudo da Evasão Acadêmica nos Cursos de Bacharelado do Instituto de Ciências Exatas da Universidade de Brasília: uma aplicação de modelos de Regressão Logística

Júlia Garcia Ribeiro

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

Júlia Garcia Ribeiro

Estudo da Evasão Acadêmica nos Cursos de Bacharelado do Instituto de Ciências Exatas da Universidade de Brasília: uma aplicação de modelos de Regressão Logística

Orientadora: Profa. Maria Teresa Leão Costa

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

Dedicatória

À minha mãe, que mesmo em meio às dificuldades, sempre me apoiou e fez tudo para que eu tivesse as condições de estudar. Sempre me garantiu o essencial no dia a dia e me deu todo o suporte que eu precisava. Com ela aprendi sobre força, generosidade e o valor de persistir, mesmo quando tudo parece difícil.

Ao meu pai, que sempre fez questão de garantir que nada me faltasse dizendo: "Se dedique aos estudos, do resto eu cuido", e cumpriu essa promessa em cada detalhe, me dando segurança para focar no estudo e no desenvolvimento da minha carreira.

Sou muito grata a vocês por todo o apoio. Se hoje chego até aqui, finalizando minha graduação na Universidade de Brasília, é porque tive vocês ao meu lado em cada passo.

Agradecimentos

Realizar minha graduação na Universidade de Brasília sempre foi um sonho, e vivenciar este ambiente de conhecimento e transformação foi uma das etapas mais marcantes da minha trajetória. Esses anos foram repletos de desafios, mas também de crescimento, e não teria chegado até aqui sem o apoio de pessoas essenciais que caminharam comigo.

Ao meu noivo, Felipe, por ser meu porto seguro e por acreditar em mim mesmo nos momentos mais difíceis. Aos meus pais, pela base incondicional e pelo incentivo constante. À minha avó, por todo o cuidado e amor que sempre teve comigo. À minha irmã, que celebrou comigo minha entrada na UnB com um orgulho que me motivava a seguir em frente.

Aos meus amigos Gabriel, Patrícia, Enzo, Isadora, Giovana e tantos outros que, mesmo distantes, estiveram presentes de alguma forma, obrigada por tornarem essa jornada mais leve e especial.

À minha orientadora, Professora Maria Teresa, que conheci desde o primeiro semestre na disciplina de Estatística Exploratória e por quem nutro profunda admiração. Agradeço pela paciência, parceria e por compartilhar seu conhecimento, fundamental para a construção deste trabalho.

Meus agradecimentos se estendem a todos os professores, colegas e profissionais que contribuíram direta ou indiretamente para esta conquista, este trabalho não seria possível sem cada um.

Vocês fazem parte dessa história.

Resumo

Este trabalho teve como objetivo investigar os fatores associados à evasão nos cursos de Estatística, Matemática e Ciência da Computação da Universidade de Brasília, considerando estudantes ingressantes entre 2011 e 2022. Além da modelagem estatística, o estudo traçou o perfil sociodemográfico e acadêmico dos estudantes de cada curso, com destaque para diferenças relacionadas ao gênero, tipo de escola de ensino médio, uso de cotas, idade de ingresso, histórico de reprovações e trajetória no curso.

O curso de Matemática apresentou a maior taxa de evasão, concentrada especialmente entre estudantes com desempenho acadêmico mais baixo e maior histórico de reprovações. Em Estatística, o tempo de permanência e o contexto da pandemia mostraram-se fatores relevantes. Já o curso de Ciência da Computação se destacou por possuir a maior disparidade de gênero, com predominância expressiva de estudantes do gênero masculino. O modelo logístico ajustado para esse curso apontou como principais variáveis associadas à evasão o número de semestres cursados, a ocorrência de menções SR e o Índice de Rendimento Acadêmico (IRA), indicando que estudantes com menor envolvimento acadêmico estão mais propensos à evasão.

Os resultados evidenciam padrões distintos de evasão entre os cursos, o que reforça a importância de políticas de permanência estudantil direcionadas às especificidades de cada contexto.

Palavras-chave: evasão universitária; regressão logística; pandemia; perfil discente; ensino superior.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

O Ministério da Educação (MEC) tem promovido a ampliação e democratização do acesso ao ensino superior no Brasil. Por meio de programas como o Programa Universidade para Todos (PROUNI), o Programa de Financiamento Estudantil (FIES) e políticas afirmativas, além da criação de Institutos Federais (IFs) e Faculdades Tecnológicas (FATECs), buscou-se expandir e interiorizar o ensino público, permitindo que estudantes de diversas classes sociais tivessem maior acesso às universidades (OLIVEIRA, 2019). Contudo, uma questão que se coloca é a permanência no curso e sua conclusão. Pois apesar dessas iniciativas, o índice de evasão no ensino superior brasileiro permanece alarmante, atingindo 57,2% nas redes pública e privada, de acordo com o Mapa do Ensino Superior no Brasil 2024, citado pelo Correio Braziliense (2024).

A evasão acadêmica, definida pelo MEC (1997) como a saída definitiva do estudante de seu curso antes da conclusão, traz implicações significativas. Para os estudantes, representa perdas pessoais e financeiras devido ao desperdício de recursos investidos sem o retorno esperado (Felicetti e Fossatti, 2014). Para as instituições, afeta a gestão acadêmica e financeira e compromete os indicadores de qualidade, tornando-se um dos principais desafios para o ensino superior no Brasil.

A evasão no ensino superior brasileiro apresenta características distintas entre as áreas de estudo, com os cursos de ciências exatas destacando-se pelos elevados índices de desistência. No estudo realizado por Garcia e Gomes (2022), os autores mostram que os cursos de ciências exatas apresentaram uma taxa de evasão de 59%, superior à média nacional a presentada no Mapa do Ensino Superior no Brasil de 2024. Esses cursos englobam áreas como engenharias, matemática, computação e estatística, que demandam um forte domínio de conceitos teóricos e habilidades quantitativas. Garcia e Gomes (2022) apontam que a principal causa dessa evasão é a dificuldade acadêmica, especialmente quando os estudantes reprovam nas disciplinas bases para o curso nos primeiros semestres. Neste contexto, este estudo se propõe a avaliar a influência dessas dificuldades no processo de evasão, considerando especialmente o impacto das reprovações iniciais.

O ano de 2020 foi marcado pela crise global provocada pela pandemia da COVID-19, que teve um impacto direto na educação. Em resposta à emergência sanitária, foi implementada uma oferta de ensino remoto, utilizando meios tecnológicos, mas sem o tempo necessário para um planejamento adequado dessa modalidade (Nunes, 2021). Isso resultou em diversos problemas, como aponta Gusso et al. (2020): a) a falta de suporte psicológico para os professores; b) a baixa qualidade do ensino, decorrente da falta de

planejamento estruturado devido ao tempo hábil insuficiente; c) a sobrecarga de trabalho para os docentes; d) o descontentamento dos estudantes; e) e a dificuldade de acesso dos alunos às tecnologias necessárias. A UNESCO (2020) descreve esse período como "a maior interrupção da aprendizagem da história", o que evidencia a magnitude do impacto. Além disso, a transição para o ensino remoto destacou as desigualdades educacionais no Brasil, já que muitos estudantes não tinham acesso ao material ou às condições adequadas para acompanhar as aulas a distância (Nunes, 2021). Nesse contexto, é razoável inferir que essas dificuldades contribuíram para o aumento nos índices de evasão durante o período, aspecto que será explorado neste estudo.

O presente estudo tem como objetivo identificar e analisar os fatores que contribuem para a evasão de estudantes do Instituto de Ciências Exatas (IE) da Universidade de Brasília. A análise abrange os três cursos de bacharelado oferecidos pelo instituto: Matemática, Estatística e Ciência da Computação.

2 Objetivos

Objetivo Geral

O presente estudo tem como principal objetivo a investigação sobre a evasão nos cursos de bacharelado em Estatística, Ciência da Computação e Matemática do Instituto de Ciências Exatas na Universidade de Brasília, bem como descobrir e quantificar os principais fatores contribuintes para este fenômeno.

Objetivos Específicos

- Traçar o perfil dos alunos dos cursos de bacharelado em Estatística, Ciência da Computação e Matemática;
- Verificar o impacto da COVID-19 na evasão destes grupos;
- Verificar se há associação entre a evasão e fatores sociodemográficos e acadêmicos dos estudantes;
- Averiguar se há diferenças expressivas na evasão entre os cursos de bacharelado de Estatística, Ciência da Computação e Matemática.

3 Regressão Logística

3.1 O Modelo de Regressão Logística Múltipla

No modelo de regressão logística binária, a variável resposta possui dois resultados possíveis: "sucesso"ou "insucesso", que são codificados como 0 e 1, com probabilidades de ocorrência π e $1-\pi$. A variável resposta Y tem como distribuição de probabilidade Bernoulli com valor esperado $E(Y) = \pi$, com $0 \leq \pi \leq 1$, sendo

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)}. \tag{3.1.1}$$

Sejam $X_1, \ldots, X_p$ as p variáveis explicativas e $\beta_0, \ldots, \beta_p$ os parâmetros do modelo. O objetivo é estimar a probabilidade de sucesso $\pi(X)$ a partir de uma função das variáveis explicativas, isto é:

$$\pi(\mathbf{X}) = P(Y = 1 | X_1, ..., X_p). \tag{3.1.2}$$

As variáveis explicativas podem ser quantitativas ou qualitativas. As covariáveis qualitativas serão transformadas em variáveis *dummy*, nas quais serão atribuídos valores 0 e 1, indicando ausência ou presença do efeito categórico. É importante ressaltar que para l níveis da variável independente, haverá l-1 variáveis *dummy*.

Além disso, para modelar a relação entre as variáveis explicativas e a probabilidade do evento de interesse, a regressão logística utiliza a transformação *logito*. O *logito* é definido como:

$$\text{logit}(\pi_i) = \log\left(\frac{\pi_i}{1-\pi_i}\right). \tag{3.1.3}$$

Essa transformação expressa a razão de chances (*odds*) do evento como uma função linear das variáveis preditoras:

$$\log\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p. \tag{3.1.4}$$

Dessa forma, a transformação *logito* garante que os valores previstos para π_i permaneçam em $(0, 1)$. Ademais, os coeficientes β_i podem ser interpretados em termos da mudança na razão de chances associada a uma variação unitária na variável X_i.

3.2 Estimação dos Parâmetros

A distribuição de probabilidade de cada observação Y_i é dada por

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \quad Y_i = 0, 1; \quad i = 1, \ldots, n. \tag{3.2.1}$$

Considerando que as observações Y_i são independentes, segue que a função de probabilidade conjunta é

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi^{Y_i}(1-\pi_i)^{1-Y_i}. \tag{3.2.2}$$

A partir do logaritmo de função de probabilidade conjunta, é possível calcular o log da função de verossimilhança:

$$\begin{aligned} \ln(g) &= \ln \prod_{i=1}^{n} \pi^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n} Y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} \ln(1-\pi_i). \end{aligned} \tag{3.2.3}$$

Dado que $1 - \pi_i = [1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p)]^{-1}$, segue que a função de log-verossimilhança é definida como

$$lnL(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_i) - \sum_{i=1}^{n} ln[1 + (\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_i)]. \tag{3.2.4}$$

Os parâmetros do modelo são estimados a partir dos valores que maximizam a função de verossimilhança, derivando a expressão 3.2.4 em relação a cada parâmetro do modelo e solucionando o sistema de equações resultante. Entretanto, como não há uma fórmula fechada para os valores de β que maximizam a função de verossimilhança, é necessário utilizar o método de Newton-Raphson para encontrar as estimativas.

Uma vez obtidas as estimativas de $\beta_0, \ldots, \beta_p$, substituem-se os valores em 3.1.1 para encontrar a função de resposta ajustada:

$$\hat{\pi} = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p)}. \tag{3.2.5}$$

Interpretação dos parâmetros

Para a interpretação dos parâmetros nos modelos de regressão logística, considera-se a transformação *logito*, representada pela equação 3.1.3. Essa transformação expressa a *odds*, que quantifica a razão entre as probabilidades de sucesso (π) e insucesso ($1 - \pi$):

$$odds = \frac{\pi}{1 - \pi}. \tag{3.2.6}$$

Se *odds* >1, a probabilidade de sucesso é maior do que a de fracasso, e o contrário se *odds* <1. O parâmetro β_j do preditor linear refere-se ao efeito do acréscimo de uma unidade em X_j sobre a *ln(odds)*, mantendo as variáveis explicativas constantes. Logo, considere x_j e x_j+1 dois valores distintos de uma variável explicativa X_j, com chances de sucesso *odds1* e *odds2*, respectivamente. Tem-se que

$$ln(odds1) - ln(odds2) = ln\frac{odds1}{odds2} = \beta_j. \tag{3.2.7}$$

A partir do exponencial da equação, obtém-se a razão de chances, ou *odds ratio*(θ), dada por

$$\theta = \frac{odds2}{odds1} = e^{\beta_j}. \tag{3.2.8}$$

A *odds ratio* representa a razão da chance de sucesso de um grupo em relação a outro. Dessa forma, se $\theta > 1$ a chance de sucesso dos indivíduos com X_j=x_j+1 é maior do que a dos indivíduos com X_j=x_j, e o contrário se $\theta < 1$. Considerando o caso em que X_j é uma das l níveis, então x_j=0 e $x_j + 1 = 1$ e, portanto, e_j^{β} quantifica o efeito multiplicativo na *odds* dada a presença do efeito categórico em relação à categoria de referência.

3.3 Inferência no Modelo de Regressão Múltipla

Quando o modelo inicial é ajustado, é necessário verificar se a relação entre a variável resposta e as variáveis explicativas é significativa. Os testes de significância mais comuns em regressão logística são o teste da Razão de Verossimilhança e o teste de Wald.

3.3.1 Teste da Razão de Verossimilhança

O teste da Razão de Verossimilhança avalia se existe ausência de regressão no modelo, ou seja, testa a significância dos p coeficientes do modelo. As hipóteses são:

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

A estatística de teste é dada por

$$G^2 = ln\left[\frac{L(R)}{L(F)}\right] = -2[lnL(R) - lnL(F)], \tag{3.3.1}$$

em que $L(R)$ e $L(F)$ são os valores da função de verossimilhança para os modelos reduzido e completo, respectivamente. O modelo reduzido é obtido sob H_0. A estatística de teste $G^2 \sim \chi^2_{p-q}$ sob H_0, sendo p-q a diferença entre o número de coeficientes dos dois modelos.

O teste da Razão de Verossimilhança também pode ser usado para testar individualmente se algum parâmetro do modelo é nulo ou testar subconjuntos de parâmetros. Nesse caso, as hipóteses são:

$$\begin{cases} H_0 : \beta_q = \beta_{q+1} = \cdots = \beta_{p-1} = 0, \\ H_1 : \beta_j \neq 0 \text{ para algum } j = q, \ldots, p-1. \end{cases}$$

O modelo é ordenado de modo que os últimos p-q coeficientes sejam aqueles a serem testados.

3.3.2 Teste de Wald

O teste de Wald avalia individualmente a significância de cada parâmetro do modelo e possui as seguintes hipóteses:

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

A estatística de teste é dada por

$$z^* = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \qquad \text{ou} \qquad z^* = \frac{\hat{\beta}_k^2}{s\{\hat{\beta}_k\}^2}. \tag{3.3.2}$$

Para n grande, $z^* \sim N(0,1)$ sob H_0. Alternativamente, $(z^*)^2 \sim \chi_1^2$ sob H_0.

3.3.3 Intervalo de Confiança para os parâmetros

O intervalo de confiança $1-\alpha$ de um parâmetro β_k é obtido por

$$\beta_k \in (\hat{\beta}_k \pm z_{1-\frac{\alpha}{2}} s\{\hat{\beta}_k\}), \tag{3.3.3}$$

em que $z_{1-\alpha/2} \sim N(0,1)$ e $s\{\hat{\beta}_k\}$ corresponde à estimativa do erro padrão de $\hat{\beta}_k$. De modo similar, o intervalo de confiança de nível $1-\alpha$ para a razão de chances θ_k é dado por

$$\theta_k \in (exp(\hat{\beta}_k \pm z_{1-\frac{\alpha}{2}} s\{\hat{\beta}_k\})). \tag{3.3.4}$$

3.4 Seleção do Modelo

A seleção de variáveis explicativas é uma etapa fundamental na construção de um modelo de regressão. Esta etapa tem como objetivo reduzir o número de covariáveis sem comprometer a qualidade do ajuste, de modo a obter um modelo mais parcimonioso. A seleção pode ser feita a partir de critérios, como *AIC* e *BIC* ou métodos automáticos.

3.4.1 Critérios de Seleção

O Critério de Informação de Akaike *AIC* e o Critério de Informação Bayesiano *BIC* medem a proximidade entre os valores estimados pelo modelo e os valores observados. Valores pequenos indicam melhor ajuste do modelo. As medidas são calculadas a partir das expressões:

$$AIC = -2lnL(\beta) + 2p, \tag{3.4.1}$$

$$BIC = -2lnL(\beta) + pln(n), \tag{3.4.2}$$

em que L(β) é a log-verossimilhança definida em 3.2.4 e p é o número de parâmetros do modelo.

3.4.2 Métodos Automáticos

Quando há um número elevado de variáveis explicativas potenciais, a comparação manual de todos os modelos possíveis se torna inviável. Nesses cenários, recorre-se a métodos computacionais iterativos para identificar os modelos mais apropriados. Entre os procedimentos automáticos mais comuns em análise de regressão, destaca-se o método *Stepwise* e suas variações:

Forward: O método inicia com um modelo vazio e adiciona, iterativamente, as variáveis preditoras que mais melhoram o modelo de acordo com o ganho na função de log-verossimilhança, até que nenhuma variável adicional melhore significativamente o ajuste.

Backward: O procedimento começa com um modelo completo que inclui todas as variáveis preditoras e remove, iterativamente, a variável menos significativa, até que todas as variáveis restantes sejam significativas.

Stepwise: Este método combina os passos de inclusão e exclusão. Ele inicia com um modelo vazio ou completo e, a cada etapa, as variáveis são incluídas ou não de acordo com o ganho na função de log-verossimilhança. O procedimento termina quando não há mais variáveis para serem adicionadas ou retiradas do modelo.

3.5 Avaliação do Modelo

Após a seleção do modelo, é fundamental avaliar tanto a qualidade do seu ajuste aos dados quanto sua capacidade preditiva. Dessa forma, diversos procedimentos devem ser aplicados a fim de assegurar a adequação e a confiabilidade do modelo.

3.5.1 Adequabilidade do Ajustamento

Para medir a qualidade do ajuste do modelo são feitos testes de adequabilidade. Caso o modelo não esteja bem ajustado, este deve ser corrigido ou descartado.

Teste χ^2 de Pearson

O Teste χ^2 de Pearson mede o desvio entre o número observado de sucessos e o número esperado ou ajustado de sucessos. Seja $\pi(x)$ definido em 3.2.5, as hipóteses de teste são:

$$\begin{cases} H_0 : \pi_i = \pi(x_i), \\ H_1 : \pi_i \neq \pi(x_i). \end{cases}$$

A estatística de teste é dada por:

$$\chi^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(o_{jk} - e_{jk})^2}{e_{jk}}, \tag{3.5.1}$$

em que o_{jk} é a frequência observada e e_{jk} é a frequência esperada. Sob H_0, tem-se que $\chi^2 \sim \chi^2_{c-p}$, sendo $c - p$ a diferença entre o número de conjuntos de valores distintos de variáveis e o total de parâmetros no modelo, respectivamente. Para a utilização do teste, as frequências esperadas devem ser maiores ou iguais a 5, e nunca inferiores a 1.

Teste de Hosmer e Lemeshow

Hosmer e Lemeshow (2000) propuseram um teste alternativo de adequabilidade de ajuste para modelos logísticos. Neste teste a variável resposta é agrupada de acordo com os valores estimados de probabilidade. A hipóteses são:

$$\begin{cases} H_0 : \pi_i = \pi(x_i), \\ H_1 : \pi_i \neq \pi(x_i). \end{cases}$$

A amostra com n observações é ordenada a partir da probabilidade estimada de sucesso, e depois dividida em g grupos. Os autores recomendam $g = 10$ grupos. A divisão pode ser feita de duas formas: a partir dos percentis das probabilidades estimadas ou com base em valores fixados das probabilidades.

No primeiro método, cada um dos g grupos deve conter $n' = n/g$ valores de

probabilidades preditas, de modo que o primeiro grupo possua os menores valores, e o g-*ésimo* grupo possua os maiores valores. No segundo método, são definidos pontos de corte nos valores $k/g, k = 1, ..., g-1$, e os grupos contêm as observações com probabilidades estimadas entre os pontos de corte próximos.

A estatística de teste é calculada a partir do χ^2 de Pearson:

$$\hat{C} = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(o_{jk} - e_{jk})^2}{e_{jk}}. \tag{3.5.2}$$

Sob H_0 tem-se que $\hat{C} \sim \chi^2_{g-2}$, sendo g o número de grupos.

3.5.2 Análise de Resíduos

Os resíduos representam a diferença entre os valores observados e os estimados pelo modelo. Quando o modelo se ajusta bem aos dados, esses resíduos tendem a ser pequenos. Assim, a análise diagnóstica dos resíduos é uma etapa essencial para avaliar a qualidade do ajuste do modelo.

Resíduos de Pearson

Os resíduos de Pearson são a razão entre a diferença dos valores observados e preditos e a estimativa do erro padrão de Y_i. Podem ser calculados por:

$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}. \tag{3.5.3}$$

Resíduos Studentizados de Pearson

Resíduos Studentizados são uma padronização dos resíduos de Pearson, e são calculados por:

$$r_{SPi} = \frac{r_i}{\sqrt{1 - h_{ii}}}, \tag{3.5.4}$$

em que h_{ii} é a i-ésima diagonal da matriz H,

$$H = W^{1/2} X (X'WX)^{-1} X' W^{1/2}, \tag{3.5.5}$$

e W é a matriz diagonal $n \times n$dos valores $\hat{\pi}_i(1 - \hat{\pi}_i)$.

Resíduos *Deviance*

Resíduos *Deviance* medem a distância das funções de máxima verossimilhança observada e estimada. O resíduo *deviance* para uma observação i é dado por:

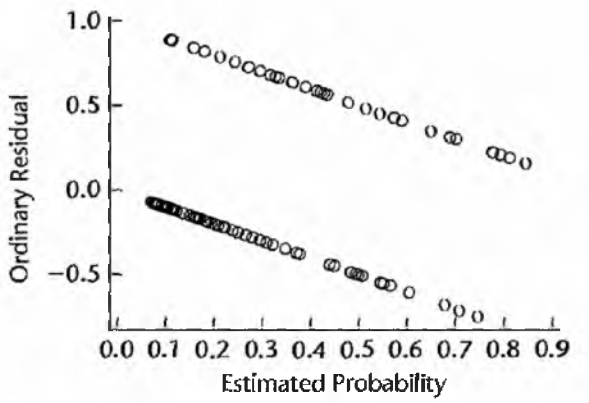
$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\sum_{i=1}^{n}[Y_i log(\hat{\pi}_i) + (1 - Y_i)log(1 - \hat{\pi}_i))]}. \tag{3.5.6}$$

A função $sign()$ indica o sinal do resultado de $Y_i - \hat{\pi}_i$, isto é, assume valor 1 se $Y_i - \hat{\pi}_i > 0$ e valor -1 se $Y_i - \hat{\pi}_i < 0$. A soma do quadrado dos resíduos *deviance* é equivalente a *Deviance* do modelo:

$$DEV(X_0, ..., X_{p-1}) = \sum_{i=1}^{n} dev_i^2 = -2\sum_{i=1}^{n}[Y_i \log(\hat{\pi}_i) + (1 - Y_i)\log(1 - \hat{\pi}_i)]\,. \tag{3.5.7}$$

Os gráficos podem representar os resíduos após serem calculados, facilitando a avaliação de potenciais desvios dos pressupostos do modelo. Considerando que, se o modelo de regressão logística estiver corretamente especificado, então $E\{Y_i\} = \pi_i$ e, assintoticamente, espera-se que $E\{e_i\} = 0$. Assim, se o modelo estiver correto, espera-se observar um gráfico de resíduos com a seguinte configuração:

Figura 1: Exemplo de gráfico de resíduos



Fonte: Neter et al., *Applied Linear Statistical Models*, p. 594.

3.5.3 Matriz de confusão e desempenho

Ao avaliar o ajuste de um modelo, é fundamental verificar seu poder preditivo. A predição de uma observação é dada pelo resultado do Modelo 3.2.5. De acordo com

a probabilidade de sucesso $\hat{\pi}_i$, a observação i é $\hat{y}_i$= 1 (sucesso) quando a probabilidade estimada $\hat{\pi}_i > \pi_0$ e $\hat{y}_i$= 0 (insucesso) quando $\hat{\pi}_i \leq \pi_0$ para determinada probabilidade de corte π_0. Normalmente adota-se π_0=0,5, mas π_0 é obtido precisamente a partir da análise da curva de ROC.

A matriz de confusão e desempenho é uma tabela de classificação cruzada da variável resposta binária Y e o resultado preditivo de $\hat{Y}$. A matriz classifica os valores em quatro classes:

- **Verdadeiro positivo (VP):** Y=1 e $\hat{Y}$=1;
- **Falso positivo (FP):** Y=0 e $\hat{Y}$=1;
- **Verdadeiro negativo (VN):** Y=0 e $\hat{Y}$=0;
- **Falso negativo (FN):** Y=1 e $\hat{Y}$=1.

Dessa forma, a matriz de confusão e desempenho mostra as frequências de ocorrência de cada classe, como demonstrado abaixo:

Tabela 1: Matriz de Confusão

		Observado	
		$Y = 1$	$Y = 0$
Previsto	$\hat{Y} = 1$	**VP**	**FP**
	$\hat{Y} = 0$	**FN**	**VN**

A partir da matriz é possível medir a acurácia do modelo, ou seja, o percentual de acerto das previsões:

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{3.5.8}$$

A sensibilidade $= P(\hat{Y} = 0 \mid Y = 1)$ mede o poder do modelo em prever os sucessos, e pode ser calculada por:

$$\text{Sensibilidade} = \frac{VP}{VP + FN}. \tag{3.5.9}$$

A especificidade $= P(\hat{Y} = 0 \mid Y = 0)$ mede o poder do modelo em prever os fracassos, e pode ser calculada por:

$$\text{Especificidade} = \frac{VN}{VN + FP}. \tag{3.5.10}$$

A métrica *F1-score* é definida como a média harmônica entre a Precisão e a Sensibilidade (também conhecida como *Recall*). Trata-se de uma medida particularmente útil em contextos de dados desbalanceados, nos quais é necessário equilibrar o impacto dos erros do tipo falso positivo e falso negativo. O *F1-score* busca oferecer uma visão mais equilibrada do desempenho do modelo, considerando tanto a capacidade de identificar corretamente os positivos quanto a confiabilidade dessas previsões.

A fórmula do *F1-score* é dada por:

$$\textit{F1-score} = \frac{2 \times \text{Precisão} \times \text{Sensibilidade}}{\text{Precisão} + \text{Sensibilidade}}. \tag{3.5.11}$$

A sensibilidade já foi apresentada na equação 3.5.9. A seguir, define-se a precisão:

$$\text{Precisão} = \frac{VP}{VP + FP}. \tag{3.5.12}$$

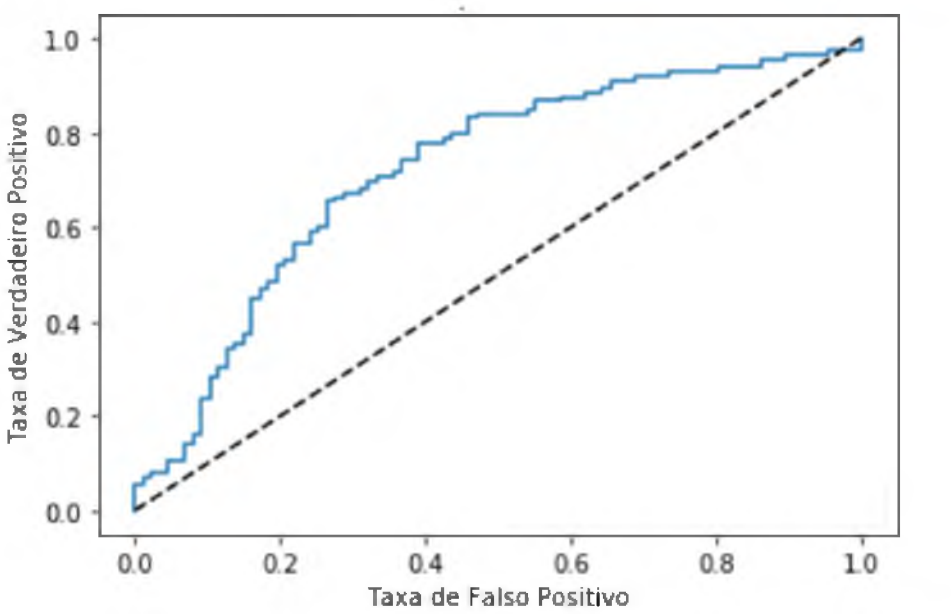
3.5.4 Curva ROC

A curva ROC ou curva de Característica de Operação do Receptor (*Receiver Operating Characteristic*) é um gráfico da sensibilidade *vs.* 1 - especificidade de todas as possíveis probabilidades de corte π_0.

A área abaixo da curva, denominada AUC, fornece uma medida do poder preditivo do modelo. Quanto maior for a AUC, melhor é a habilidade do modelo em discriminar corretamente a variável resposta em sucesso ou fracasso. Segundo Hosmer e Lemeshow (2000), a AUC pode ser interpretada como:

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

Figura 2: Exemplo da curva ROC



A partir da análise da Figura 2, o ponto de corte ideal pode ser escolhido como o valor que maximiza a sensibilidade e a especificidade, geralmente próximo ao canto superior esquerdo da curva ROC.

4 Metodologia

4.1 Conjunto de dados

Neste estudo utilizou-se a base de dados fornecida pela Secretaria de Tecnologia da Informação (STI) a pedido do Instituto de Ciências Exatas ao Decanato de Graduação. Os dados foram extraídos via Sistema de Informações Acadêmicas de Graduação (SIGRA) e Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA). A base abrange o período de 2011 a 2023/2. São ao todo três bases de dados, referentes a cada curso de bacharelado do Instituto de Ciências Exatas e possuem dados acadêmicos e sociodemográficos dos alunos matriculados nos cursos durante o período.

As bases 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.

Cada registro da base de dados original corresponde a uma disciplina cursada por um estudante. Durante o processo de tratamento dos dados, foi identificado que alguns estudantes apareciam mais de uma vez na base com diferentes identificadores, mas com as demais informações repetidas. Por essa razão, foi necessário realizar uma etapa de identificação e correção desses casos, garantindo que cada estudante fosse considerado apenas uma vez na nova base. A partir disso, foram construídas duas tabelas: uma tabela "raw", que manteve a estrutura original da base, e uma tabela derivada, criada a partir da primeira. Todas as novas variáveis utilizadas nas análises foram geradas com base na tabela "raw". Na tabela derivada, cada linha passou a representar um estudante.

4.2 Criação de Variáveis

A partir dos dados disponíveis na base, foram criadas as novas variáveis de interesse para o estudo. Além disso, as variáveis já presentes foram agrupadas ou padronizadas, sempre que necessário, com o objetivo de garantir maior consistência e adequação às análises propostas.

Evasão

Indica se o estudante evadiu ou não do curso. A variável foi derivada da coluna original "Forma de saída do curso" e categorizada em dois valores: "Sim" e "Não". A categoria "Sim" foi atribuída aos estudantes que interromperam a graduação antes da conclusão. Já a categoria "Não" inclui tanto os estudantes que concluíram integralmente o curso quanto aqueles que, até o momento da extração dos dados, permaneciam regularmente matriculados. A variável foi classificada da seguinte forma:

Tabela 2: Classificação da evasão a partir da forma de saída do curso

Forma de saída	**Evasão**
Ativo	Não
Formatura	
Integralização de discente	
Novo vestibular	Sim
Mudança de curso	
Desligamento - Abandono	
Desligamento - Rendimento	
Desligamento - Voluntário	
Desligamento - Intercâmbio	
Efetivação de novo cadastro	
Reprovado 3 vezes	
Solicitação espontânea	
Falecimento	

Para esta classificação foi utilizado o conceito de evasão acadêmica definida pelo MEC (1997) como a saída definitiva do estudante de seu curso antes da conclusão.

Pandemia

A variável Pandemia foi criada para categorizar os estudantes com base no período de ingresso e conclusão de seus cursos em relação ao contexto da pandemia de COVID-19. Essa variável divide os estudantes em três grupos distintos:

- **Pré-Pandemia:** Estudantes que ingressaram e concluíram ou evadiram do curso antes do início da pandemia.
- **Transição:** Estudantes que ingressaram antes da pandemia, mas cuja trajetória acadèmica (conclusão ou evasão) ocorreu durante o período pandèmico.
- **Pandemia:** Inclui estudantes que ingressaram durante o período da pandemia e que já tenham concluído ou evadido, também contém aqueles que ingressaram antes da pandemia e estavam ativos no curso durante o período pandèmico. Em resumo, abrange todos os estudantes que, de alguma forma, tiveram sua trajetória acadèmica na universidade afetada pela pandemia.

Para a seção descritiva do trabalho foram considerados também os estudantes Pós-Pandemia, ou seja, que ingressaram na universidade após o período pandèmico, a partir de 2023/1. Embora a universidade tenha voltado suas atividades presenciais em 2022/2, era um período de transição e ainda com restrições (como algumas flexibilidades nas regras acadèmicas como trancamentos, etc), por isso o semestre de 2023/1 foi considerado como o período pós-pandemia. Porém, a categoria não foi considerada na modelagem por conter poucos estudantes.

Idade de Ingresso

Foi calculada a partir da variável "Data de nascimento"e "Período de ingresso". É expressa em anos completos. É calculada pela diferença em anos entre a data de nascimento e o ano de período de ingresso no curso.

Semestres Cursados

Representa o número de semestres efetivamente cursados pelo estudante até sua saída do curso, para estudantes ainda ativos, a saída do curso foi considerada como o 2º semestre de 2023, momento da extração dos dados. Inicialmente, essa variável foi calculada com base na diferença entre o "Período de ingresso no curso" e o "Período de saída do curso". No entanto, como alguns estudantes apresentaram semestres sem matrícula em disciplinas na grade curricular, foi realizada uma verificação adicional para

que contabilizasse nesta variável apenas o número de semestres onde foi cursada ao menos uma disciplina. É uma variável quantitativa discreta.

Interrupções

Refere-se ao número de semestres em que o estudante não cursou nenhuma disciplina ao longo de sua trajetória no curso. Essa variável foi derivada durante o processo de construção da variável "Semestres Cursados", por meio da identificação e contagem dos períodos em que não houve registro de matrícula ativa em disciplinas, ou seja, quando os estudantes não cursaram nenhuma disciplina no semestre.

Trancamentos

Contém "Sim" ou "Não", sendo verificado a partir da variável "Menção na Disciplina". Toda vez em que eram atribuídas as menções TR ou TJ (Trancamento ou Trancamento Justificado) em uma disciplina. É uma variável nominal binária.

Menções SR

Número de menções SR (Sem Rendimento) que o estudante obteve durante o seu período de graduação. Foi contabilizada a partir da variável "Menção na Disciplina".

Começou outra Graduação

Para alguns estudantes, a base de dados apresenta discrepâncias entre o período de ingresso no curso de Estatística e o período em que cursaram disciplinas do primeiro semestre. Nesses casos, foram identificados estudantes que já haviam iniciado outro curso na universidade antes de ingressarem em Estatística.

Além disso, foi realizada uma segunda verificação com base na variável "forma de ingresso na UnB", considerando como indicativos de uma graduação anterior as seguintes categorias: Dupla Diplomação, Dupla Habilitação, Duplo Curso, Mudança de Habilitação e Portador de Diploma de Curso Superior.

A variável recebeu o nome "Começou outra graduação" porque, com os dados disponíveis, não é possível confirmar se esses estudantes concluíram os cursos anteriores. Ainda assim, essa variável foi incluída por se tratar de uma informação potencialmente relevante para o modelo final.

Taxa de Reprovação

A taxa de reprovação indica a proporção entre o total de créditos com reprovação e o total de créditos cursados pelo aluno ao longo de sua permanência no curso. A taxa varia de 0 a 1. O cálculo da taxa é feito da seguinte forma:

$$\text{Taxa de reprovação} = \frac{\text{Total de créditos com reprovação}}{\text{Total de créditos cursados}} \tag{4.2.1}$$

Currículo

Foi realizada uma análise da grade curricular de cada curso com o intuito de identificar mudanças significativas ao longo do tempo. A partir dessa análise, foi feita uma classificação binária dos currículos como "Antigo" ou "Novo", utilizando como critério o semestre de ingresso do discente no respectivo curso.

Para o curso de Estatística essa mudança foi identificada a partir de 2014/1. Enquanto em Ciência da Computação foi a partir de 2015/2, dado que houveram diversas mudanças principalmente na oferta do 1º semestre do curso. Contudo, no curso de Matemática não foi observada nenhuma mudança significativa da grade ao longo dos anos, por isso foi atribuída a classificação Novo para todas as observações do curso.

Reprovou Obrigatória do 1º Semestre

Foi realizada uma análise dos currículos dos três cursos ao longo dos anos para definir as disciplinas do primeiro semestre de cada um. Após esta etapa, foi criada uma variável binária que mostra se o estudante reprovou em disciplinas obrigatórias do primeiro semestre de seu curso (quando foram atribuídas as menções MI, II e SR) ou não.

Tabela 3: Matérias obrigatórias do 1° semestre. Bacharelado do IE - UnB

Curso	**Matérias obrigatórias do 1° semestre**
Ciência da Computação (Currículo Antigo)	Cálculo 1, Computação Básica, Física 1 e Física 1 Experimental
Ciência da Computação (Currículo Novo)	Cálculo 1, Algoritmos e Programação de Computadores, Informática e Sociedade e Introdução aos Sistemas Computacionais
Estatística (Currículo Novo)	Cálculo 1, Computação em Estatística 1, Estatística Exploratória, Introdução à Ciência da Computação e Introdução à Probabilidade
Matemática (Currículo Novo)	Cálculo 1 e Introdução à Ciência da Computação

Renda por Local

O nível de renda foi definido com base na Região Administrativa (RA) de residência dos estudantes, identificada a partir do Código de Endereçamento Postal (CEP). Os CEPs foram utilizados para localizar a respectiva RA, por meio de uma consulta à API pública ViaCEP[1]. A partir dessa identificação, cada RA foi classificada em grupos de renda de acordo com a categorização proposta pela Codeplan (PDAD 2021)[2], conforme apresentado nas Tabelas 4 e 5.

Tabela 4: Grupo de renda das RAs do Distrito Federal

Grupo de renda	**Região Administrativa**
Renda Alta	Plano Piloto, Lago Sul, Lago Norte, Jardim Botânico, Sudoeste/Octogonal, Park Way, Águas Claras
Renda Média-Alta	Sobradinho, Guará, Cruzeiro, Vicente Pires, Taguatinga, Núcleo Bandeirante, Candangolândia, SIA, Arniqueira
Renda Média-Baixa	Gama, Riacho Fundo I, Samambaia, Santa Maria, Ceilândia, Riacho Fundo II, Sobradinho II
Renda Baixa	Fercal, Brazlândia, Planaltina, Recantos das Emas, Paranoá, São Sebastião, Varjão, Itapoã, Sol Nascente , Estrutural/SCIA

Fonte: PDAD 2021 - Codeplan

[1]Interface de programação que permite o acesso automatizado a dados por meio de requisições web. Link de exemplo para consultar um CEP na API ViaCEP

[2]PDAD 2021 - Codeplan

Além das RAs do Distrito Federal, foram identificados diversos alunos vindos do Entorno. Para incluí-los na análise, foi consultado o site do IBGE[3], onde foi verificado o salário médio mensal dos trabalhadores formais de cada município. Com base nesses valores, a classificação das faixas de renda também foi feita de acordo com os critérios estabelecidos no Relatório da Codeplan (PDAD 2021)[1].

Tabela 5: Grupo de renda das regiões do Entorno

Grupo de renda	**Região do Entorno**
Renda Baixa	Águas Lindas, Alexânia, Cidade Ocidental, Cocalzinho, Cristalina, Formosa, Luziânia, Novo Gama, Padre Bernardo, Planaltina, Santo Antônio do Descoberto e Valparaíso

Fonte: PDAD 2021 - IBGE

Distância de deslocamento

A partir do endereço aproximado de residência de cada estudante, obtido por meio da API pública ViaCEP, foi utilizada a API do Google Maps para calcular a distância média entre a residência e a Universidade de Brasília, campus Darcy Ribeiro, considerando o trajeto por transporte rodoviário. Todo esse processo foi realizado no software R, com o uso de funções que automatizaram tanto a consulta aos endereços quanto o cálculo das distâncias. A variável resultante é quantitativa contínua.

4.3 Métodos de Análise

Para atingir os objetivos do estudo, foi usada análise descritiva para traçar o perfil dos estudantes dos cursos de bacharelado de Estatística, Ciência da Computação e Matemática. Após esta etapa, a regressão logística será utilizada para identificar os fatores associados à evasão, uma vez que a variável reposta (se evadiu ou não) é qualitativa binária.

O tratamento da base de dados, a criação das variáveis e a análise descritiva foram feitos no software *R*, utilizando a versão 4.4.2. As análises e a construção dos modelos serão conduzidas no *SAS OnDemand for Academics*.

[3]Site do IBGE

5 Resultados

5.1 Análise Descritiva

Esta seção apresenta a análise descritiva univariada das variáveis deste estudo. São exploradas características sociodemográficas, bem como variáveis relacionadas ao percurso acadèmico dos estudantes. Essa etapa tem como objetivo oferecer uma visão geral do perfil dos estudantes dos cursos de Estatística, Ciência da Computação e Matemática da Universidade de Brasília, entre os anos de 2011 a 2023.

Tabela 6: Distribuição dos estudantes segundo informações pessoais. Bacharelados do IE - UnB, 2011-2023

	Estatística	**Ciência da Computação**	**Matemática**
Gênero			
Feminino	34,63% (457)	11,46% (183)	22,62% (114)
Masculino	65,37% (863)	88,54% (1414)	77,38% (390)
Renda			
Renda Baixa	6,74% (89)	5,39% (86)	7,14% (36)
Renda Média-Baixa	9,47% (125)	9,64% (154)	10,91% (55)
Renda Média-Alta	25,15% (332)	22,23% (355)	27,18% (137)
Renda Alta	49,55% (654)	52,31% (836)	45,04% (227)
Etnia/Cor			
Amarela	3,29% (33)	2,53% (31)	1,20% (4)
Branca	46,61% (467)	45,27% (555)	45,95% (153)
Parda	28,44% (285)	28,55% (350)	27,93% (93)
Preta	5,78% (58)	7,91% (97)	8,11% (27)

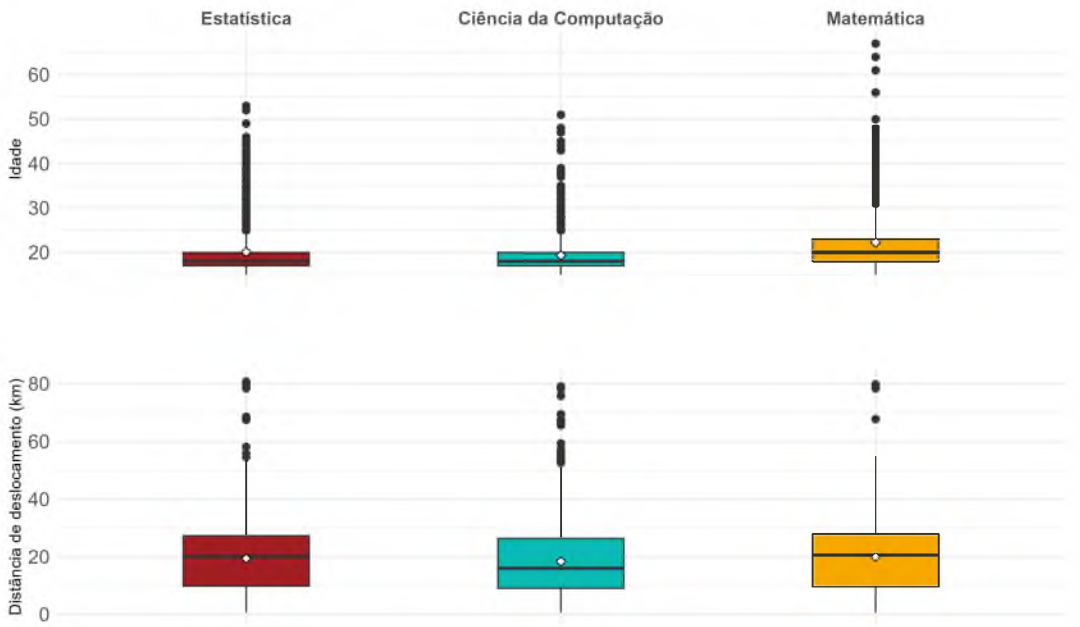
Observa-se uma predominância de estudantes do gènero masculino nos três cursos analisados. Em Ciència da Computação, essa diferença é ainda mais acentuada, com 88,54% dos alunos do gènero masculino. Em Matemática, os homens também são maioria, representando 77,38% do total. O curso de Estatística apresenta a distribuição mais equilibrada entre os cursos avaliados, com 34,63% de estudantes do gènero feminino e 65,37% do masculino.

No que diz respeito à variável renda, destaca-se a predominância de estudantes provenientes de regiões classificadas como de renda alta e média-alta. Em Ciència da Computação, mais da metade dos estudantes (52,31%) pertence ao grupo de renda

alta, enquanto em Estatística e Matemática, essa proporção é de 49,55% e 45,04%, respectivamente. Os grupos de renda média-baixa e baixa representam uma parcela menor dos estudantes nos três cursos, com destaque para Estatística, onde apenas 6,74% dos alunos são de renda baixa. Para esta análise, os estudantes com CEPs localizados fora do Distrito Federal ou da região do Entorno foram considerados com valor nulo para a variável de renda, a fim de garantir a consistência da classificação baseada no contexto socioeconômico regional.

Quanto à distribuição por etnia, a maioria dos estudantes se autodeclara branca, com percentuais próximos entre os cursos: 46,61% em Estatística, 45,27% em Ciência da Computação e 45,95% em Matemática. Em seguida, destaca-se a categoria parda, representando aproximadamente 28% dos alunos nos três cursos. A participação de estudantes autodeclarados pretos varia entre 5,78% (Estatística) e 8,11% (Matemática). A menor representatividade é observada na categoria amarela, que corresponde a menos de 4% dos estudantes em todos os cursos analisados. Para todos os cursos há também a classificação "Não Informado", em que contém o número de estudantes que não declararam sua etnia, essa categoria, portanto, foi excluída da tabela.

Figura 3: Distribuição das variáveis Idades (anos) e Distância de Deslocamento (km), por curso. Bacharelado do IE - UnB, 2011-2023



Em relação à idade, observa-se que o curso de Matemática possui uma distribuição mais dispersa, com maior presença de estudantes mais velhos, evidenciada pela maior quantidade de outliers acima dos 50 anos. Os cursos de Ciência da Computação

e Matemática apresentam distribuições mais concentradas em faixas etárias mais jovens, com a mediana de idade próxima a 20 anos em ambos os casos.

No que diz respeito à distância de deslocamento, nota-se uma maior variabilidade entre os estudantes de Estatística e Matemática, com medianas superiores a 20 km e presença de valores extremos acima de 60 km. Em Ciência da Computação, a distância média de deslocamento é um pouco menor, com menor dispersão e uma concentração mais evidente de estudantes que residem a distâncias inferiores a 20 km da Universidade.

Figura 4: Mapa do local de residência de estudantes do curso de Estatística. Bacharelado do IE - UnB, 2011-2023

Ciência da Computação Matemática

Figura 5: Mapas do local de residência de estudantes dos cursos de Ciência da Computação e Matemática. Bacharelado do IE - UnB, 2011-2023

A análise espacial dos locais de residência dos estudantes dos cursos do Instituto de Ciências Exatas da UnB revela uma forte concentração em regiões centrais do Distrito Federal. No caso do curso de Estatística, observa-se uma alta densidade de estudantes residentes no Plano Piloto, além de uma presença significativa em regiões próximas, como Guará, Águas Claras e Taguatinga.

Já no curso de Ciência da Computação, o padrão é semelhante, com uma concentração ainda mais acentuada no Plano Piloto e em áreas próximas, como Vicente Pires, Taguatinga, Guará e Ceilândia, indicando uma distribuição levemente mais dispersa do que a observada em Estatística.

Para o curso de Matemática, embora a concentração no Plano Piloto também se mantenha como o principal polo de residência, há uma presença proporcionalmente maior em regiões periféricas como Ceilândia, Samambaia e Santa Maria, refletindo uma distribuição mais descentralizada entre os estudantes do curso.

Além disso, a análise do local de moradia também mostrou que a participação de estudantes provenientes do Entorno do Distrito Federal é baixa em todos os cursos, representando aproximadamente 1,6% em Estatística, 1,8% em Ciência da Computação e 2,6% em Matemática.

Para visualizar detalhadamente as informações sobre o local de residência por curso, consulte a Tabela 32, disponível no Apêndice deste trabalho.

Tabela 7: Distribuição dos estudantes segundo informações de entrada na universidade. Bacharelados do IE - UnB, 2011-2023

	Estatística	**Ciência da Computação**	**Matemática**
Tipo de Escola no Ensino Médio			
Escola Particular	51,28% (583)	53,33% (745)	51,40% (165)
Escola Pública	48,72% (554)	46,67% (652)	48,60% (156)
Forma de Ingresso na UnB			
Vestibular	44,55% (588)	49,96% (669)	48,64% (251)
PAS	35,95% (397)	14,83% (199)	14,80% (75)
ENEM	10,26% (119)	7,24% (97)	1,94% (10)
SISU	10,67% (109)	7,89% (101)	3,49% (16)
Outras formas	18,57% (245)	20,08% (269)	31,13% (161)
Começou outra graduação			
Sim	11,67% (154)	17,85% (285)	18,99% (98)
Não	88,33% (1166)	82,15% (1312)	81,01% (418)
Utilizou cotas			
Sim	21,74% (287)	29,93% (478)	14,73% (76)
Não	78,26% (1033)	70,07% (1119)	85,27% (440)

Em relação ao tipo de escola de origem, os cursos de Estatística e Matemática possuem uma distribuição relativamente equilibrada entre estudantes provenientes de escolas públicas e particulares. Em Ciência da Computação, observa-se uma ligeira predominância de estudantes oriundos de escolas particulares.

Em relação à forma de ingresso, destaca-se que o Vestibular foi a principal via de acesso para os três cursos, seguido pelo Programa de Avaliação Seriada (PAS), especialmente em Estatística, onde essa modalidade representou cerca de 36% dos ingressantes. Observa-se também uma presença relevante de estudantes que ingressaram por outras formas de seleção[4], sobretudo em Matemática, onde essa categoria atinge cerca de 31%. Esse resultado pode estar relacionado ao elevado número de estudantes que realizam mudança de habilitação entre as modalidades de licenciatura e bacharelado dentro do próprio curso de Matemática ao longo da graduação.

Quanto à experiência prévia no ensino superior, a maioria dos estudantes nos três cursos não iniciou outra graduação antes de ingressar na UnB. Contudo, os cursos de Ciência da Computação e Matemática apresentam proporções mais altas de estudantes que já haviam iniciado outra graduação, com aproximadamente 18% e 19%, respectivamente.

No que se refere à utilização de políticas de ação afirmativa, os dados indicam que o curso de Ciência da Computação concentra a maior proporção de estudantes que ingressaram por meio de cotas (aproximadamente 30%), seguido por Estatística (cerca de 22%) e Matemática (15%). Destaca-se que, nos três cursos, a maior parte dos estudantes ingressou pela ampla concorrência.

Tabela 8: Distribuição dos estudantes cotistas segundo tipo de cota declarada. Bacharelados do IE - UnB, 2011-2023

Tipo de Cota	**Estatística**	**Ciência da Computação**	**Matemática**
Negro	29,27% (84)	7,11% (34)	59,21% (45)
Pública Alta Renda - PPI	25,44% (73)	17,36% (83)	11,84% (9)
Pública Baixa Renda - PPI	9,41% (27)	13,18% (63)	7,89% (6)
Pública Alta Renda - Não PPI	10,10% (29)	-	1,32% (1)
Pública Baixa Renda - Não PPI	6,97% (20)	9,83% (47)	7,89% (6)

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

Em Estatística, destacam-se os cotistas autodeclarados negros (29,27%) e os oriundos de escolas públicas de alta renda - PPI (25,44%). As demais categorias apresentam participação menor, mas com distribuição relativamente equilibrada.

[4]Como por convênio cultural, ingresso por portador de diploma de curso superior, etc.

No curso de Ciência da Computação, observa-se maior concentração de cotistas de escolas públicas de baixa renda, tanto PPI (13,18%) quanto não PPI (9,83%). A presença de cotistas negros é mais reduzida (7,11%) e não há registros na categoria "Pública Alta Renda - Não PPI".

Em Matemática, chama atenção a predominância de cotistas autodeclarados negros (59,21%), indicando forte adesão às ações afirmativas. As outras categorias aparecem com proporções próximas entre si, sem grandes variações.

Tabela 9: Distribuição dos estudantes segundo período de ingresso na universidade. Bacharelados do IE - UnB, 2011-2023

Período de Ingresso	**Estatística**	**Ciência da Computação**	**Matemática**
Pré-Pandemia	65,6% (765)	64,6% (914)	79,1% (400)
Transição	12,1% (141)	12,0% (169)	6,1% (31)
Pandemia	30,7% (358)	29,8% (421)	14,4% (73)
Pós-Pandemia	3,0% (35)	3,0% (42)	2,2% (11)

Em relação ao período de ingresso, observa-se que a maioria dos estudantes dos três cursos iniciou sua graduação antes da pandemia de Covid-19. Esse resultado era esperado, considerando que a base de dados abrange um período longo, de 2011 a 2023, o que resulta em um maior número de estudantes classificados como pré-pandemia. Esse percentual é elevado em Matemática, onde aproximadamente 79% dos estudantes ingressaram no período pré-pandèmico. Em Ciência da Computação e Estatística, essa proporção é de cerca de 65% em ambos os cursos. Quanto aos ingressantes durante a pandemia, eles representam aproximadamente 30% em Estatística, 30% em Ciência da Computação e 14% em Matemática. Já os estudantes classificados como pós-pandèmicos correspondem a um percentual reduzido, inferior a 3% nos três cursos.

Tabela 10: Distribuição dos estudantes segundo características acadêmicas. Bacharelados do IE - UnB, 2011-2023

	Estatística	Ciência da Computação	Matemãtica
Currículo			
Currículo Antigo	42,05% (555)	53,35% (852)	..
Currículo Novo	57,95% (765)	46,65% (745)	100,00% (516)
Reprovou obrigatória no 1º semestre			
Sim	48,18% (636)	47,84% (764)	23,45% (121)
Não	51,82% (684)	52,16% (833)	76,55% (395)
Fez disciplina de verão			
Sim	25,57% (282)	30,76% (399)	21,38% (99)
Não	74,43% (821)	69,24% (898)	78,62% (364)

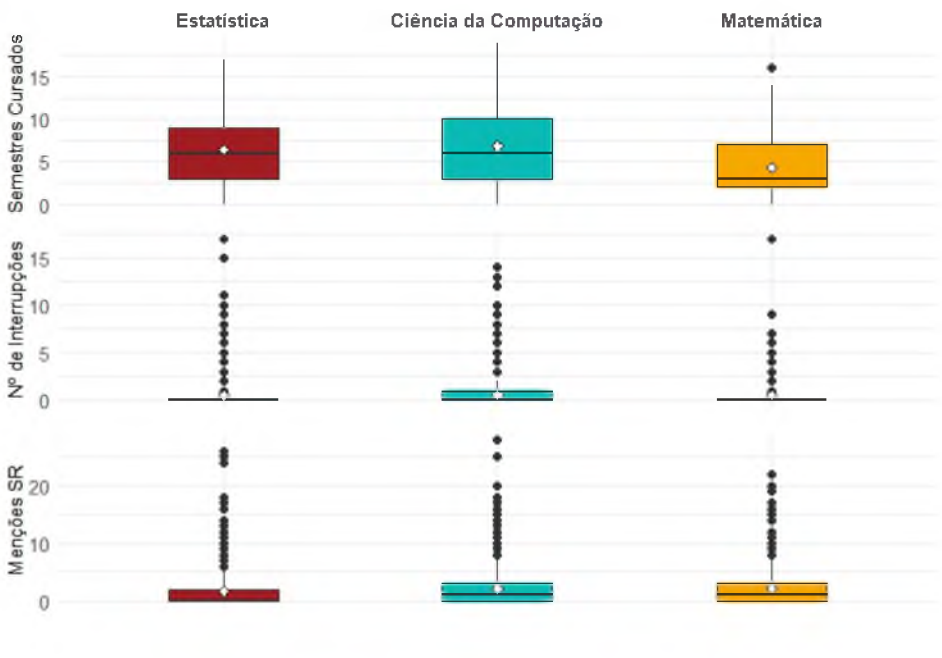
Nota: Sinal convencional utilizado: .. Não se aplica dado numérico.

No que se refere ao tipo de currículo, observa-se que os cursos de Estatística e Ciência da Computação ainda possuem uma parcela considerável de estudantes vinculados ao currículo antigo, com 42,05% e 53,35%, respectivamente. Contudo, no curso de Matemática todos os estudantes analisados pertencem ao currículo novo, pois não houve alterações significativas no currículo do curso ao longo dos anos analisados.

Com relação à reprovação em disciplinas obrigatórias no primeiro semestre, os cursos de Estatística e Ciência da Computação apresentam percentuais semelhantes, com aproximadamente 48% dos estudantes tendo reprovado ao menos uma disciplina obrigatória nesse período. Em Matemática, esse percentual é consideravelmente menor, com apenas 23,45% dos estudantes nessa situação. Esse resultado pode estar relacionado ao fato de que, no primeiro semestre do curso de Matemática, há apenas duas disciplinas obrigatórias, enquanto nos demais cursos a carga de obrigatórias é maior, o que naturalmente amplia as chances de reprovação.

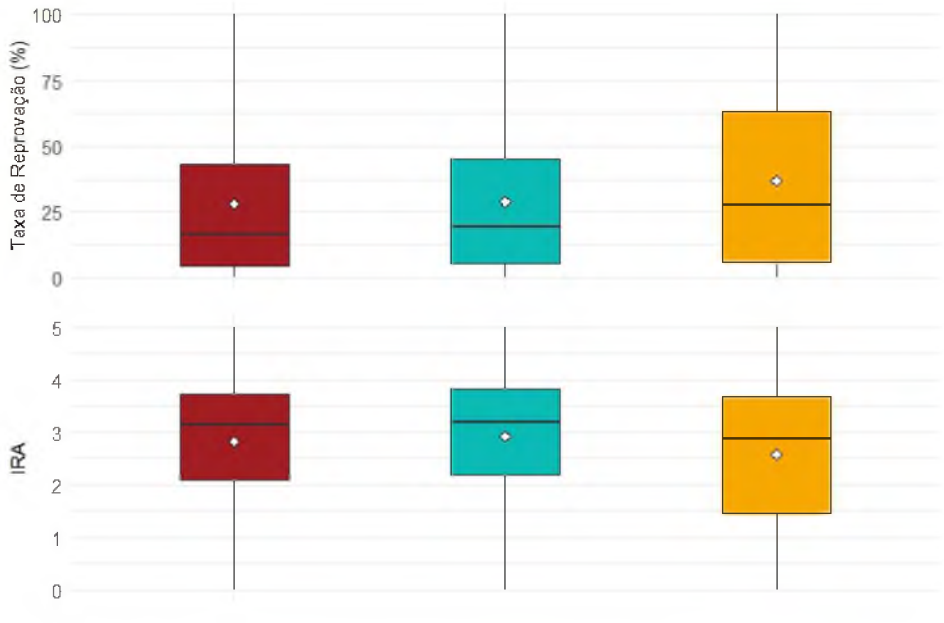
Quanto à participação em disciplinas de verão, os percentuais variam entre os cursos. Ciência da Computação apresenta a maior participação (30,76%), seguida de Estatística (25,57%) e Matemática (21,38%). De modo geral, a oferta de disciplinas de verão foi limitada, especialmente nos períodos posteriores à pandemia, quando a universidade enfrentava um atraso do calendário acadêmico e por consequência, havia menor oferta de componentes curriculares nesse formato. Além disso, o contexto de greve ocorrido em 2024 também contribuiu para a redução das ofertas de verão.

Figura 6: Distribuição das variáveis Semestres Cursados, Nº de Interrupções e Menções SR, por curso. Bacharelado do IE - UnB, 2011-2023



A Figura 6 mostra que os cursos de Estatística e Ciência da Computação possuem distribuição semelhante em termos de semestres cursados, com medianas próximas e leve variação nos extremos. O curso de Matemática apresenta uma mediana inferior em relação aos outros cursos, sugerindo menor tempo médio de permanência. Quanto ao número de interrupções, há maior concentração de estudantes com zero ou poucas interrupções nos três cursos, embora haja casos extremos com número elevado, especialmente em Estatística. A variável Menções SR também se mostra concentrada em valores baixos, com distribuição assimétrica e presença de outliers em todos os cursos, mais visível em Estatística.

Figura 7: Distribuição das variáveis Taxa de Reprovação e IRA, por curso. Bacharelado do IE - UnB, 2011-2023



Em relação à taxa de reprovação, os três cursos apresentam ampla variação, com mediana mais elevada em Matemática, indicando que no curso parte significativa dos estudantes acumula reprovações ao longo da graduação. Já em relação ao IRA, os cursos de Estatística e Ciência da Computação apresentam distribuições muito próximas, com medianas entre 3,0 e 3,5. O curso de Matemática, por sua vez, apresenta maior dispersão e maior presença de estudantes com IRA mais baixo.

Tabela 11: Distribuição dos estudantes segundo ocorrência de evasão. Bacharelados do IE - UnB, 2011-2023

Evasão	**Estatística**	**Ciência da Computação**	**Matemática**
Sim	35,47% (468)	42,76% (683)	63,75% (329)
Não	64,53% (852)	57,24% (914)	36,25% (187)

Com o objetivo de analisar com maior detalhamento a evasão nos cursos, as formas de saída dos estudantes foram agrupadas em quatro categorias principais:

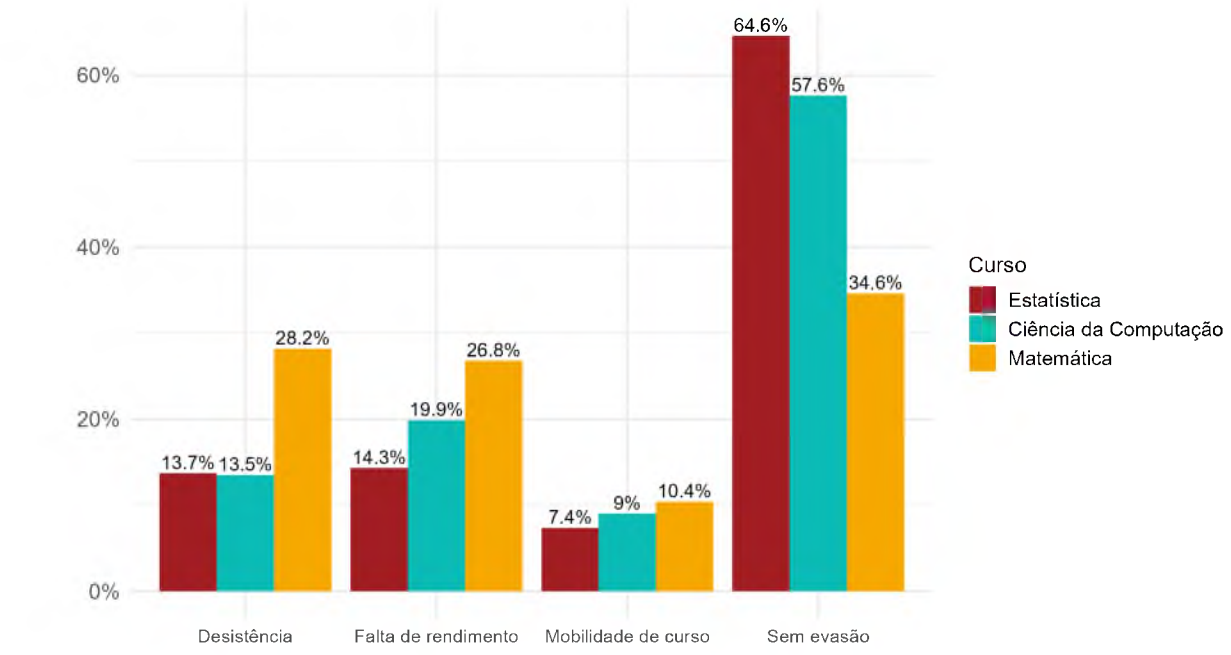
- Desistência: inclui casos como abandono sem matrícula, desligamento por abandono e desligamento voluntário, refletindo situações em que o estudante optou por interromper o curso.
- Falta de rendimento: foram classificados estudantes desligados por não cumprimento

de condição, reprovação em três vezes na mesma disciplina obrigatória e jubilamento, caracterizando evasão por baixo desempenho acadêmico.

- Mobilidade de curso: abrange casos de mudança de curso, transferência e novo vestibular, indicando saída do curso atual com continuidade acadêmica em outro.
- Sem evasão: compreende estudantes ativos, formados ou que integralizaram o curso, não sendo caracterizados como evadidos.

A categoria Falecimento não foi inclusa, pois há 3 casos para os três cursos, sendo um caso extremamente raro.

Figura 8: Distribuição percentual por categoria de forma de saída. Bacharelado do IE - UnB, 2011-2023



A análise da evasão nos cursos do Instituto de Ciências Exatas da UnB revela diferenças significativas entre os cursos tanto no percentual geral de evasão quanto na distribuição das formas de saída.

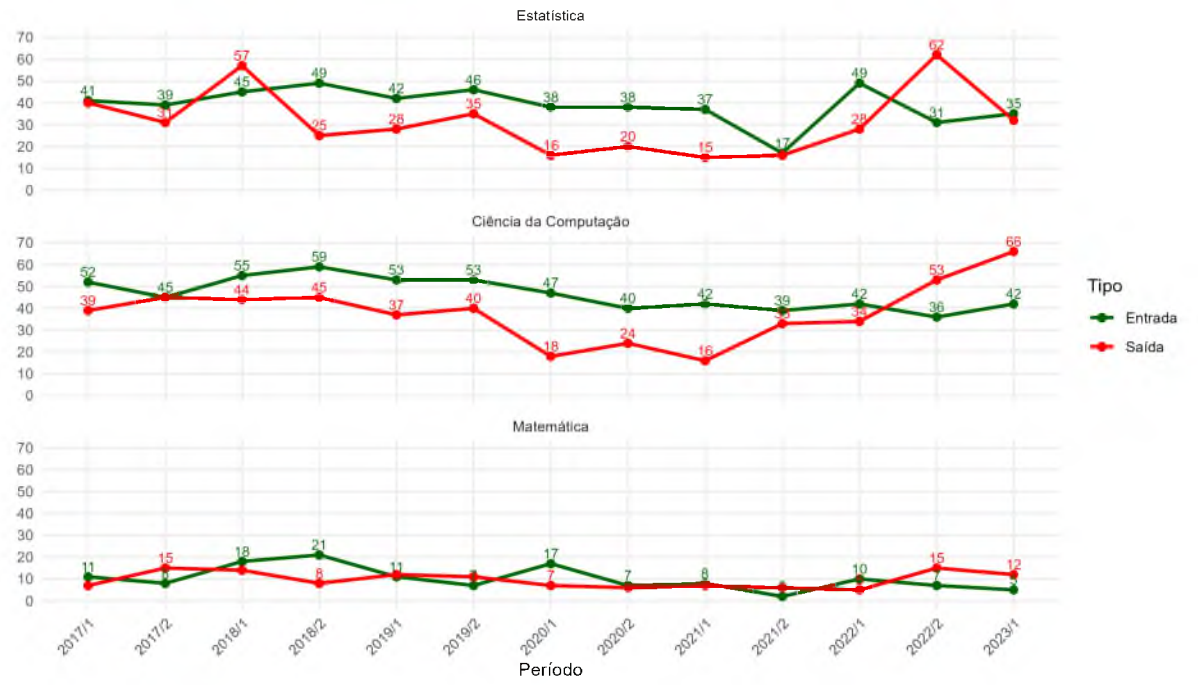
O curso de Matemática apresentou a maior taxa de evasão entre os três, com 63,75% dos estudantes tendo evadido o curso. A maior parte dessas saídas ocorreu por falta de rendimento (28,2%) e desistência voluntária (26,8%), o que sugere dificuldades acadêmicas associadas ao desempenho, além de desmotivação ou insatisfação com o curso. Apenas 34,6% dos estudantes de Matemática permaneceram ou concluíram o curso, o menor percentual de "sem evasão" entre os três cursos.

Já o curso de Ciência da Computação apresentou uma taxa intermediária de evasão (42,76%), com destaque também para falta de rendimento (19,9%) e desistência (13,5%). Entretanto, uma parcela maior dos estudantes (57,6%) permaneceu no curso ou o concluiu, refletindo um cenário menos crítico que o da Matemática.

O curso de Estatística, por sua vez, foi o que apresentou a menor taxa de evasão (35,47%). As principais causas de evasão neste curso foram desistência (13,7%) e falta de rendimento (14,3%), com mobilidade de curso representando apenas 7,4%, o menor valor entre os três cursos.

Para visualizar detalhadamente as informações sobre as formas de saída por curso, consulte a Tabela 33, disponível no Apêndice deste trabalho.

Figura 9: Ingressos e Saídas por Período e Curso (2017/1 a 2023/1). Bacharelado do IE - UnB, 2011-2023



A Figura 9 apresenta a evolução dos ingressos e das saídas dos cursos de bacharelado, no período compreendido entre 2017/1 e 2023/1. Observa-se que o curso de Ciência da Computação manteve, ao longo do tempo, um elevado e relativamente estável número de ingressantes, variando entre 45 e 59 estudantes por período até 2022/1. A partir de 2019/2, no entanto, há uma tendência de queda nos ingressos, com mínima em 2022/2. As saídas (aqui consideradas como o total de desligamentos do curso, sejam por conclusão ou evasão), por sua vez, cresceram de forma consistente, com destaque para os períodos

de 2022/2 e 2023/1, em que o número de egressos superou o de ingressos, atingindo o pico de 66 saídas.

No curso de Estatística, o número de ingressantes manteve-se estável até 2020/2, com valores entre 38 e 49 alunos. Entretanto, a partir de 2021/2, nota-se uma redução significativa no ingresso, chegando a apenas 17 estudantes. Ainda assim, houve uma recuperação pontual em 2022/1. Quanto às saídas, observa-se uma maior variabilidade, com destaque para os picos nos períodos de 2018/1 (57 saídas) e 2022/2 (62 saídas), indicando momentos de maior conclusão ou evasão. Em geral, o curso apresenta um equilíbrio entre ingressos e saídas, embora, nos últimos períodos, as saídas tenham se tornado superiores aos ingressos.

Já o curso de Matemática apresenta os menores volumes absolutos tanto de ingresso quanto de saída entre os três cursos analisados. Os ingressos oscilaram entre 2 e 21 alunos por período, com picos em 2018/2 e 2020/1. A partir de 2021/2, observa-se uma redução acentuada nesse indicador. As saídas, que em geral variavam entre 4 e 11 estudantes, superaram os ingressos em diversos períodos, como em 2022/2 e 2023/1. Tal cenário sugere uma situação crítica, na qual o número de alunos que deixam o curso é superior ao de novos ingressantes, indicando um possível problema de evasão.

De maneira geral, os dados revelam trajetórias distintas entre os cursos. Enquanto Ciência da Computação mantém alta atratividade e mostra crescimento no número de egressos, Estatística apresenta estabilidade com oscilações nas saídas, e Matemática enfrenta dificuldades tanto na captação quanto na retenção de estudantes. A partir de 2021/2, tornam-se mais evidentes os sinais de desequilíbrio entre ingressos e saídas, especialmente nos cursos de Estatística e Matemática, o que pode refletir mudanças no perfil dos alunos, possivelmente, impactos residuais da pandemia da COVID-19. As adaptações ao ensino remoto e as dificuldades enfrentadas por parte dos estudantes durante esse período podem ter influenciado tanto o ingresso quanto a permanência.

5.2 Análise Bivariada

Análise Descritiva Bivariada - por Evasão

O estudo da análise descritiva bivariada a seguir tem como foco a variável de interesse central do estudo: a evasão. Nessa etapa, a evasão é examinada em relação às demais variáveis explicativas como desempenho acadêmico e características institucionais. O objetivo é identificar possíveis associações e padrões que possam contribuir para a compreensão dos fatores relacionados ao abandono dos cursos, servindo de base para a

construção do modelo proposto neste estudo.

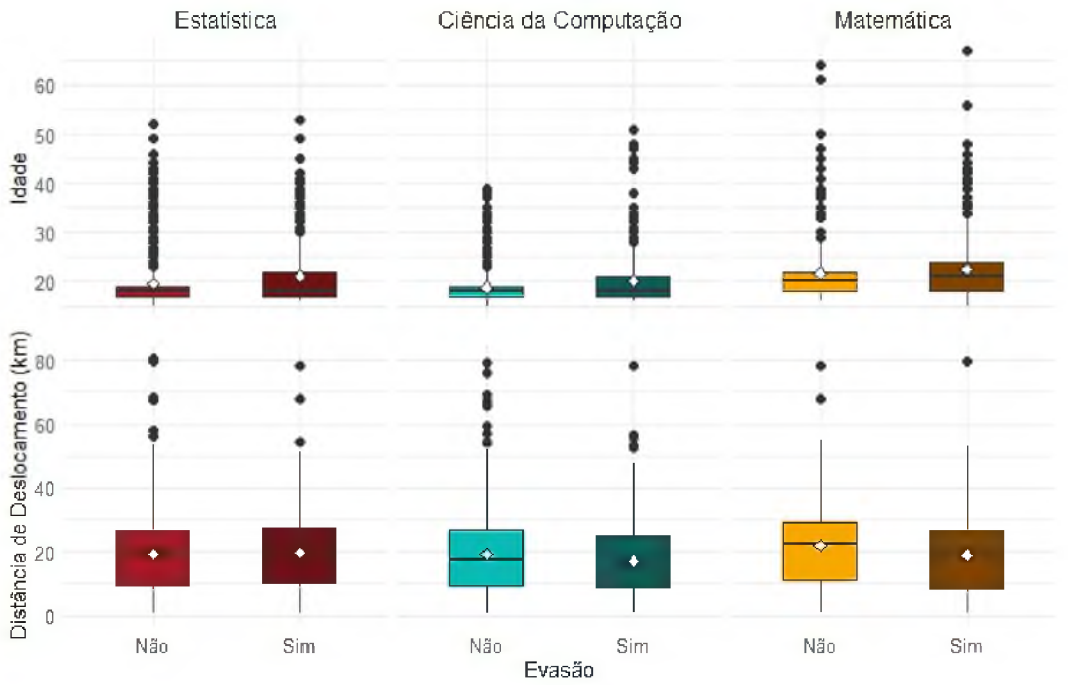
Tabela 12: Percentual de evasão por gênero e faixa de renda. Bacharelados do IE - UnB, 2011-2022

	Estatística	**Ciência da Computação**	**Matemática**
Gênero			
Feminino	30,85% (141)	36,61% (67)	63,16% (72)
Masculino	37,89% (327)	43,56% (616)	65,13% (254)
Renda			
Renda Baixa	33,71% (30)	39,53% (34)	61,11% (22)
Renda Média-Baixa	40,00% (50)	31,17% (48)	58,18% (32)
Renda Média-Alta	36,45% (121)	44,51% (158)	59,85% (82)
Renda Alta	33,79% (221)	43,90% (367)	67,84% (154)

No recorte por gênero, a evasão é consistentemente maior entre estudantes do sexo masculino nos três cursos analisados. Em Estatística, 37,89% dos homens evadiram, frente a 30,85% das mulheres; em Ciência da Computação, os percentuais são de 43,56% e 36,61%, respectivamente. Em Matemática, embora os índices de evasão sejam elevados para ambos os gêneros, os valores são bastante próximos: 65,13% entre homens e 63,16% entre mulheres. Esses resultados podem estar associados não apenas a fatores individuais, mas também ao fato de que os cursos apresentam um perfil majoritariamente masculino, o que pode influenciar a composição geral das taxas de evasão.

No que se refere à renda, observa-se que estudantes provenientes de faixas de renda mais altas representam a maior parte do corpo discente nos três cursos (conforme observado na Tabela 6), enquanto os de renda média-baixa e baixa constituem uma minoria. Em Estatística, por exemplo, apenas cerca de 16% dos estudantes pertencem às categorias de renda média-baixa e baixa; em Ciência da Computação, esse grupo representa aproximadamente 15%, e em Matemática, 18%. No entanto, apesar de numericamente menores, esses grupos apresentam percentuais expressivos de evasão. Em Estatística, estudantes de renda média-baixa possuem a maior taxa de evasão (40,00%), e os de renda baixa também figuram com 33,71%. Em Matemática, a evasão entre estudantes de renda baixa atinge 61,11%, e na Ciência da Computação, 39,53%.

Figura 10: Distribuição das variáveis Idade de Ingresso e Distância de Deslocamento, por curso e evasão. Bacharelado do IE - UnB, 2011-2022



A Figura 10 mostra a distribuição das variáveis idade de ingresso e distância de deslocamento por curso e status de evasão. De modo geral, a idade de ingresso tende a ser ligeiramente mais alta entre os estudantes que evadiram, especialmente nos cursos de Estatística e Matemática. No entanto, essa diferença não é tão acentuada em Ciência da Computação. Já a distância de deslocamento apresenta distribuição semelhante entre evadidos e não evadidos, sugerindo que, isoladamente, essa variável pode não ter impacto expressivo sobre a evasão nos cursos analisados.

Tabela 13: Percentual de evasão segundo categorias institucionais. Bacharelados do IE - UnB, 2011-2022

		Estatística	**Ciência da Computação**	**Matemática**
Tipo de Escola	Particular	38,25% (223)	45,64% (340)	56,36% (93)
	Pública	31,41% (174)	32,21% (210)	49,36% (77)
Começou outra Graduação	Não	36,11% (421)	42,15% (553)	69,14% (289)
	Sim	30,52% (47)	45,61% (130)	40,82% (40)
Período de Ingresso	Pandemia	7,02% (25)	4,75% (20)	13,89% (10)
	Pré-Pandemia	52,43% (401)	65,52% (599)	76,76% (307)
	Transição	26,95% (38)	25,44% (43)	32,26% (10)
Utilizou Cotas	Não	34,37% (355)	45,05% (504)	65,67% (289)
	Sim	39,37% (113)	37,45% (179)	52,63% (40)

Na Tabela 13 em relação ao tipo de escola, a evasão tende a ser maior entre estudantes oriundos de escolas particulares, especialmente em Ciência da Computação (45,64%) e Matemática (56,36%). No entanto, essa diferença é menos pronunciada em Estatística. Para a variável "Começou outra Graduação", observa-se que, em todos os cursos, estudantes que não haviam cursado outra graduação antes apresentam maiores percentuais de evasão, com destaque para Matemática (69,14%). Isso pode indicar que experiências prévias no ensino superior estão associadas a maior permanência.

No que se refere ao uso de cotas, destaca-se que os estudantes cotistas apresentam taxas de evasão mais elevadas em Estatística, enquanto em Matemática e Ciência da Computação o maior percentual está entre os não cotistas.

Em relação ao período de ingresso, observa-se que a maioria dos estudantes dos três cursos iniciou a graduação antes da pandemia de Covid-19, o que é esperado dado o recorte temporal da base de dados (2011–2023). Ainda assim, chama atenção o grupo de estudantes classificados como de transição, que foram aqueles que ingressaram antes da pandemia e concluíram ou evadiram durante esse período. Embora numericamente menores, esses estudantes apresentam percentuais relevantes de evasão, especialmente em Estatística (26,95%) e Ciência da Computação (25,44%), sugerindo que o contexto pandêmico pode ter impactado negativamente a trajetória acadêmica dos discentes ativos naquele período.

Tabela 14: Percentual de evasão segundo características acadêmicas. Bacharelados do IE - UnB, 2011-2022

	Estatística	**Ciência da Computação**	**Matemática**
Currículo			
Antigo	41,44% (230)	57,86% (493)	..
Novo	31,11% (238)	25,50% (190)	63,76% (329)
Reprovou obrigatória 1º semestre			
Sim	54,87% (349)	62,30% (476)	82,64% (100)
Não	17,40% (119)	24,85% (207)	57,97% (229)
Fez disciplina de verão			
Sim	23,05% (65)	30,08% (120)	53,54% (53)
Não	46,17% (379)	57,91% (520)	73,08% (266)

Nota: Sinal convencional utilizado: .. Não se aplica dado numérico.

Em relação ao currículo, nos cursos de Estatística e Ciência da Computação observa-se que a evasão é mais elevada entre estudantes vinculados ao currículo antigo, 41,44% e 57,86%, respectivamente em comparação aos que cursaram o currículo novo. No curso de Matemática não é possível fazer essa comparação.

No que diz respeito à reprovação em disciplinas obrigatórias no primeiro semestre, verifica-se uma forte tendência com maiores taxas de evasão. Estudantes que reprovaram nesse período inicial apresentam taxas consideravelmente mais altas de evasão: 54,87% em Estatística, 62,30% em Ciência da Computação e 82,64% em Matemática. Esses dados sugerem que dificuldades acadêmicas logo no início do curso podem ser um fator relevante de risco para o abandono. Abaixo pode-se observar as frequências de reprovações em disciplinas do primeiro semestre de cada curso:

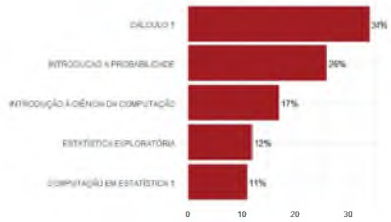


Figura 11: Estatística

Figura 12: Ciência da Computação

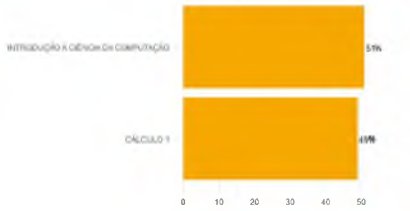
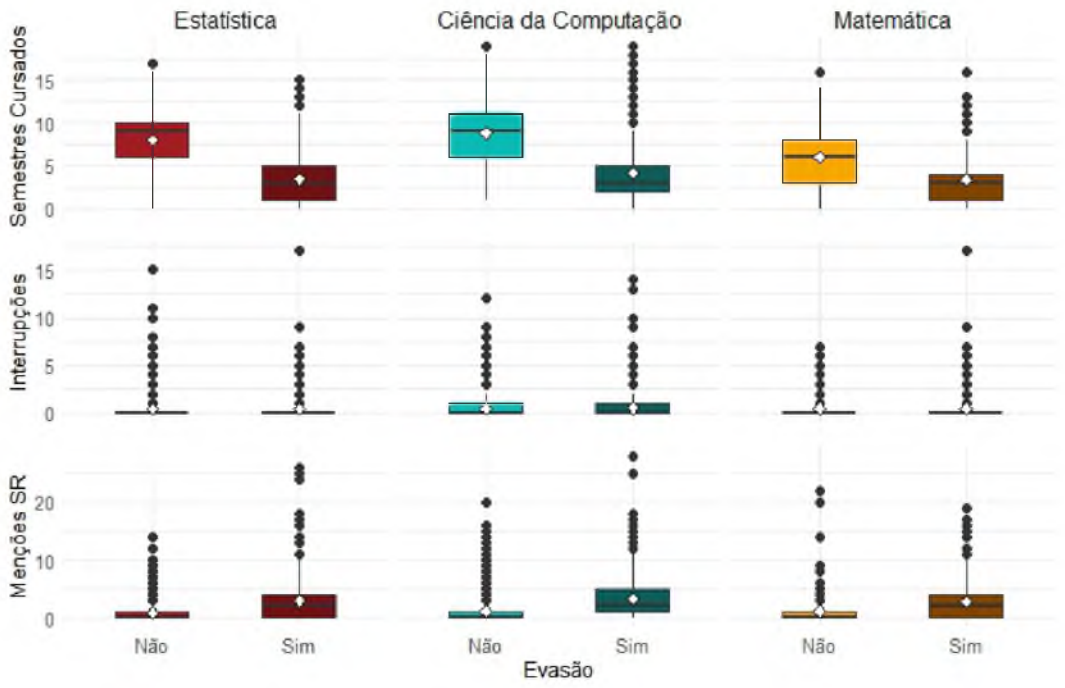


Figura 13: Matemática

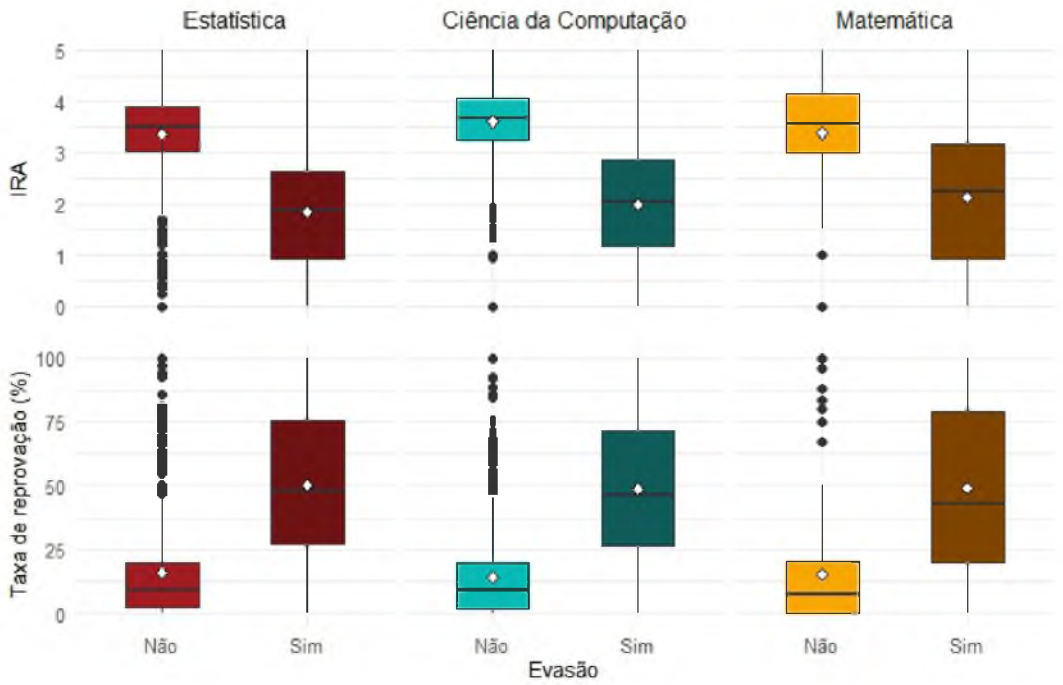
A evasão é maior entre estudantes que não cursaram disciplinas de verão em todos os cursos. Em Estatística, a diferença chega a quase o dobro (46,17% contra 23,05%), padrão também observado em Ciência da Computação (57,91% contra 30,08%) e Matemática (73,08% contra 53,54%). Isso sugere que a participação nessa modalidade pode contribuir para a permanência, ao facilitar a recuperação ou o avanço acadêmico.

Figura 14: Distribuição das variáveis Semestres Cursados, Interrupções e Menções SR, por curso e evasão. Bacharelado do IE - UnB, 2011-2022



Na Figura 14, observa-se que estudantes que evadiram tendem a cursar menos semestres, o que é esperado, dado que não completaram a trajetória acadêmica. Em todos os cursos, há uma diferença clara entre os grupos, com menor média de semestres cursados entre os evadidos. Quanto ao número de interrupções (semestres sem cursar disciplinas), nota-se que estudantes que evadiram apresentam maior concentração de valores acima da média em comparação aos que permaneceram, especialmente em Estatística. No caso das menções SR, associadas a rendimento insatisfatório por ausência, o padrão é semelhante: estudantes evadidos concentram mais registros dessa menção nos três cursos.

Figura 15: Distribuição das variáveis Taxa de Reprovação e IRA, por curso e evasão. Bacharelado do IE - UnB, 2011-2022



Na Figura 15, a taxa de reprovação mostra-se visivelmente mais elevada entre os estudantes evadidos. Esse padrão é nítido em todos os cursos, sugerindo que altos índices de reprovação podem estar diretamente relacionados à evasão. Por outro lado, o IRA, tende a ser mais baixo entre os estudantes evadidos. Isso reforça a associação entre desempenho acadêmico insatisfatório e maior propensão ao abandono da graduação.

Análise de associação entre variáveis explicativas e evasão

Antes da construção dos modelos, é fundamental investigar a associação entre as variáveis explicativas e a variável de interesse (evasão), a fim de identificar os fatores mais relevantes para o estudo. A tabela seguinte apresenta os resultados dos testes aplicados para verificar essa associação. Entre variáveis quantitativas e qualitativas, foram utilizados modelos de regressão logística simples. Já entre variáveis qualitativas, utilizou-se o teste do qui-quadrado de Pearson.

Associação com Evasão	**Estatística**		**Ciência da Computação**		**Matemática**	
	Estatística	p-valor	Estatística	p-valor	Estatística	p-valor
IRA	307.20	< **0.0001**	393.44	< **0.0001**	74.92	< **0.0001**
Gênero	6.46	**0.0110**	3.19	0.0737	0.14	0.6986
Idade	23.52	< **0.0001**	43.13	< **0.0001**	1.49	0.2214
Renda	2.14	0.5430	9.65	**0.0217**	3.39	0.3353
Distância	0.16	0.6819	8.83	**0.0029**	5.54	**0.0185**
Tipo Escola	5.85	**0.0156**	26.26	< **0.0001**	1.57	0.2089
Cotas	2.46	0.1167	7.88	**0.0050**	4.77	**0.0288**
Currículo	14.99	**0.0001**	170.04	< **0.0001**	..	..
Pandemia	221.92	< **0.0001**	458.83	< **0.0001**	121.57	< **0.0001**
Começou outra Grad.	1.85	0.1732	1.14	0.2839	27.55	< **0.0001**
Interrupções	0.0567	0.8117	0.4187	0.5176	0.0075	0.9311
Reprovou obr.	202.26	< **0.0001**	228.38	< **0.0001**	24.39	< **0.0001**
Taxa de reprovação	273.67	< **0.0001**	362.36	< **0.0001**	79.55	< **0.0001**
Menções SR	134.24	< **0.0001**	142.41	< **0.0001**	26.59	< **0.0001**
Trancamentos	20.32	< **0.0001**	72.89	< **0.0001**	13.57	**0.0002**
Semestres cursados	336.20	< **0.0001**	376.82	< **0.0001**	64.59	< **0.0001**
Verão	46.62	< **0.0001**	85.60	< **0.0001**	13.87	**0.0002**

Nota: Sinal convencional utilizado: .. Não se aplica dado numérico.

Os resultados revelam que em todos os cursos, variáveis acadêmicas apresentaram associação estatisticamente significativa ($p < 0{,}05$) com a evasão em todos os cursos analisados. Entre elas destacam-se: o Índice de Rendimento Acadêmico (IRA), currículo, o período da pandemia, a reprovação em disciplinas obrigatórias, a taxa de reprovação, trancamentos e se cursou verão. Esses fatores, fortemente associados à evasão, indicam que aspectos relacionados ao desempenho e à trajetória acadêmica possuem papel central na permanência do estudante na universidade.

Outras variáveis apresentaram associação significativa em dois dos três cursos. É o caso do tipo de escola cursada no ensino médio e se utilizou cotas (significativas em Estatística e Computação, e Computação e Matemática, respectivamente), além das menções SR e da quantidade de semestres cursados. Esses resultados sugerem que, embora não sejam determinantes universais, esses fatores influenciam a evasão em contextos específicos.

Ademais, algumas variáveis mostraram associação estatisticamente significativa apenas em um curso. A idade e o gênero apresentaram associação com a evasão apenas no curso de Estatística, enquanto a renda e a distância da universidade foram significativas apenas em Ciência da Computação. Já a variável "começou outra graduação"teve associação apenas no curso de Matemática. Por fim, variáveis como "interrupções" e "começou outra graduação" (nos cursos de Estatística e Computação) não apresentaram associação significativa com a evasão.

Esses resultados reforçam a importância de considerar tanto fatores acadêmicos quanto características individuais na análise de estudos sobre evasão acadêmica.

Associação entre as variáveis explicativas

Além de analisar a relação entre as variáveis explicativas e a variável de interesse, é igualmente importante verificar a correlação entre as próprias variáveis explicativas. Dessa forma, foram utilizados os mesmos testes aplicados anteriormente, mas em relação à associação entre variáveis quantitativas, foi utilizado o teste de correlação de Spearman, que é apropriado quando as variáveis não apresentam distribuição normal.

A análise das associações entre variáveis explicativas revelou padrões importantes relacionados à trajetória acadêmica dos estudantes e destacou a necessidade de atenção à presença de multicolinearidade nos modelos.

No curso de Estatística, observou-se uma forte correlação negativa entre o IRA e a taxa de reprovação ($\rho = -0{,}9034$), evidenciando que estudantes com menor desempenho geral acumulam maiores índices de reprovação. Além disso, a variável reprovou obrigatória no 1º semestre apresentou associação altamente significativa com as menções SR ($p < 0{,}0001$), sugerindo sobreposição de informação entre indicadores de baixo desempenho inicial.

Nos cursos de Estatística e Ciência da Computação, a variável utilizou cotas mostrou associação significativa com o tipo de escola de ensino médio do estudante ($p < 0{,}0001$), indicando um ponto de atenção a essas variáveis no mesmo modelo.

No curso de Ciência da Computação, destacou-se ainda a associação significativa entre a taxa de reprovação e a pandemia ($p < 0{,}0001$), possivelmente refletindo os impactos do ensino remoto e das mudanças na dinâmica acadêmica durante esse período.

No curso de Matemática, o IRA também se mostrou fortemente associado a variáveis de desempenho, como taxa de reprovação ($\rho = -0{,}7996$) e reprovou obrigatória no 1º semestre ($p < 0{,}0001$), reforçando sua centralidade como indicador de rendimento acadêmico. Além disso, a variável semestres cursados apresentou associação significativa com trancamentos e verão ($p < 0{,}0001$), sugerindo que estudantes com maior tempo de curso estão mais expostos a situações de interrupção e à necessidade de cursar disciplinas em períodos alternativos

Essas correlações apontam para a importância de uma seleção criteriosa das variáveis nos modelos finais, a fim de evitar redundâncias e garantir a estabilidade das estimativas. A consideração da multicolinearidade foi, portanto, fundamental na definição das especificações finais de cada modelo ajustado.

5.3 Modelagem

O objetivo deste estudo é identificar os principais fatores associados à evasão de discentes dos cursos de bacharelado do Instituto de Ciências Exatas (IE), por meio de modelos de regressão logística. Nesse sentido, foi ajustado, em cada curso individualmente, um modelo inicial contendo todas as potenciais variáveis explicativas. Em seguida, foi aplicado o procedimento de seleção do tipo *stepwise*. Com base nas variáveis selecionadas, foram conduzidos testes da razão de verossimilhança, com nível de significância de 5%, para avaliar individualmente a contribuição de cada uma, bem como possíveis interações entre os parâmetros. A partir desses resultados, foram construídos modelos com maior relevância estatística para explicar o fenômeno de evasão em cada contexto.

A base de treinamento é composta por 70% dos dados, selecionados aleatoriamente, enquanto a base de validação é composta pelos 30% restantes. Durante o processo de modelagem, foi necessário recodificar algumas variáveis categóricas para evitar problemas decorrentes da baixa frequência em certas categorias e da ausência de significância estatística. A variável renda, por exemplo, originalmente composta por quatro categorias (Renda Alta, Média-Alta, Média-Baixa e Baixa), foi reestruturada em apenas duas: Renda Alta e Renda Baixa. Essa simplificação teve como objetivo facilitar a interpretação do modelo, eliminando categorias intermediárias com pouca representatividade e sem efeito estatisticamente significativo.

De forma semelhante, a variável Pandemia, inicialmente dividida em três categorias (Pré-pandemia, Transição e Pandemia), também foi recodificada. Nesse processo, as categorias Transição e Pandemia foram reunidas sob o rótulo "Sim", representando os períodos que já passaram pela pandemia, enquanto a categoria Pré-pandemia foi reclassificada como "Não", indicando aqueles que não passaram pela pandemia. Essa simplificação resultou em duas categorias finais, tornando a análise mais clara e consistente, além de refletir os resultados preliminares dos modelos, que mostraram que o impacto das fases intermediárias não se distinguia significativamente do observado durante a pandemia, ao passo que a comparação com o período pré-pandêmico se mostrava estatisticamente mais relevante.

Além disso, em análises anteriores foi constatada uma forte correlação entre as variáveis IRA e taxa de reprovação. Com o intuito de evitar problemas de multicolinearidade, optou-se por ajustar dois modelos distintos para cada curso, cada um incluindo apenas uma dessas variáveis. Essa abordagem permitiu avaliar o efeito isolado de cada fator no risco de evasão, contribuindo para uma interpretação mais precisa dos resultados.

5.3.1 Modelo do curso de Estatística

Considerando que as variáveis IRA e taxa de reprovação apresentam alta correlação entre si e o processo de seleção do modelo para o curso de Estatística usando o critério *stepwise*, foram testadas separadamente as duas variáveis nos modelos, com todas as demais variáveis explicativas. Ambos os caminhos (utilizando IRA ou taxa de reprovação) resultaram na mesma estrutura de modelo, evidenciada na Tabela 15.

Tabela 15: Modelo do *Stepwise* para o curso de Estatística

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	5,3227	0,5927	80,6432	$< 0{,}0001$
Menções SR	0,4641	0,0857	29,3487	$< 0{,}0001$
Semestres cursados	-0,7431	0,0715	108,0562	$< 0{,}0001$
Pandemia (Sim)	-4,5912	0,5177	78,6411	$< 0{,}0001$
Reprovou obrigatória (Não)	-1,4758	0,3841	14,7616	0,0001

A partir do modelo reduzido obtido pelo procedimento *stepwise*, foram realizadas etapas adicionais de avaliação por meio do teste de razão de verossimilhança, com nível de significância de 5%. Nessa fase, testou-se a inclusão individual de variáveis explicativas adicionais, bem como a introdução de possíveis interações entre as variáveis já selecionadas. Duas interações entre variáveis já presentes mostraram-se estatisticamente significativas: entre semestres cursados e pandemia, e entre reprovação em disciplina obrigatória no 1º semestre e pandemia. A inclusão desses termos de interação não apenas aumentou a qualidade do ajuste do modelo, como também resultou em alterações nas estimativas dos coeficientes em comparação ao modelo anterior, indicando que os efeitos dessas variáveis variam conforme o contexto pandêmico. Isso reforça a importância de considerar interações para capturar a complexidade do fenômeno da evasão.

Adicionalmente, a variável utilizou cotas foi incluída no modelo com base no teste da razão de verossimilhança, apresentando significância estatística na fase de construção do modelo. No entanto, ao aplicar esse modelo à base de validação, observou-se que sua inclusão não contribuiu para a melhoria do desempenho preditivo. Quando a variável foi retirada, o valor do teste de Hosmer e Lemeshow aumentou, indicando melhor aderência do modelo aos dados e reforçando a decisão de excluí-la da versão final. O AIC do modelo é 233,460.

Tabela 16: Modelo final para o curso de Estatística, 2011-2023

Parâmetro	Estimativa	Erro padrão	Estatística	p-valor
Intercepto	7,5878	0,9765	60,3839	< 0,0001
Menções SR	0,5697	0,0926	37,8783	< 0,0001
Semestres cursados	-0,9812	0,1063	85,2082	< 0,0001
Semestres × Pandemia	0,4037	0,1485	7,3945	0,0065
Pandemia (Sim)	-8,2273	1,1169	54,2637	< 0,0001
Pandemia × Reprovou obrigatória	3,3208	0,8127	16,6982	< 0,0001
Reprovou obrigatória (Não)	-2,6826	0,5445	24,2723	< 0,0001

O teste de Hosmer e Lemeshow foi utilizado para testar adequabilidade do ajuste do modelo. De acordo com os resultados obtidos na Tabela 17, o p-valor é superior ao nível de significância de 5%. Portanto, não há evidências para rejeitar a hipótese nula de que o modelo está bem ajustado aos dados.

Tabela 17: Teste de adequação de ajuste - Modelo Estatística

Teste	Estatística	p-valor
Hosmer e Lemeshow	10,9028	0,2073

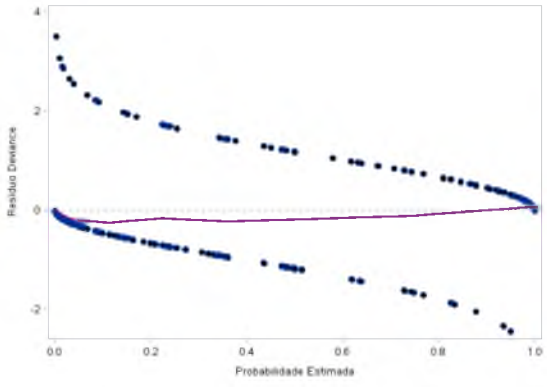


Figura 16: Resíduos *Deviance*

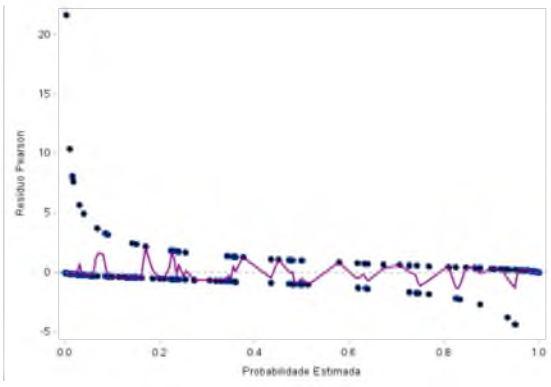


Figura 17: Resíduos de *Pearson*

Além disso, os gráficos de resíduos 16 e 17 fornecem evidências adicionais de adequação. O gráfico dos resíduos *deviance* apresenta valores simetricamente distribuídos ao redor da linha zero, com ausência de padrões sistemáticos, o que é um indicativo de que o modelo não apresenta grandes desvios estruturais.

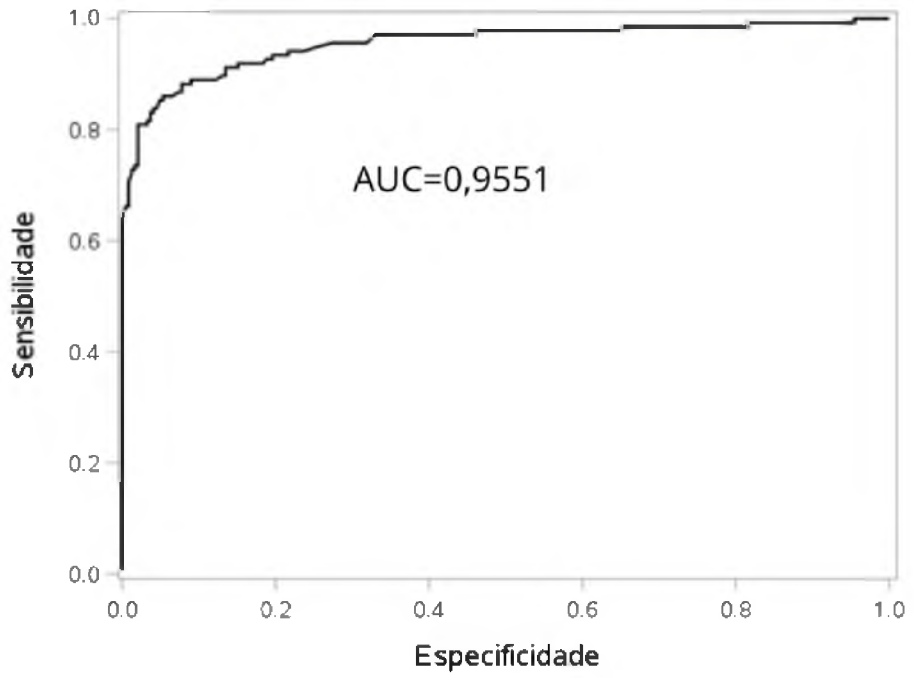


Figura 18: Curva ROC - Modelo de Estatística

A curva ROC do modelo ajustado revela excelente capacidade preditiva, com AUC = 0,9551. Isso indica que o modelo é altamente eficaz em discriminar a evasão. Foi utilizado o ponto de corte de 0,55.

Tabela 18: Matriz de Confusão

		Observado	
		$Y = 1$	$Y = 0$
Previsto	$\hat{Y} = 1$	114	11
	$\hat{Y} = 0$	22	249

A matriz de confusão gerada com a base de validação mostrou que o modelo apresenta um desempenho preditivo consistente. A acurácia geral foi de 91,4%, com sensibilidade de 83,8% e especificidade de 95,8%. Além disso, a precisão foi de 91,2% e o *F1-score* alcançou 87,3%, refletindo um equilíbrio eficaz entre a identificação correta dos casos de evasão e a redução de falsos positivos.

Tabela 19: Razão de chances e IC 95% - Modelo final Estatística

Variável Explicativa	Razão de Chances	IC 95%
Menções SR	1,768	[1,475; 2,121]
Semestres cursados	0,375	[[0,304; 0,463]
Semestres × Pandemia	1,497	[1,117; 2,007]
Pandemia (Sim)	0,00027	[0,00004; 0,002]
Pandemia × Reprovou obrigatória	27,713	[5,57; 137,659]
Reprovou obrigatória	0,068	[0,024; 0,195]

De acordo com Hosmer e Lemeshow (2000), em modelos de regressão logística com termos de interação, os coeficientes das variáveis que participam dessas interações não devem ser interpretados isoladamente. A interpretação adequada deve considerar o efeito conjunto das variáveis envolvidas na interação. No modelo ajustado para evasão no curso de Estatística, há duas interações significativas: entre "Semestres cursados"e "Pandemia (Sim)", e entre "Reprovou disciplina obrigatória"e "Pandemia (Sim)".

A interação entre "Semestres cursados"e "Pandemia"revela que o impacto do tempo de curso sobre a evasão se modificou após a pandemia. No período Pré-Pandemia, cada semestre adicional cursado está associado a uma redução de aproximadamente 62,5% na chance de evasão (*Odds Ratio* (OR) = 0,375). No entanto, com a pandemia, esse efeito torna-se menos pronunciado: a razão de chances combinada é de aproximadamente 0,561 (0,375 × 1,497), o que representa uma redução de 43,9% na chance de evasão a cada semestre adicional. Esse enfraquecimento do efeito pode estar relacionado ao fato de que, atualmente, os estudantes tendem a permanecer mais tempo no curso antes de evadir, talvez pela flexibilização dos critérios para desligamento devido a pandemia. Diferente do comportamento pré-pandêmico, em que os casos de evasão ocorriam predominantemente nos primeiros semestres.

Já a interação entre "Reprovou obrigatória"e "Pandemia"mostra uma mudança substancial no padrão de evasão: enquanto, no período pré-pandêmico, reprovar uma disciplina obrigatória estava associada a uma redução da evasão (OR = 0,068), após a pandemia, esse efeito é revertido, passando de uma redução da evasão para um aumento de aproximadamente 88,5% na chance de evasão (OR combinada = 1,885; 0,068 × 27,713). Essa mudança pode estar relacionada às flexibilizações acadêmicas implementadas durante e um período após a pandemia, como a possibilidade de trancamento ou retirada de disciplinas a qualquer momento do semestre. Assim, casos de reprovação, mesmo com essas facilidades disponíveis, podem indicar um desengajamento mais severo, sugerindo que o aluno já está propenso à evasão.

Adicionalmente, a variável "Menções SR", que não participa de interações, pode ser interpretada diretamente: cada menção SR adicional, mantendo-se constantes as demais variáveis do modelo, está associada a um aumento de 76,8% na chance de evasão (OR = 1,768; IC 95%: [1,475; 2,121]), o que reforça a influência do baixo desempenho acadêmico na decisão de evadir do curso.

Como apresentado nas seções de análise descritiva e de associação, as variáveis acadêmicas demonstram inter-relação. No modelo estatístico referente ao curso de Estatística, a inclusão simultânea das variáveis Menções SR e Reprovou Obrigatória resultou em uma melhora na análise dos resíduos, quando comparado a modelos em que uma dessas variáveis era omitida. Por outro lado, as variáveis IRA e Taxa de Reprovação não foram incorporadas ao modelo, pois, conforme ilustrado na Figura 15, ambas apresentam valores extremos que comprometem sua estabilidade e dificultam a avaliação de seu potencial discriminativo em relação à evasão, no contexto da base de dados do curso de Estatística.

5.3.2 Modelo do curso de Ciência da Computação

Assim como no modelo anterior, para o curso de Ciência da Computação, foram considerados modelos principais que foram construídos separadamente: um com a variável IRA e outro com a taxa de reprovação.

Tabela 20: Modelo com IRA do *Stepwise* para o curso de Ciência da Computação

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	10,2042	1,2625	65,3254	$< 0{,}0001$
IRA	-1,4810	0,3268	20,5435	$< 0{,}0001$
Menções SR	0,2613	0,0731	12,7678	0,0004
Semestres cursados	-0,6857	0,0752	83,1799	$< 0{,}0001$
Segundo grau - Particular	1,1040	0,4412	6,2622	0,0123
Pandemia (Sim)	-4,1575	0,6187	45,1497	$< 0{,}0001$
Reprovou obrigatória (Não)	-0,8079	0,3712	4,7363	0,0295
Trancamentos (Não)	-1,2231	0,5171	5,5940	0,0180

Tabela 21: Modelo com Taxa de Reprovação do *Stepwise* para o curso de Ciência da Computação

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	5,6022	0,8995	38,7878	$< 0{,}0001$
Taxa de reprovação (%)	0,0230	0,0110	4,3724	0,0365
Menções SR	0,3325	0,0722	21,2198	$< 0{,}0001$
Semestres cursados	-0,7430	0,0778	91,2743	$< 0{,}0001$
Segundo grau - Particular	1,0127	0,4258	5,6562	0,0174
Pandemia (Sim)	-3,9134	0,5510	50,4406	$< 0{,}0001$
Reprovou obrigatória (Não)	-0,8813	0,3962	4,9486	0,0261
Trancamentos (Não)	-1,6619	0,5008	11,0124	0,0009

Ao longo do processo de testes adicionais de razão de verossimilhança, algumas variáveis, como Tipo de Escola, Trancamentos e Taxa de Reprovação, inicialmente significativas, foram sendo excluídas em razão da perda de significância estatística (p-valor elevado) à medida que outras variáveis mais explicativas eram inseridas nos modelos. Além disso, os modelos com IRA e com Taxa de Reprovação foram comparados também por meio do teste de razão de verossimilhança.

Em uma das versões intermediárias do modelo, estava incluída a variável Reprovou Obrigatória 1º semestre e sua interação com Pandemia. Apesar de ambas apresentarem significância estatística, a análise dos resíduos revelou um ajuste inadequado: os resíduos *deviance* e de *Pearson* apresentavam assimetrias e padrões indesejáveis. Suspeitou-se que a variável Menções SR (menção atribuída diretamente à reprovação) poderia estar capturando a mesma informação da variável de reprovação, resultando em sobreposição de efeitos (multicolinearidade implícita).

Para investigar essa hipótese, foram testadas duas abordagens: a retirada da variável reprovou obrigatória (e sua interação) mantendo Menções SR; a retirada de Menções SR e reintrodução da variável reprovou obrigatória com a interação.

Ao comparar ambas as estratégias, observou-se que a primeira configuração produziu um comportamento muito mais adequado dos resíduos, além de melhorar o desempenho global do modelo na base de validação. Essa evidência levou à decisão de manter Menções SR como variável explicativa e excluir Reprovou Obrigatória da versão final. O AIC do modelo é 274,076.

Tabela 22: Modelo final do curso de Ciência da Computação, 2011-2023

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	14,7099	1,9637	56,1136	$< 0{,}0001$
Menções SR	0,1971	0,0718	7,5393	0,0060
Semestres cursados	-0,6738	0,0693	94,6236	$< 0{,}0001$
IRA	-2,8208	0,4860	33,6833	$< 0{,}0001$
Pandemia	-12,4440	1,9878	39,1895	$< 0{,}0001$
IRA $\times$ Pandemia	2,3250	0,5434	18,3061	$< 0{,}0001$

O teste de Hosmer e Lemeshow apresentou estatística de 13,2033 com p-valor igual a 0,1050. Como esse valor é maior que o nível de significância adotado (5%), não se rejeita a hipótese nula de que o modelo ajusta-se bem aos dados. Isso indica que as probabilidades estimadas pelo modelo não diferem significativamente das observadas, reforçando a adequação geral do ajuste.

Tabela 23: Teste de adequação de ajuste - Modelo Ciência da Computação

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	13,2033	0,1050

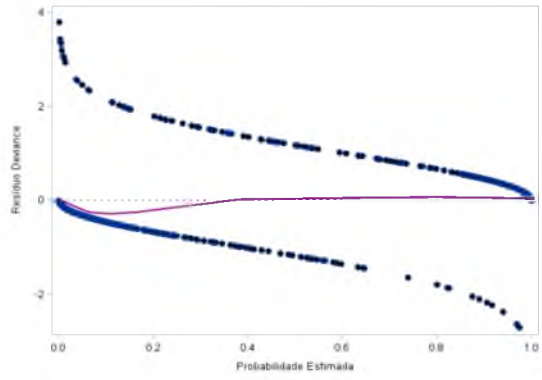


Figura 19: Resíduos *Deviance*

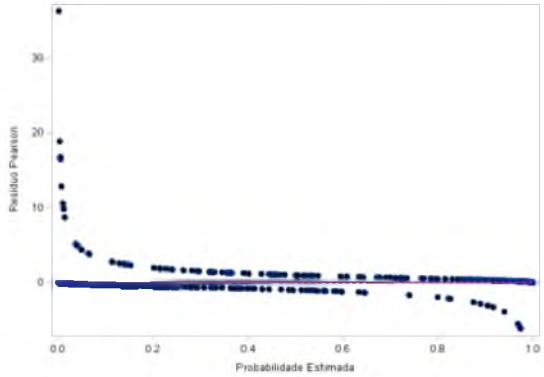


Figura 20: Resíduos de *Pearson*

O gráfico dos resíduos 19 *deviance* exibe uma distribuição relativamente simétrica em torno de zero e sem tendência sistemática. No gráfico 20, a curva suavizada de *Lowess* segue próxima do eixo horizontal, indicando ausência de grandes desvios.

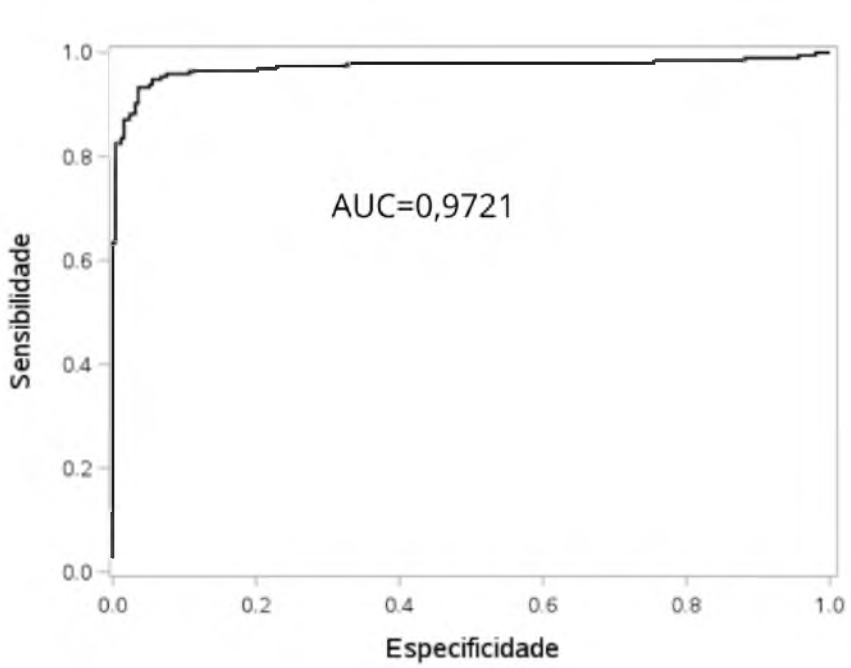


Figura 21: Curva ROC - Modelo de Ciência da Computação

A curva ROC (gráfico 21) apresentada para o modelo do curso de Ciência da Computação demonstra um excelente desempenho discriminativo do modelo. Além disso, o formato da curva, que se aproxima rapidamente do canto superior esquerdo, indica que o modelo mantém alta sensibilidade com baixa taxa de falsos positivos. O melhor ponto de corte é 0,6.

Tabela 24: Matriz de Confusão

		Observado	
		$Y = 1$	$Y = 0$
Previsto	$\hat{Y} = 1$	171	7
	$\hat{Y} = 0$	31	270

A matriz de confusão revela que o modelo obteve uma acurácia de 92,1%, sensibilidade de 84,7% e especificidade de 97,5%. A precisão foi de 96,1%, refletindo uma baixa taxa de falsos positivos, e o *F1-score* alcançou 89,9%, evidenciando um bom equilíbrio entre sensibilidade e precisão. Esses resultados indicam que o modelo possui excelente capacidade para detectar estudantes evadidos.

Tabela 25: Razão de chances e IC 95% - Modelo Final Ciência da Computação

Variável Explicativa	Razão de Chances	IC 95%
Menções SR	1,218	[1,058; 1,402]
Semestres cursados	0,509	[0,443; 0,585]
IRA	0,060	[0,023; 0,159]
Pandemia (Sim)	0,000004	[0,000001; 0,00016]
IRA × Pandemia	10,23	[3,53; 29,66]

No modelo ajustado para evasão no curso de Ciência da Computação, identificou-se uma interação significativa entre o Índice de Rendimento Acadêmico (IRA) e Pandemia. Conforme orientam Hosmer e Lemeshow (2000), a interpretação de variáveis envolvidas em interações deve considerar o efeito conjunto entre elas. No período pré-pandêmico, o IRA apresentou forte efeito protetivo contra a evasão, com uma razão de chances de 0,060, indicando uma redução de aproximadamente 94% na chance de evasão a cada ponto acréscimo no índice. No entanto, a partir da pandemia, esse efeito foi consideravelmente reduzido: a razão de chances combinada é de aproximadamente 0,614 (0,060 × 10,23), o que representa uma redução de cerca de 38,6% na chance de evasão associada ao IRA. Essa mudança de intensidade pode refletir o impacto das alterações institucionais ocorridas durante e após a pandemia, bem como transformações nas percepções sociais sobre o desempenho estudantil. Atualmente, o IRA pode não exercer o mesmo peso simbólico ou psicológico que exercia anteriormente, já que muitos estudantes passaram a relativizar sua importância diante de um contexto mais amplo de bem-estar, saúde mental e reorganização de prioridades acadêmicas.

Além disso, a variável "Menções SR" mostra que cada menção SR adicional está associada a um aumento de 21,8% na chance de evasão (OR = 1,218; IC 95%: [1,058; 1,402]), indicando que o acúmulo de registros sem rendimento pode indicar um importante fator de risco para a evasão.

Já o número de semestres cursados também apresentou efeito significativo, com uma razão de chances de 0,509, o que corresponde a uma redução de aproximadamente 49,1% na chance de evasão a cada semestre cursado, sugerindo que estudantes mais avançados no curso têm menor propensão a evadir.

5.3.3 Modelo do curso de Matemática

Para o curso de Matemática, foi adotado um procedimento semelhante ao aplicado aos demais cursos, com o ajuste de dois modelos iniciais: um com a variável IRA e outro

com a Taxa de Reprovação.

Tabela 26: Modelo com Ira do *Stepwise* para o curso de Matemática

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	5,2139	1,4727	12,5343	0,0004
Idade de ingresso	0,0969	0,0350	7,6471	0,0057
IRA	-1,0651	0,2421	19,3581	$< 0{,}0001$
Semestres cursados	-0,3632	0,0874	17,2700	$< 0{,}0001$
Pandemia (Sim)	-2,8409	0,7371	14,8526	0,0001
Verão (Não)	-1,8622	0,7326	6,4613	0,0110

Tabela 27: Modelo com Taxa de Reprovação do *Stepwise* para o curso de Curso de Matemática

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	2,2989	0,6632	12,0165	0,0005
Taxa de reprovação (%)	0,0363	0,0086	16,7815	$< 0{,}0001$
Semestres cursados	-0,3677	0,0804	20,9018	$< 0{,}0001$
Trancamentos (Não)	-1,7868	0,5083	12,3557	0,0004

O modelo com IRA (Tabela 26) incluiu, além dessa variável, os termos: idade de ingresso, semestres cursados, pandemia, e a variável verão. Todos os parâmetros se mostraram estatisticamente significativos ao nível de 5%.

Já o modelo alternativo, com a Taxa de Reprovação (Tabela 27), identificou como significativas as variáveis semestres cursados, trancamentos e a própria taxa de reprovação. Este modelo apresentou bom ajuste inicial, mas menor número de variáveis explicativas relevantes.

Após o teste de razão de verossimilhança para comparação entre os modelos, foi selecionado o modelo final, apresentado na Tabela 28. Este modelo inclui a variável Taxa de reprovação, Semestres cursados e Pandemia, todas estatisticamente significativas e relevantes para explicar a evasão. O AIC do modelo é 159,64.

Tabela 28: Modelo final do curso de Matemática, 2011-2023

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	1,2380	0,4336	8,1534	0,0043
Taxa de reprovação (%)	0,0357	0,0090	15,4181	$< 0,0001$
Semestres cursados	-0,2929	0,0686	18,2311	$< 0,0001$
Pandemia	-1,8746	0,5103	13,4923	0,0002

Como o p-valor mostrado na Tabela 29 é muito superior a 0,05, não há evidência para rejeitar a hipótese nula de que o modelo se ajusta bem aos dados observados.

Tabela 29: Teste de adequação de ajuste - Modelo Matemática

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	7,9363	0,4397

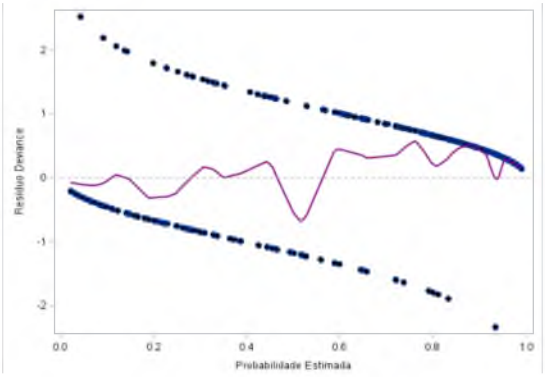


Figura 22: Resíduos *Deviance*

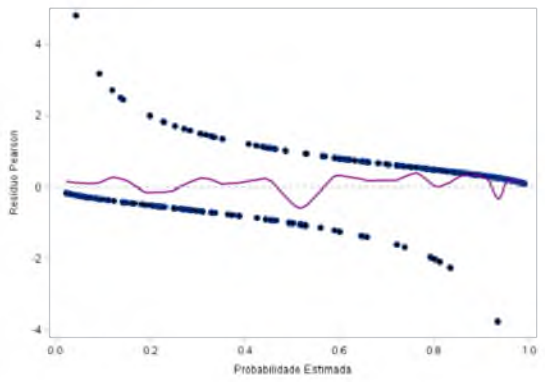


Figura 23: Resíduos de *Pearson*

A análise gráfica dos resíduos para o modelo de Matemática indica que os ajustes se mostram adequados. Na Figura 22, os resíduos do tipo *deviance* apresentam curva suavizada próxima da linha zero, com dispersão relativamente simétrica. Já os resíduos de *Pearson*, ilustrados na Figura 23, se caracterizam pela ausência de padrões sistemáticos ou variações abruptas.

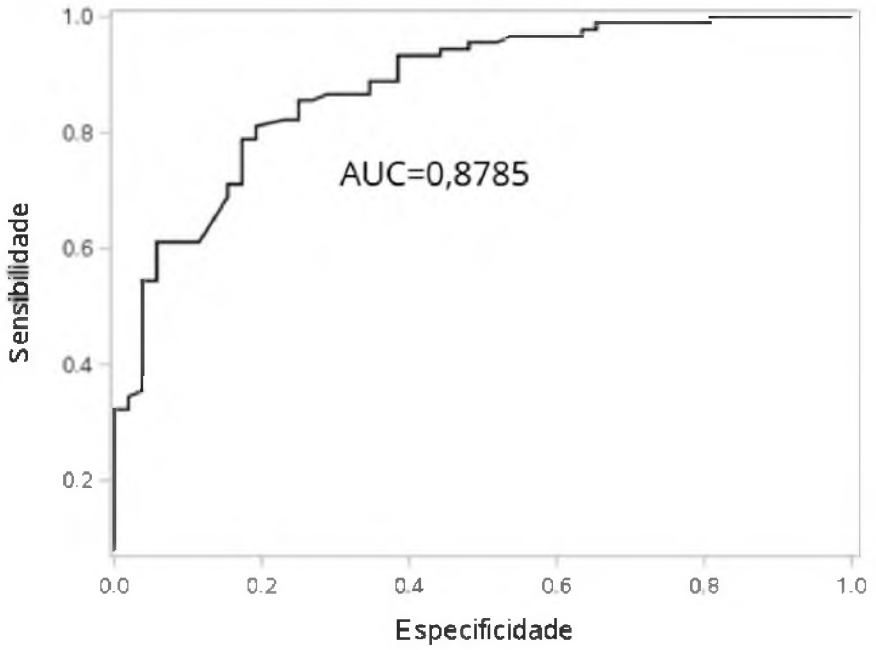


Figura 24: Curva ROC - Modelo de Matemática

A Figura 24 apresenta a curva ROC do modelo ajustado para o curso de Matemática. A área sob a curva (AUC) foi de 0,8785, indicando um alto poder discriminatório do modelo. O ponto de corte é de 0,5.

Tabela 30: Matriz de Confusão

		Observado	
		$Y = 1$	$Y = 0$
Previsto	$\hat{Y} = 1$	70	9
	$\hat{Y} = 0$	24	51

A matriz de confusão mostrou uma acurácia de 78,6%, com sensibilidade de 74,5% e especificidade de 85,0%. A precisão foi de 88,6%, indicando uma boa taxa de acertos entre os casos classificados como evasão, enquanto o *F1-score* alcançou 80,8%, revelando um equilíbrio consistente entre precisão e sensibilidade. Esses resultados sugerem que o modelo apresenta desempenho satisfatório para identificar estudantes evadidos, com boa capacidade de distinguir também os estudantes que permaneceriam.

Tabela 31: Razão de chances e IC 95% - Modelo Final Matemática

Variável Explicativa	**Razão de Chances**	**IC 95%**
Taxa de Reprovação (%)	1,036	[1,018; 1,055]
Semestres cursados	0,746	[0,641; 0,869]
Pandemia	0,153	[0,056; 0,419]

No modelo ajustado para evasão no curso de Matemática, a variável "Taxa de Reprovação (%)" apresentou efeito positivo e estatisticamente significativo sobre a evasão, com razão de chances de aproximadamente 1,036. Isso indica que, para cada ponto percentual a mais na taxa de reprovação, a chance de evasão aumenta em cerca de 3,6%.

Já a variável "Semestres cursados" apresentou uma razão de chances de 0,746, sugerindo que cada semestre adicional cursado está associado a uma redução de 25,4% na chance de evasão, reforçando a tendência de que estudantes mais avançados têm menor propensão a evadir.

A variável "Pandemia" foi negativamente associada à evasão, com uma razão de chances de 0,153. Isso indica que, com a pandemia, as chances de evasão foram 84,7% menores, o que pode refletir os dados da base usada na construção do modelo, em que há, em sua maioria, estudantes evadidos do período pré-pandemia.

Todos os parâmetros incluídos no modelo apresentaram significância estatística (p-valor $< 0{,}01$), reforçando a robustez dos achados.

6 Conclusão

Este trabalho teve como objetivo investigar os fatores associados à evasão dos cursos de bacharelado em Estatística, Ciência da Computação e Matemática do Instituto de Ciências Exatas no Instituto de Ciências Exatas da UnB, a partir de uma abordagem quantitativa que combinou análise descritiva do perfil dos estudantes e modelagem estatística via regressão logística.

Inicialmente, traçou-se o perfil sociodemográfico dos discentes, revelando uma predominância de estudantes jovens, com idade em torno de 20 anos, provenientes em sua maioria de regiões administrativas como Asa Norte, Asa Sul, Águas Claras, Guará e Taguatinga. A distribuição geográfica dos estudantes mostrou-se relativamente similar entre os cursos, com baixa representação de estudantes do Entorno do Distrito Federal. Observou-se também que a maioria dos discentes ainda se encontra em fase inicial ou intermediária do curso, com destaque para os que cursam entre o 3º e o 6º semestre.

No que diz respeito à modelagem estatística, foram ajustados modelos de regressão logística específicos para cada curso, a partir de um modelo inicial com todas as potenciais variáveis explicativas e seleção via procedimento *stepwise*.

Para o curso de Estatística, o modelo final identificou como variáveis mais relevantes as menções SR, os semestres cursados, a pandemia e a reprovação em disciplinas obrigatórias no primeiro semestre, além de interações entre pandemia e essas duas últimas variáveis. O modelo demonstrou que o risco de evasão aumenta conforme o estudante reprova disciplinas essenciais no início da graduação, especialmente no contexto pós-pandemia.

No curso de Ciência da Computação, o modelo final incluiu as variáveis IRA, menções SR, semestres cursados, pandemia e a interação entre IRA e pandemia, demonstrando que estudantes com menor desempenho acadêmico recente são mais propensos à evasão, particularmente no cenário pandêmico.

Já para o curso de Matemática, o modelo final selecionado apresentou como preditores significativos a taxa de reprovação, semestres cursados e pandemia. Observou-se que estudantes com desempenho acadêmico mais baixo (maior taxa de reprovação) têm maior propensão a evadir.

Em síntese, os resultados obtidos permitiram alcançar os objetivos propostos, identificando padrões e fatores relevantes associados à evasão nos três cursos. Evidenciase, de modo consistente, que o desempenho acadêmico, o tempo de permanência no curso

e os impactos da pandemia são elementos centrais na explicação do fenômeno da evasão. Esses achados podem subsidiar a Comissão de Evasão da universidade na criação de políticas institucionais de acompanhamento e permanência de discentes, especialmente direcionadas aos estudantes em risco de evasão nos períodos iniciais do curso e com baixo rendimento acadêmico.

Apesar das contribuições relevantes deste trabalho, algumas limitações devem ser consideradas. As bases de dados extraídas dos sistemas institucionais da Universidade de Brasília apresentaram diversas inconsistências, como registros duplicados, ausência de padronização entre variáveis e uma quantidade expressiva de dados faltantes. Além disso, a escassez de informações sobre o perfil sócio-demográfico e cultural dos estudantes, da sua formação antes de ingressar na universidade e interesse no curso, se recebe bolsas ou outras formas de apoio institucional, restringiu a análise de fatores contextuais potencialmente associados à evasão. Também seria fundamental contar com dados sobre os recursos materiais disponíveis aos discentes, como acesso a equipamentos e internet, especialmente no caso de estudantes que cursaram parte da graduação durante o período pandêmico, quando o ensino remoto foi adotado. Sugere-se, portanto, que formulários de ingresso na universidade passem a incluir perguntas sobre tais condições, possibilitando análises mais precisas. Para estudos futuros, recomenda-se a construção de indicadores que avaliem o avanço do discente com base no fluxo curricular previsto, além da incorporação de variáveis socioeconômicas mais completas.

Referências

AGRESTI, A. *An introduction to categorical data analysis*. 3. ed. New York: John Wiley & Sons, 2019.

AMBIEL, Rodolfo Augusto Matteo; DE OLIVEIRA BARROS, Leonardo. *Relações entre evasão, satisfação com escolha profissional, renda e adaptação de universitários. Revista Psicologia: Teoria e Prática*, v. 20, n. 2, p. 254-267, 2018.

AMBIEL, Rodolfo Augusto Matteo; DOS SANTOS, Acácia Aparecida Angeli; DALBOSCO, Simone Nenê Portela. Motivos para evasão, vivências acadêmicas e adaptabilidade de carreira em universitários. *Psico*, v. 47, n. 4, p. 288-297, 2016.

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.

Berkeley Research. (2024). Five years after COVID, U.S. university students still lag in key areas. University of California Berkeley. Disponível em: ⟨https://vcresearch.berkeley.edu/news/five-years-after-covid-us-university-students-still-lag-key-areas⟩. Acesso em: 6 jul. 2025.

CORREIO BRAZILIENSE. Ensino superior no Brasil tem 57% de evasão na rede pública e privada. *Correio Braziliense*, 24 maio 2024. Disponível em: ⟨https://www.correiobraziliense.com.br/euestudante/ensino-superior/2024/05/6852929-ensino-superior-no-brasil-tem-57-de-evasao-na-rede-publica-e-privada.html⟩. Acesso em: 6 jan. 2025.

DA SILVA PORTO, Vanessa et al. Ensino a distância e evasão no ensino superior no Brasil: evidências no contexto da pandemia. *Revista EDaPECI*, v. 24, n. 3, p. 25-38, 2024.

DE OLIVEIRA, Carlos Henrique Mendes et al. Busca dos fatores associados à evasão: um estudo de caso no Campus Universitário da UFC em Crateús. *Revista Internacional de Educação Superior*, v. 5, p. e019006-e019006, 2019.

FIOR, Camila Alves. Evasão do ensino superior e papel preditivo do envolvimento acadêmico. *Amazônica-Revista de Psicopedagogia, Psicologia escolar e Educação*, v. 13, n. 1, jan-jun, p. 9-32, 2021.

GARCIA, Léo Manoel Lopes da Silva; GOMES, Raquel Salcedo. Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica. *Revista Educar Mais*,

Pelotas, RS, v. 6, p. 937-957, 2022.

GUSSO, Hélder Lima et al. Ensino superior em tempos de pandemia: diretrizes à gestão universitária. *Educação & Sociedade*, v. 41, p. e238957, 2020.

HAIDAR, Ana Clara Arrais. Análise da evasão nos cursos de graduação da Universidade de Brasília utilizando modelos de regressão logística. 2023. Trabalho de Conclusão de Curso (Graduação em Estatística) – Universidade de Brasília, Brasília, 2023.

HOSMER, D. W.; LEMESHOW, S. *Applied Logistic Regression*. [S.l.]: John Wiley and Sons Inc, New York, 2000.

KUTNER, M. H.; NACHTSHEIM, C. J.; NETER, J.; LI, W. *Applied Linear Statistical Models*. 5. ed. New York: McGraw-Hill/Irwin, 2005.

NUNES, Renata Cristina. *Um olhar sobre a evasão de estudantes universitários durante os estudos remotos provocados pela pandemia do COVID-19. Research, Society and Development*, v. 10, n. 3, p. e1410313022-e1410313022, 2021.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA (UNESCO). COVID-19: como a Coalizão Global de Educação da UNESCO está lidando com a maior interrupção da aprendizagem da história. *UNESCO*, 2020. Disponível em: ⟨https://pt.unesco.org/news⟩. Acesso em: 5 jan. 2025.

SILVA, Paulo César Azevedo et al. Permanência em educação superior em uma universidade comunitária do Sul do Brasil. *Revista Educação e Cultura Contemporânea*, v. 18, n. 53, p. 338-352, 2021.

SILVA, Glauco Peres da. Análise de evasão no ensino superior: uma proposta de diagnóstico de seus determinantes. *Avaliação: Revista da Avaliação da Educação Superior (Campinas)*, v. 18, p. 311-333, 2013.

Apêndice

Tabela 32: Distribuição percentual de estudantes por Região Administrativa e por curso

RA	Estatística	Ciência da Computação	Matemática
Plano Piloto	26,4	29,9	27,9
Águas Claras	8,8	7,6	6,0
Taguatinga	7,9	5,7	9,2
Guará	6,9	6,0	9,2
Sobradinho	6,8	6,6	6,7
Sudoeste	4,5	4,9	3,6
Ceilândia	4,0	3,3	4,5
Jardim Botânico	3,2	2,8	1,7
Lagos Sul e Norte	5,7	8,3	5,8
Cruzeiro	2,5	2,3	1,7
Planaltina	2,5	1,4	2,8
Park Way	2,4	1,5	1,7
Samambaia	2,3	2,3	2,8
Paranoá	1,7	0,4	0,4
Entorno	1,6	1,8	2,6
Gama	1,6	1,2	2,1
Núcleo Bandeirante	1,3	1,4	1,5
Octogonal	1,3	1,1	1,5
Riacho Fundo I	1,2	2,2	1,3
Vicente Pires	1,2	1,8	0,9
Santa Maria	1,1	1,5	1,1
Setor Noroeste	1,1	1,2	0,2
Recanto das Emas	1,0	1,1	2,1
São Sebastião	1,0	1,2	0,9
Vila Planalto	0,6	0,6	0,4
Candangolândia	0,4	0,4	0,2
Estrutural	0,4	0,3	0,2
Brazlândia	0,2	0,5	0,4

Tabela 33: Distribuição percentual (e valores absolutos) das formas de saída por curso.

Forma de saída	**Estatística**	**Ciência da Computação**	**Matemática**
Abondono (Nenhuma Matríccula)	3.79% (50)	3.88% (62)	2.91% (15)
Ativo	27.73% (366)	27.49% (439)	11.24% (58)
Concluído	6.67% (88)	6.76% (108)	5.43% (28)
Deslig - não cumpriu condição	12.20% (161)	15.65% (250)	23.06% (119)
Desligamento - Abandono	7.35% (97)	6.14% (98)	15.31% (79)
Desligamento Voluntário	2.27% (30)	3.13% (50)	8.53% (44)
Efetivação de novo cadastro	0.98% (13)	0.63% (10)	0.58% (3)
Falecimento	0.08% (1)	0.13% (2)	–
Formatura	27.58% (364)	19.66% (314)	15.50% (80)
Integralização de Discente	2.58% (34)	3.26% (52)	1.36% (7)
Mudança de Curso	0.38% (5)	0.44% (7)	1.36% (7)
Novo Vestibular	5.83% (77)	7.14% (114)	7.56% (39)
Repr 3 vezes na mesma disc obr	2.12% (28)	4.01% (64)	2.71% (14)
Solicitação Espontânea	0.30% (4)	0.25% (4)	0.58% (3)
Transferência	0.15% (2)	0.56% (9)	0.58% (3)
Cancelamento Judicial	–	0.06% (1)	–
Decisão Administrativa	–	0.06% (1)	–
Desligamento-Força de Convênio	–	0.06% (1)	0.19% (1)
Desligamento Jubilamento	–	0.06% (1)	0.19% (1)
Desligamento por Força de Intercâmbio	–	0.31% (5)	0.19% (1)
Mudança de Turno	–	0.06% (1)	–
Mudança de Habilitação	–	–	2.71% (14)
Outros	–	0.06% (1)	–
Transferência para outra IEs	–	0.19% (3)	–