Universidade de Brasília
Departamento de Estatística

Estudo sobre Evasão Feminina no Instituto de Ciências Exatas (IE) da Universidade de Brasília: uma aplicação de Regressão Logística

Ana Clara Arrais Haidar

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

Brasília
2023

Ana Clara Arrais Haidar

Estudo sobre Evasão Feminina no Instituto de Ciências Exatas (IE) da Universidade de Brasília: uma aplicação de Regressão Logística

Orientadora: Profa. Maria Teresa Leão Costa

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

Brasília
2023

Resumo

O estudo tem por objetivo identificar os principais fatores associados à evasão de estudantes do gênero feminino dos cursos de bacharelado do Instituto de Ciências Exatas da Universidade de Brasília, por meio do ajuste de modelos de regressão logística. O escopo do projeto abrange todas as mulheres ingressantes nos cursos de Ciência da Computação, Estatística e Matemática, no período de 2011 a 2019. O desejo de focar na experiência feminina parte de dois principais motivadores: primeiro, os cursos de exatas são tradicionalmente dominados por alunos do gênero masculino; segundo, a literatura aponta que a taxa de evasão das alunas mulheres nesses cursos é maior que a dos homens.

Nesse trabalho, define-se evasão como a saída do curso de origem por qualquer meio diferente da formatura. Foram levadas em consideração características acadêmicas e algumas características socioeconômicas das alunas, e os resultados finais revelaram que, nos três cursos analisados, a evasão está fortemente associada com o baixo desempenho acadêmico e com aspectos institucionais. No curso da Estatística, a idade também se mostrou um fator relevante para descrever o fenômeno de evasão, e na Ciência da Computação foi observado que alunas de escola pública têm menores chances de evasão que as alunas de escola particular.

Palavras-chaves: Gênero, Evasão, Ensino Superior, Ciências Exatas, Regressão Logística

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A participação de mulheres em ciência, tecnologia, engenharia e matemática (*science, technology,engineering and mathematics* - STEM) vem crescendo nos últimos anos, mas o domínio masculino ainda prevalece. Meninos e meninas concluem o ensino médio igualmente capacitados para seguir uma graduação em STEM na universidade, no entanto, mulheres são muito menos propensas a escolher seguir carreira nessas áreas do que os homens.

De todos os estudantes matriculados em cursos relacionados à STEM no ensino superior no mundo, apenas **35%** são mulheres. Entre as alunas de graduação do gênero feminino, só **30%** optam por cursar STEM. As matrículas são especialmente baixas em tecnologias da informação e comunicação - TIC (**3%**) e ciências naturais, matemática e estatística (**5%**) (UNESCO, 2018).

No Brasil a situação não é diferente. Segundo o Instituto Semesp (2020), **57%** dos estudantes no ensino superior brasileiro são do sexo feminino. Apesar de serem maioria nas universidades, as mulheres ocupam **48,6%** das vagas na área de ciências naturais, matemática e estatística, e apenas **13,2%** em computação e TIC. De acordo com dados da UNESCO (2018), a média mundial de participação feminina nesses campos é de **55%** e **28%**, respectivamente.

Compreender as razões da sub-representação de mulheres na educação em ciências exatas e como contorná-las é fundamental. Existem dois estereótipos prevalentes com relação ao gênero e as áreas de STEM: "os meninos são melhores em matemática e em ciências do que as meninas" e "carreiras em ciência e engenharia são domínios masculinos (Hill; Corbett; Rose, 2010).

Ao contrário do que sugerem esses esteriótipos, pesquisas sobre o desenvolvimento do cérebro, neurociência, genética e hormônios mostram que a disparidade de gênero em STEM não é resultado de fatores biológicos, ou habilidades inatas. Embora existam diferenças sexuais em certas funções fisiológicas, elas têm pouco ou nenhum impacto na habilidade acadêmica e no mecanismo neural de aprendizagem (UNESCO, 2018).

As ideias estereotipadas sobre o desempenho feminino nas ciências exatas são falsas, mas sua influência na autopercepção das mulheres é real e permanece muito forte. O Programa Internacional de Avaliação de Estudantes (*Program for International Student Assessment* - PISA) de 2015 mostrou que as meninas têm uma menor autoeficácia [1] em

[1] Autoeficácia designa em psicologia a convicção de ser capaz de realizar uma tarefa específica

ciências e matemática dos que os meninos. Em seu estudo sobre gênero e o processo de escolha de carreira, Correll (2001) confirmou que meninas avaliam suas habilidades matemáticas como sendo inferiores a de meninos, mesmo possuindo um histórico de desempenho matemático equivalente.

Essa falta de autoconfiança feminina é observada já na ensino fundamental, e tende a aumentar no ensino médio e no ensino superior. Assim, crenças sobre papéis de gênero são peças fundamentais na sub-representação feminina nas áreas de ciências exatas. Além da reduzida participação das mulheres nesses campos, também chama atenção a elevada taxa de abandono na universidade. O estudo de Ellis et al.(2016) sugere que mulheres tem **1,5** mais chance de deixar os campos de STEM do que os homens. Não apenas isso, mais de **32%** das mulheres estudantes de STEM mudam para cursos de outras áreas antes de concluir a graduação, enquanto **26%** dos homens o fazem (Chen, 2013).

As causas da alta evasão de alunas em STEM não estão limitadas à baixa autoeficácia ou a estigmas sociais, e devem ser entendidas levando-se em conta a complexidade de questões sócio-culturais, econômicas e acadêmicas. A Comissão Especial de Estudo sobre a Evasão nas Universidades Públicas Brasileiras (ANDIFES; ABRUEM; SESU/MEC, 1996) classifica os fatores que levam à evasão em três grupos: *fatores individuais*, que englobam as características pessoais do aluno, como condição socio-econômica e formação escolar; *fatores internos às instituições*, como metodologias de ensino, corpo docente e estruturas curriculares; por fim, *fatores externos*, como crises econômicas, taxa de emprego e perspectivas de remuneração.

A respeito dos cursos de STEM em particular, o estudo de Chen (2013) concluiu que o principal preditor de evasão é a baixa performance acadêmica, considerando alunos que abandonam totalmente a graduação. Esse resultado leva em conta também o baixo desempenho no ensino médio, indicando a importância da formação básica para o sucesso no ensino superior. Entre os alunos que trocam STEM por cursos em outras áreas, a pesquisadora sugere que os fatores mais determinantes para essa decisão são: cursar poucas disciplinas desafiadoras de matemática no primeiro ano de curso, pegar poucos créditos no primeiro ano de curso, e ter baixo desempenho em disciplinas de STEM em comparação à disciplinas de outros campos.

No caso específico das estudantes do gênero feminino, o sentimento de isolamento é apontado como um possível fator adicional para a alta taxa de evasão. Mulheres tendem a persistir mais nos estudos em STEM nas instituições em que o percentual de alunas nessas áreas é maior, embora não necessariamente uma elevada participação feminina no curso indique menores taxas de abandono entre mulheres (Griffith, 2010).

A Universidade de Brasília possui diversos cursos em áreas de STEM, dentre os quais os ofertados pelo Instituto de Ciências Exatas (IE). O IE engloba três departamentos - Matemática (MAT), Ciência da Computação (CIC) e Estatística (EST) - e oferece sete opções de graduação: Ciência da Computação, Licenciatura em Computação, Engenharia Mecatrônica, Engenharia da Computação, Bacharelado em Matemática, Licenciatura em Matemática e Bacharelado em Estatística. Nesses cursos é observado o mesmo fenômeno de predominância masculina discutido anteriormente, levantando o interesse de estudar a evasão das estudantes mulheres.

Assim, o presente estudo objetiva identificar e analisar os fatores associados à evasão de estudantes do gênero feminino no Instituto de Ciências Exatas (IE) da Universidade de Brasília. Serão considerados os três cursos de Bacharelado ofertados pelo instituto: Matemática, Estatística e Ciência da Computação.

2 Referencial Teórico

2.1 Regressão Logística

Modelos de regressão são amplamente utilizados para descrever a relação entre um fenômeno de interesse (variável resposta) e os fatores que podem ajudar a explicá-lo (variáveis explicativas). Em muitos casos a variável resposta é categórica, ou seja, assume valores não numéricos. Nessas situações, o modelo mais utilizado é o de regressão logística.

No modelo de regressão logística binária, a variável resposta apresenta dois resultados possíveis classificados em "sucesso"e "fracasso", respectivamente codificadas como 1 e 0 e com probabilidades de ocorrência π e $1-\pi$. A variável resposta Y segue distribuição Bernoulli com valor esperado $E(Y) = \pi$, $0 \leq \pi \leq 1$, sendo

$$\pi = \frac{exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p}, \tag{2.1.1}$$

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

$$\pi(\boldsymbol{X}) = P(Y = 1|X_1, \ldots, X_p). \tag{2.1.2}$$

As variáveis explicativas podem ser quantitativas ou qualitativas. Para a inserção no modelo, variáveis qualitativas precisam ser transformadas em variáveis *dummy*, também denominadas variáveis indicadoras, que assumem valores 1 e 0 para indicar a presença ou ausência de determinado efeito categórico. Uma variável independente qualitativa tendo l níveis requer a criação de $l-1$ variáveis *dummy*.

2.1.1 Estimação dos parâmetros

A distribuição de probabilidade de cada observação Y_i é dada por

$$f_i(Y_i) = \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \qquad Y_i = 0, 1; \quad i = 1, \ldots, n. \tag{2.1.3}$$

Uma vez que as Y_i observações são independentes, segue que a função de proba-

bilidade conjunta é

$$g(Y_1, \ldots, Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i}. \tag{2.1.4}$$

A partir do logaritmo da função de probabilidade conjunta, é possível calcular a função de máxima verossimilhança:

$$\begin{aligned} ln\, g(Y_1, \ldots, Y_n) &= ln \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_i)^{1-Y_i} \\ &= \sum_{i=1}^{n} Y_i\, ln\left(\frac{\pi_i}{1-\pi_i}\right) + \sum_{i=1}^{n} ln(1-\pi_i). \end{aligned} \tag{2.1.5}$$

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

$$ln\, L(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i) - \sum_{i=1}^{n} ln[1 - exp(\beta_0 + \beta_1 X_i + \ldots + \beta_p X_i)]. \tag{2.1.6}$$

Os parâmetros do modelo são estimados a partir da função de máxima verossimilhança. Contudo, não existe solução fechada para os valores de $\beta_0, \ldots, \beta_p$ que maximize a função de verossimilhança, de modo que as estimativas são obtidas por meio de métodos numéricos iterativos, como *Newton-Raphson* ou *Score*. Usualmente, os cálculos requerem auxílio de programas computacionais.

Uma vez obtidas as estimativas de $\beta_0, \ldots, \beta_p$, substituem-se os valores em (2.1.1) para encontrar a função de resposta ajustada:

$$\hat{\pi}_i = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_i + \ldots + \hat{\beta}_p X_i)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_i + \ldots + \hat{\beta}_p X_i)}. \tag{2.1.7}$$

2.1.2 Interpretação dos parâmetros

A interpretação dos parâmetros de modelos de regressão logística não é intuitiva como em modelos de regressão linear. Costuma-se aplicar a transformação logito no modelo para obter o preditor linear:

$$logito(\pi) = ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 X_1 + \ldots + \beta_p X_p. \tag{2.1.8}$$

A transformação logito é uma função da chance de sucesso, ou *odds*, uma medida que quantifica a razão entre as probabilidades de sucesso (π) e fracasso ($\pi - 1$)

$$odds = \frac{\pi}{1-\pi}. \tag{2.1.9}$$

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 o $ln(odds)$, mantendo as outras variáveis explicativas constantes. Assim, considere x_j e $x_j + 1$ dois valores distintos de uma variável explicativa X_j, com chances de sucesso $odds_1$ e $odds_2$, respectivamente. Tem-se que

$$ln(odds_1) - ln(odds_2) = ln\left(\frac{odds_2}{odds_1}\right) = \beta_j. \tag{2.1.10}$$

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

$$\theta = \frac{odds_2}{odds_1} = e^{\beta_j}. \tag{2.1.11}$$

A *odds ratio* indica 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 $l-1$ variáveis *dummy* de uma variável qualitativa com l níveis, então $x_j = 0$ e $x_j + 1 = 1$ e, portanto, e^{β_j} quantifica o efeito multiplicativo na *odds* dada a presença do efeito categórico em questão.

2.1.3 Intervalo de Confiança dos parâmetros

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

$$\beta_k = \hat{\beta}_k \pm z_{1-\alpha/2}\, s\{\hat{\beta}_k\}, \tag{2.1.12}$$

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

$$\theta_k = exp(\hat{\beta}_k \pm z_{1-\alpha/2}\, s\{\hat{\beta}_k\}). \tag{2.1.13}$$

2.2 Testes de Significância

Uma vez ajustado o modelo inicial, 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.

2.2.1 Teste de Razão de Verossimilhança

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

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

A estatística de teste é dada por

$$G^2 = ln \left[\frac{L(R)}{L(F)} \right] = -2[ln\, L(R) - ln\, L(F)], \tag{2.2.1}$$

em que *L(R)* e *L(F)* são os valores da função de máxima 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 parâmetros 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:

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

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

2.2.2 Teste de Wald

O teste de Wald verifica individualmente a significância de cada parâmetro do modelo. As hipóteses são:

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

A estatística de teste z^* é dada por

$$z^* = \frac{\hat{\beta}_k}{s\{\hat{\beta}_k\}} \quad ou \quad (z^*)^2 = \frac{\hat{\beta}_k^2}{s\{\hat{\beta}_k\}^2}. \tag{2.2.2}$$

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

2.3 Seleção do Modelo

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

2.3.1 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 verdadeiras valores médios observados. Valores pequenos indicam melhor ajuste do modelo, contudo o modelo selecionado não necessariamente deve ser aquele com menores *AIC* e *BIC*. As medidas são calculadas a partir das expressões

$$AIC = -2ln\, L(\beta) + 2p, \tag{2.3.1}$$

$$BIC = -2ln\, L(\beta) + p\, ln(n), \tag{2.3.2}$$

em que $L(\beta)$ é a log-verossimilhança definida em (2.1.6) e p é o número de parâmetros do modelo.

2.3.2 Métodos Automáticos

Quando o total de potenciais variáveis explicativas é muito grande, torna-se inviável comparar cada um dos modelos possíveis de forma manual. Nesses casos, utilizam-se métodos iterativos computacionais para selecionar os modelos mais adequados. Os métodos automáticos mais utilizados em regressão são o *Stepwise* e suas variantes.

Stepwise: a primeira variável explicativa adicionada ao modelo é aquela mais correlacionada com a variável resposta. As variáveis seguintes são incluídas ou não de acordo com o ganho na função de log-verossimilhança, até que a adição de novas variáveis não seja mais significativa no modelo com base no nível de significância adotado. A cada etapa da seleção, verifica-se se existem variáveis no modelo que podem ser retiradas antes de seguir para a próxima etapa. O procedimento termina quando não há mais variáveis para serem adicionadas ou retiradas do modelo.

Forward: é uma simplificação do método *Stepwise*, em que não há a verificação se existem variáveis que devem ser retiradas do modelo. Variáveis já adicionadas não são excluídas e permanecem no modelo final selecionado.

Backward: o procedimento de eliminação *Backward* é o oposto do *Forward*, na medida em que inicia o modelo com todas as variáveis explicativas possíveis e elimina uma a uma com base na menor perda na função de log-verossimilhança.

2.4 Avaliação do Modelo

Uma vez selecionado o modelo, é necessário verificar a qualidade do ajustamento aos dados e seu poder preditivo.

2.4.1 Adequabilidade do Ajustamento

A qualidade do ajuste do modelo é medida a partir de testes de adequabilidade. Caso se verifique que o modelo não está bem ajustado, então ele 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 número esperado ou ajustado de sucessos. Considerando $\pi(x)$ definido em (2.1.1), 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:

$$\chi^2 = \sum_{j=1}^{c} \sum_{k=0}^{1} \frac{(o_{jk} - e_{jk})^2}{e_{jk}}. \tag{2.4.1}$$

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. O teste requer que as frequências esperadas sejam maiores ou iguais a 5, e nunca inferiores a 1.

Teste *Deviance*

O teste *Deviance* é similar ao teste de Razão de Verossimilhança, porém as hipóteses consideradas são diferentes:

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

A estatística de teste é dada por

$$G^2 = ln \left[\frac{L(R)}{L(F)} \right] = -2[ln\, L(R) - ln\, L(F)], \tag{2.4.2}$$

em que *L(R)* e *L(F)* são os valores da função de máxima verossimilhança para os modelos reduzido e completo, respectivamente. O modelo reduzido é obtido sob H_0.

Assim como no teste χ^2 de Pearson, a estatística de teste $G^2 \sim \chi^2_{c-p}$ sob H_0, em que $c - p$ é a diferença entre o número de conjuntos de valores distintos de variáveis e o total de parâmetros no modelo, respectivamente.

Teste de Hosmer e Lemeshow

Hosmer e Lemeshow (2000) propuseram um teste alternativo de adequabilidade de ajuste para modelos logísticos. Nesse teste, a variável resposta é agrupada com base nos valores estimados de probabilidade. As 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 probabilidade 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, $k = 1, \ldots, g-1$, e os grupos contêm todos as observações com probabilidades estimadas entre os pontos de corte adjacentes.

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

$$\hat{C} = \sum_{j=1}^{g} \sum_{k=0}^{1} \frac{(o_{jk} - e_{jk})^2}{e_{jk}}. \tag{2.4.3}$$

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

2.4.2 Análise de Resíduos

Os resíduos de um modelo indicam a distância entre os valores observados e os valores esperados. Modelos adequadamente ajustados aos dados possuem valores residuais pequenos. Assim, para avaliar a qualidade do ajuste de um modelo, é preciso fazer uma análise diagnóstica dos resíduos.

Resíduos de Pearson

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

$$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1 - \hat{\pi}_i)}}. \tag{2.4.4}$$

Resíduos Studentizados de Pearson

Resíduos studentizados são uma padronização dos resíduos de Pearson, e podem ser calculados por:

$$r_{SPi} = \frac{r_i}{\sqrt{1 - h_{ii}}}. \tag{2.4.5}$$

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

$$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}, \tag{2.4.6}$$

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:

$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\sum_{i=1}^{n}[Y_i log(\hat{\pi}_i) + (1 - Y_i)log(1 - \hat{\pi}_i)]}. \tag{2.4.7}$$

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 ao modelo *Deviance*:

$$\sum_{i=1}^{n} dev_i^2 = DEV(X_0, \ldots, X_{p-1}) = -2\sum_{i=1}^{n}[Y_i log(\hat{\pi}_i) + (1 - Y_i)log(1 - \hat{\pi}_i)]. \tag{2.4.8}$$

Uma vez calculados, os resíduos podem ser representados em gráficos que permitem avaliar possíveis afastamentos dos pressupostos do modelo.

2.4.3 Matriz de confusão e desempenho

Ao avaliar o ajuste de um modelo, é interessante verificar seu poder preditivo. A matriz de confusão e desempenho é uma tabela de classificação cruzada da variável resposta binária Y e o resultado predito $\hat{Y}$. Em regressão logística, a predição de uma observação i é $\hat{y}_i = 1$ (sucesso) quando a probabilidade estimada $\hat{\pi}_i > \pi_0$ e $\hat{y}_i = 0$ (fracasso) quando $\hat{\pi}_i \leq \pi_0$, para determinada probabilidade de corte π_0. Normalmente adota-se $\pi_0 = 0,5$. 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 $\hat{Y} = 0$.

Assim, 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.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \tag{2.4.9}$$

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

$$\text{Sensibilidade} = \frac{VP}{VP + FN} \tag{2.4.10}$$

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

$$\text{Especificidade} = \frac{VN}{VN + FP} \tag{2.4.11}$$

2.4.4 Curva ROC

A curva ROC, ou curva de Característica de Operação do Receptor *(Receiver Operating Characteristic)* é um gráfico da sensibilidade *vs.* 1 - especifidade 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 entre sucesso e 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.

2.5 Associação entre variáveis

Medir a associação entre duas variáveis é importante para entender o efeito que uma variável tem sobre a outra. Em uma modelagem, por exemplo, é de interesse saber o relação entre a variável resposta e as variáveis explicativas, bem como as interações entre as mesmas.

2.5.1 Coeficiente de Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida utilizada para verificar a intensidade da relação linear entre duas variáveis quantitativas. Tem-se que

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{2.5.1}$$

em que $-1 \leq r_{xy} \leq 1$. Se $r_{xy} > 0$, diz-se que as variáveis são diretamente proporcionais, e se $r_{xy} < 0$ as variáveis são inversamente proporcionais.

2.5.2 Teste χ^2 de Independência

O teste χ^2 de independência é utilizado para medir se existe ou não associação entre duas variáveis qualitativas X e Y. As hipóteses são:

$$\begin{cases} H_0 : \pi_{ij} = \pi_{i+}\pi_{+j} \\ H_1 : \pi_{ij} \neq \pi_{i+}\pi_{+j} \end{cases}$$

A estatística de teste é dada por:

$$\chi^2 = \sum_{j=1}^{c}\sum_{k=0}^{1}\frac{(o_{jk} - e_{jk})^2}{e_{jk}}, \tag{2.5.2}$$

em que o_{jk} é a frequência observada e e_{jk} é a frequência esperada. Sob H_0, tem-se que $\chi^2 \sim \chi^2_{(r-1)(s-1)}$, sendo r o total de categorias em X e s o total de categorias em Y. O teste requer que as frequências esperadas sejam maiores ou iguais a 5, e nunca inferiores a 1.

2.5.3 Teste Exato de Fisher

É um teste alternativo ao teste de χ^2 de independência, usado para tamanhos pequenos de amostra ou tabelas de contingência com frequências menores que 5. O teste é usualmente aplicado em tabelas 2×2, mas pode ser realizado em tabelas maiores com auxílio computacional.

Tabela 2: Tabela 2×2 do Teste Exato de Fisher

X	Y		Total
	Sucesso	Insucesso	
1	a	b	a + b
2	c	d	c + d
Total	a + c	b + d	n

As hipóteses de teste são definidas por:

$$\begin{cases} H_0 : \pi_1 = \pi_2 \\ H_1 : \pi_1 \neq \pi_2 \end{cases}$$

A estatística de teste T equivale à frequência a, ou o número de sucessos no nível 1 de X, e segue distribuição hipergeométrica de acordo com a equação abaixo:

$$P(T = a) = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}}. \tag{2.5.3}$$

3 Metodologia

3.1 Banco de dados

Nesse estudo foram utilizados os dados disponibilizados pela UnB via Sistema de Informações Acadêmicas de Graduação (SIGRA) e Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA). São ao todo três bancos - um para cada curso de bacharelado do Instituto de Ciências Exatas (IE) -, que contêm informações sociodemográficas e acadêmicas dos estudantes, totalizando 24 variáveis em comum.

Os bancos dos cursos de bacharelado em Estatística e em Matemática originalmente continham dez variáveis a mais que o banco do Ciência da Computação. Dessas, apenas duas demonstravam relevância para o estudo, "Raça"e "Modalidade da Disciplina", e as demais foram retiradas. Ambas as variáveis foram criadas no banco da Ciência da Computação e preenchidas com valores vazios. Assim, cada banco de dados passou a conter 26 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;

O banco do curso de Ciência da Computação possui 123.126 linhas, com dados desde 1991 a 2019; o banco de Estatística possui 10.996 linhas, com dados desde 1998 a 2019; e o de Matemática, 3.284 linhas com dados desde 1993 a 2019.

Com o objetivo de ter uma visão atual desses cursos, foi delimitado como escopo de estudo as estudantes do gênero feminino que ingressaram no curso entre 2011 e 2019. O corte no ano de 2011 se deve à limitação imposta pelas estruturas curriculares dos cursos. O curso de Ciência da Computação possui 8 currículos[2] disponíveis no acesso público do

[2]Currículos do Bacharelado em Ciência da Computação da UnB, disponíveis aqui.

SIGAA da UnB, datados de 2011 a 2021. O curso de Estatística possui 5 currículos[3], datados de 1997 a 2021. Por fim, o curso de Matemática possui 3 currículos[4], datados de 2011 a 2016. Uma vez que informações curriculares anteriores a 2011 só estão disponíveis para o curso de Estatística, esse foi o ano escolhido para o recorte temporal.

Ainda em relação às bases de dados, há uma inconsistência que deve ser pontuada: nos bancos da Estatística e da Matemática existem muitas informações faltantes a respeito das alunas ativas, isto é, alunas que até o semestre de 2019/2 ainda não haviam deixado o curso. No caso particular dessas alunas, verificou-se que as variáveis "escola", "chamada de ingresso na UnB", "período de ingresso na UnB", "período de ingresso no curso", "forma de ingresso no curso"e "créditos aprovados no período"estão vazias em sua totalidade. O mesmo não ocorre no banco da Ciência da Computação.

Para contornar esse problema, os períodos de ingresso das alunas ativas da Estatística e da Matemática foram inferidos a partir dos registros mais antigos da variável "período que cursou disciplina"de cada aluna. As três bases de dados continham, ainda, alunas duplicadas que precisaram ser excluídas.

Assim, as bases de dados finais contém informações das estudantes do gênero feminino do cursos de bacharelado do IE que ingressaram entre 2011 e 2019, totalizando 88 alunas da Ciência da Computação, 136 alunas da Estatística e 45 alunas da Matemática.

3.2 Criação e manipulação de variáveis

A partir dos dados originais, foram criadas novas variáveis de interesse para o estudo. Algumas variáveis já existentes foram agrupadas de forma diferente, segundo a necessidade de padronização de categorias.

Raça

Variável presente nos bancos originais, porém foi alterada. Segundo o Instituto Brasileiro de Geografia e Estatística (IBGE), a raça negra é definida como o conjunto de pessoas que se autodeclaram pretas e pardas. Assim, as raças "preta", "parda"e "negra"foram agrupadas em uma única categoria, "negra".

[3]Currículos do Bacharelado em Estatística da UnB, disponíveis aqui.
[4]Currículos do Bacharelado em Matemática da UnB, disponíveis aqui.

Forma de Ingresso no Curso

Variável presente nos bancos originais, mas passou por modificações. As categorias "Dupla Habilitação", "Dupla Diplomação"e "Duplo Curso"foram todas agrupadas em "Duplo Curso"; as categorias "Mudança de Habilitação"e "Mudança de curso"foram agrupadas em "Mudança de curso"; por fim, "Transferência Facultativa"e "Transferência Obrigatória"foram agrupadas em "Transferência".

Forma de saída do curso

Variável presente nos bancos originais, porém foi alterada. As categorias "Desligamento - não cumpriu condição"e "Reprovou 3 vezes na mesma disciplina obrigatória"refletem a mesma situação - o desligamento da aluna devido ao seu rendimento acadêmico - e, portanto, foram agrupadas em "Desligamento - Rendimento".

Evasão

Indica se a aluna evadiu ou não o curso. Foi criada a partir da variável "forma de saída do curso", e recebeu dois valores: "sim"e "não". O agrupamento foi feito de acordo com a Tabela 3 a seguir:

Tabela 3: Forma de saída do curso e evasão. Bacharelado do IE - UnB, 2011-2019

Forma de saída	**Evasão**
Ativo	Não
Formatura	
Novo vestibular	Sim
Mudança de curso	
Desligamento - Abandono	
Desligamento - Rendimento	
Desligamento - Voluntário	
Desligamento - Intercâmbio	

O conceito adotado nesse trabalho é o de evasão de curso sugerido por ANDIFES et al. (1996), que diz respeito ao aluno que se desliga do curso superior por abandono, desistência, transferência de curso ou exclusão mediante norma institucional.

Nesse sentido, alunas transferidas não foram consideradas no estudo, dado que não há informações quanto ao curso de destino dessa alunas e a transferência para outra

instituição de ensino não necessariamente significa que a aluna mudou de curso - portanto, não há como determinar se houve ou não evasão.

Evasões anteriores

Indica quantas vezes a aluna evadiu o curso anteriormente. Foi criada com base na análise do histórico acadêmico das alunas, e nas variáveis período de ingresso no curso, período de saída do curso e evasão. Foram identificadas as alunas que já possuíam alguma passagem no curso anterior ao período de ingresso mais recente, depois contabilizadas os ocorrências de evasão.

Idade ao ingressar

Idade da aluna ao ingressar no curso, expressa em anos completos. Foi criada a partir das variáveis data de nascimento e período de ingresso do curso. Considerou-se o dia 1º de março como referência para as alunas que entraram no primeiro semestre do ano, e dia 1º de agosto para as que entraram no segundo semestre. A idade foi calculada pela diferença em anos entre as datas de nascimento e de ingresso no curso.

Semestres cursados

Quantidade de semestres cursados desde o ingresso até a saída do curso. Foi criada a partir da diferença entre as variáveis período de ingresso do curso e período de saída do curso. No caso das alunas ainda ativas no curso, considerou-se o período de saída 2019/2 para realizar o cálculo.

Trancamentos

Quantidade de trancamentos de disciplina realizados ao longo da permanência no curso. Com base na variável menção na disciplina, somaram-se todas as vezes que uma aluna recebeu as menções TR ou TJ (Trancamento ou Trancamento Justificado) em uma mesma disciplina ou disciplinas diferentes.

Menções SR

Quantidade de menções SR recebidas ao longo da permanência no curso. Com base na variável menção na disciplina, somaram-se todas as vezes que uma aluna recebeu menção SR (Sem Rendimento) em uma mesma disciplina ou disciplinas diferentes.

A menção SR é dada quando a aluna não atinge a frequência mínima na disciplina. Nesse sentido, pode ser um bom indicativo de abandono da matéria - prática comum entre alunas que evadem o curso.

Taxa de reprovação

A taxa de reprovação indica a proporção entre o total de créditos reprovados e o total de créditos cursados pela aluna ao longo de sua permanência no curso (PINTO, 2022). A taxa varia de 0 a 1 e foi criada a partir da variáveis menção na disciplina e créditos da disciplina.

$$\text{Taxa de reprovação} = \frac{\text{Total de créditos com reprovação}}{\text{Total de créditos cursados}} \tag{3.2.1}$$

Currículo

Qual currículo estava vigente ao longo da permanência no curso: "antigo"ou "novo". A variável foi criada com base na análise das estruturas curriculares dos cursos de Bacharelado do IE e dos históricos acadêmicos das alunas.

Para o curso de Ciência da Computação, currículos vigentes a partir do semestre 2015/2 foram considerados como novos, e os anteriores como antigos. Para o curso de Estatística, a divisão foi feita no semestre 2014/1. Já na Matemática, não foram identificadas mudanças significativas na estrutura disciplinar no curso, de modo que todos os currículos a partir de 2011 foram considerados como novos.

Os bancos de dados originais não contêm a informação da estrutura curricular, portanto essa informação foi inferida a partir das variáveis período de ingresso no curso, período de saída do curso e mínimo de créditos para a formatura. Essa última varia de acordo com o currículo em vigor.

Reprovou Cálculo 1

Indica se a aluna reprovou alguma vez a disciplina Cálculo 1. Foi criada a partir da variável menção na disciplina e assume dois valores: "sim"e "não". Pretende-se verificar se a reprovação em Cálculo 1 - matéria obrigatória do 1º semestre nos três cursos - é um indicativo que a base matemática da aluna esta abaixo do nível exigido no curso e, portanto, que enfrentará maiores dificuldades para se formar.

No banco do curso de Matemática algumas alunas não possuem o registro da disciplina Cálculo 1 no histórico acadêmico, de tal forma que não é possível saber se houve ou não reprovação. Esses casos em particular foram agrupados em "sem informação".

Reprovou matéria obrigatória do 1º semestre

Indica quantas vezes a alunas reprovou alguma disciplina obrigatória do 1º semestre de curso. Foi criada a partir das variáveis menção na disciplina e nome da disciplina. É uma extensão da variável reprovou cálculo 1 ao considerar as outras matérias obrigatórias exigidas no 1º semestre de cada curso, e não apenas Cálculo 1. O interesse nessa variável é poder avaliar se um baixo desempenho nas matérias introdutórias do curso resulta em um maior risco de evasão.

Tabela 4: Matérias obrigatórias do 1º semestre. Bacharelado do IE - UnB, 2011-2019

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

Cursou Verão

Indica se a aluna cursou alguma vez uma disciplina em semestre de verão. É uma variável dicotômica e assume dois valores: "sim"e "não". Utilizou-se a variável "período que cursou a disciplina"para identificar se a disciplina foi cursada em período de verão ou não.

Local por renda

Nível de renda médio da Região Administrativa (RA) de residência. Foi utilizada a variável CEP para identificar as RAs de residência e, a partir dessa informação, definir os grupos de renda. O agrupamento foi feito de acordo com a Tabela 5:

Tabela 5: Grupo de renda das RAs de residência. Bacharelado do IE - UnB, 2011-2019

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/Pôr do Sol, Estrutural/SCIA

Fonte: PDAD 2021 - Codeplan[5]

Foi utilizado o pacote `BeautifulSoup` do Python para realizar buscas por CEP em um site[6] e extrair os bairros e logradouros correspondentes. A partir da informação do bairro, foi possível determinar as RAs de residência e os grupos de renda.

Os bancos contêm ainda localidades de fora do Distrito Federal: Luziânia (GO), Novo Gama (GO), Valparaíso (GO), Cidade Ocidental (GO), Formosa (GO), Uberlândia (MG), Barreiras (BA), Recife (BA) e Rio de Janeiro (RJ). Essa localidades e CEPs inválidos foram classificadas como "Sem informação"de renda.

Distância de deslocamento

Distância de deslocamento em quilômetros do local de residência até a Universidade de Brasília, criada com base no proposto por Côrtes (2023). A distância foi calculada a partir do CEP, com auxílio da função `mapdist()` do pacote `ggmap` do R, desenvolvido por Kahle e Wickham (2013).

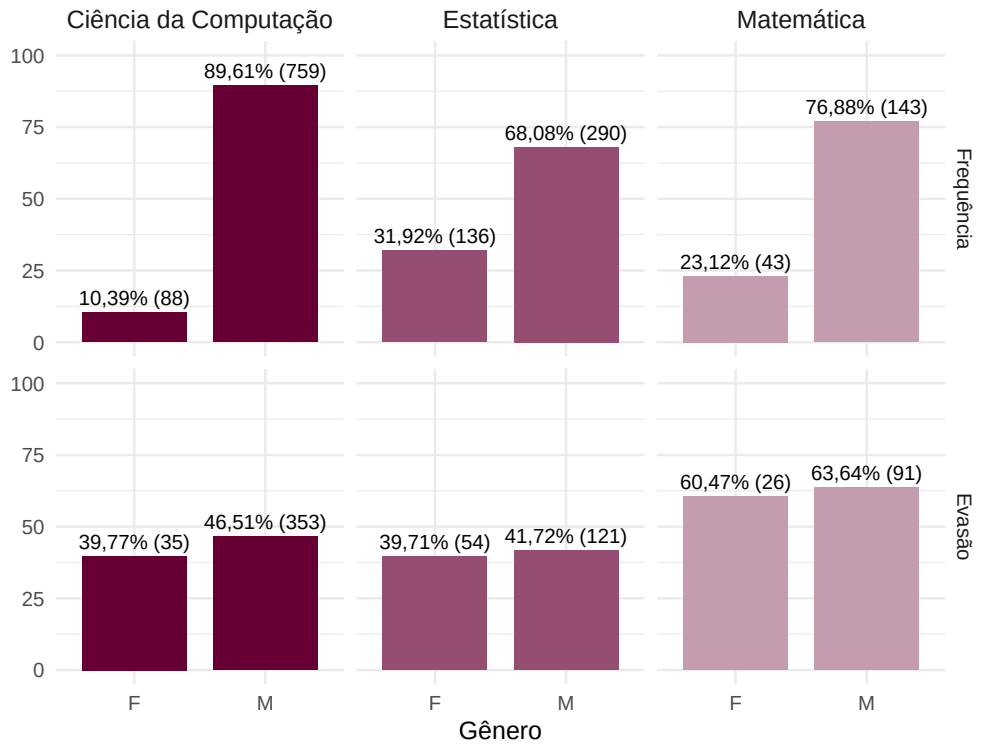
A função permite determinar um local de partida - os endereços de cada aluna, com os bairros e o logradouros extraídos a partir dos CEPs -, e um local de chegada - a Universidade de Brasília. A distância calculada considera o trajeto percorrido por carro.

[5]PDAD 2021 - Codeplan, disponível aqui.
[6]Endereço eletrônico disponível aqui.

4 Resultados

A parte inicial desse estudo será focada em avaliar as frequências e as taxas de evasão dos homens e mulheres dos cursos de bacharelado do IE.

Gráfico 1: Distribuição dos estudantes e % de evasão segundo gênero.
Bacharelados do IE - UnB, 2011-2019



Com base no Gráfico 1, fica evidente a predominância masculina nos cursos de bacharelado do Instituto de Exatas da UnB. A maior participação feminina é observada no curso de Estatística (31,92%), seguida da curso de Matemática (23,12%) e Ciência da Computação (10,39%), com o menor percentual de estudantes mulheres entre os três cursos.

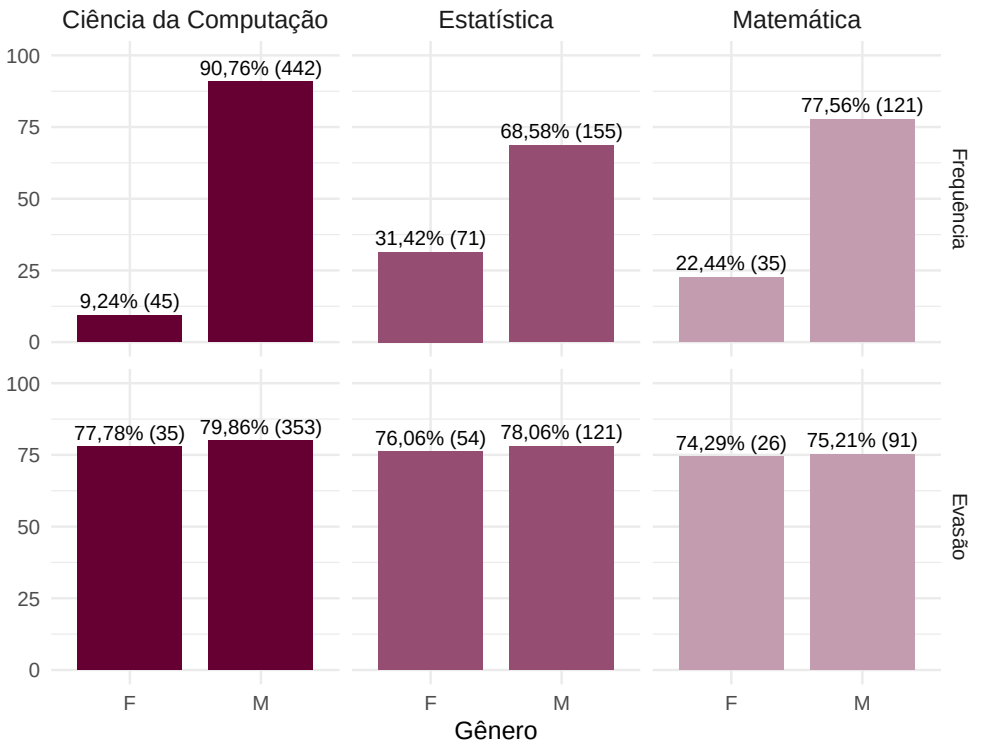
As taxas de evasão são bastante similares entre os gêneros, desafiando a noção de que mulheres evadem mais áreas de STEM do que os homens. Em todos os três cursos, a evasão de estudantes do gênero feminino foi levemente inferior a dos estudantes de gênero masculino, a maior diferença sendo de 6,74 pontos percentuais no curso de Ciência da Computação.

No período de 2011 a 2019, mulheres evadiram os cursos de Ciência da Computação e Estatística em taxas muito parecidas - 39,77% e 39,71%, respectivamente. A Matemática destoa dos outros dois cursos, com taxa de evasão de 60,47%. Esse resultado

pode ser explicado em parte pela baixa proporção de alunas ativas na Matemática em comparação aos outros cursos.

Para ter uma noção do impacto dos alunos ativos nesses resultados, foi feita uma análise apenas com os alunos não ativos, isto é, aqueles que até o fim do estudo haviam evadido ou se formado.

Gráfico 2: Distribuição dos estudantes não ativos e % de evasão segundo gênero.
Bacharelados do IE - UnB, 2011-2019

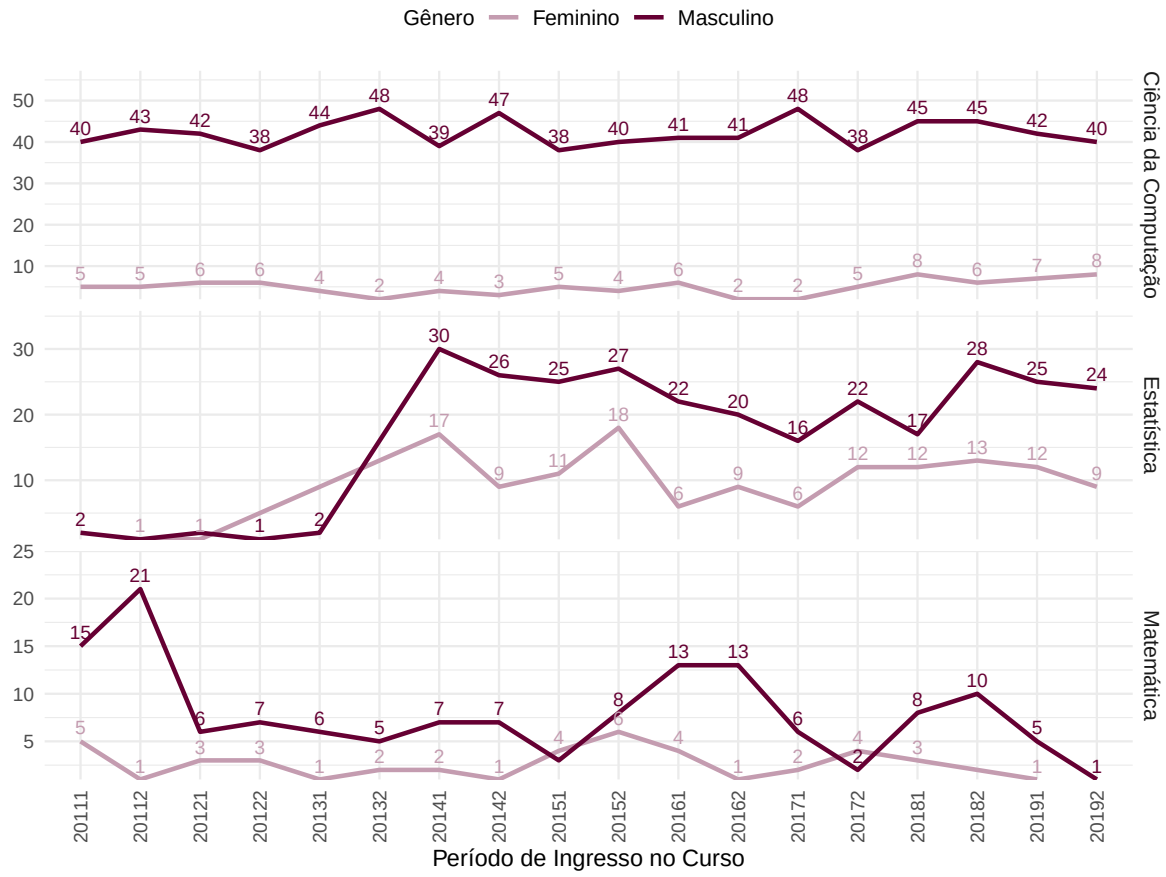


Analisando somente os alunos não ativos, nota-se que as proporções de homens e mulheres nos cursos se mantêm aproximadamente as mesmas observadas anteriormente. As taxas de evasão, em contrapartida, são bem diferentes.

Quando são desconsiderados os alunos ativos, as proporções de evasão dos cursos de Ciência da Computação e Estatística quase dobram em relação ao resultado anterior. As taxas nos três cursos variam de 74,29% a 79,86%, índices elevadíssimos de evasão. Nota-se também que as taxas são muito similares entre os cursos e os gêneros.

Dando continuidade à análise de todos os estudantes, isto é, ativos e não ativos, o gráfico a seguir mostra o número de ingressos por ano e gênero em cada curso:

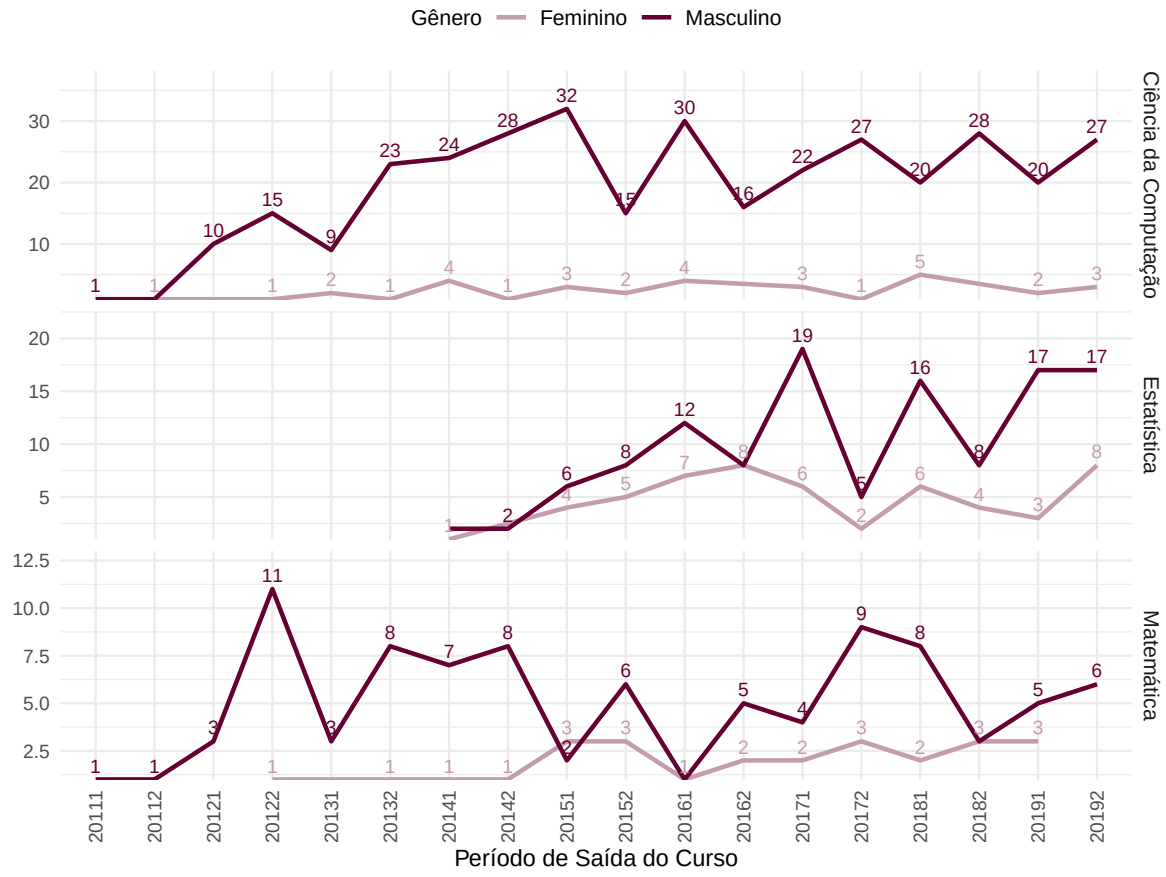
Gráfico 3: Linha do tempo do período de ingresso no curso, por gênero e curso.
Bacharelados do IE - UnB, 2011-2019



O número de ingressos no curso de Ciência da Computação segue um padrão bem linear ao longo do tempo, como pode ser verificado no Gráfico 3. Esse comportamento é diferente nos cursos da Estatística e Matemática, em que se observam maiores variações na quantidade de ingressantes por ano ao longo do período analisado. Especificamente no curso de Estatística, há uma diferença brusca de ingressos entre os períodos anteriores ao semestre 2014/1 e os seguintes. Quase nenhum aluno, homem ou mulher, entrou no curso entre 2011/1 e 2013/2. Esse resultado provavelmente reflete algum problema no banco de dados da Estatística, e merece ser investigado em outras oportunidades.

Os semestres com o maior número de ingressos de mulheres no cursos foram: 2018/1 e 2019/2 na Ciência da Computação, ambos com 8 ocorrências; 2015/2 na Estatística, com 18 ocorrências; 2015/2 na Matemática, com 6 ocorrências.

Gráfico 4: Linha do tempo do período de saída do curso de alunos evadidos, por gênero e curso. Bacharelados do IE - UnB, 2011-2019



O Gráfico 4 mostra a quantidade de alunos que evadiram os cursos por semestre, durante o período de estudo. Considerando que os alunos analisados entraram nos cursos a partir de 2011/1, espera-se que nos semestres iniciais existam poucos casos de evasão. De fato, no primeiro ano de curso dos primeiros ingressantes (período de 2011/1 até 2012/1), poucos alunos evadiram em comparação ao restante da série. Na Matemática esse cenário muda já em 2012/2, no semestre seguinte, quando há um pico de evasões no curso.

O número de alunos homens que evadiram o curso de Ciência da Computação sobe bastante a partir de 2013/2 a mantém-se elevado, com algumas oscilações, até o final do período. Já as mulheres evadiram o curso em quantidades bem semelhantes entre os anos analisados. Chama atenção o semestre 2014/1, em que ingressaram tantas mulheres quanto evadiram (4). O semestre com mais evasões foi 2018/1 (5), quando também entrou o maior número de alunas (8).

Na Estatística não houveram evasões antes do semestre 2014/1, o que não surpreende dada a pequena quantidade de alunos que entraram no curso nesse período. O

semestre de 2016/2 foi quando a maior quantidade de alunas evadiu do curso (8), e o semestre seguinte teve a maior saída de estudantes homens (19).

4.1 Análise Descritiva

Previamente à modelagem, é importante analisar as características sociais e acadêmicas das alunas do IE, bem como suas relações com a evasão de curso.

4.1.1 Variáveis Explicativas

Essa seção dedica-se a traçar o perfil das alunas dos curso de Ciência da Computação, Estatística e Matemática, com base nas variáveis explicativas de interesse.

Tabela 6: Distribuição das estudantes segundo características pessoais e sistema de cotas. Bacharelados do IE - UnB, 2011-2019

	Ciência da Computação	**Estatística**	**Matemática**
Raça			
Amarela	-	6,62% (9)	-
Branca	-	30,88% (42)	44,19% (19)
Negra	-	29,41% (40)	37,21% (16)
Sem informação	100% (88)	33,09% (45)	18,6% (8)
Local por renda			
Renda Alta	55,68% (49)	38,97% (53)	44,19% (19)
Renda Média-Alta	18,18% (16)	24,26% (33)	27,91% (12)
Renda Média-Baixa	12,5% (11)	19,12% (26)	11,63% (5)
Renda Baixa	5,68% (5)	9,56% (13)	6,98% (3)
Sem informação	7,95% (7)	8,09% (11)	9,3% (4)
Escola			
Particular	65,91% (58)	26,47% (36)	51,16% (22)
Pública	34,09% (30)	24,26% (33)	27,91% (12)
Sem informação	-	49,26% (67)	20,93% (9)
Sistema de cotas			
Não	67,05% (59)	58,09% (79)	88,37% (38)
Sim	32,95% (29)	41,91% (57)	11,63% (5)
Total	100% (88)	100% (136)	100% (43)

- indica frequência nula.

O banco da Ciência da Computação não fornece qualquer informação a respeito da raça das alunas. Nos cursos da Estatística e Matemática, respectivamente 33,09% e 18,6% da alunas não possuem registro de raça. As estudantes com identificação de raça são majoritariamente brancas ou negras. A Estatística é a única graduação com alunas amarelas, porém em pequena proporção - 6,62%.

Alunas residentes em locais de renda alta são as mais frequentes nos três cursos, seguidas das alunas residentes em locais de renda média-alta, renda média-baixa e, desconsiderando os dados sem informação, renda baixa.

Aproximadamente dois terços das alunas da Ciência da Computação, 65,91%, completaram o segundo grau em escola particular. Na Matemática também é observada alta frequência de alunas oriundas de ensino particular, 51,16%, ao passo que apenas 27,91% vieram de ensino público. Muitas alunas da Estatística não têm registro do tipo de escola de origem; das que possuem, quase metade é de escola particular e a outra metade de escola pública.

Nos três cursos a maioria da alunas não é cotista. A maior parcela de alunas com cota é observada na curso da Estatística, representando 41,91%, e a menor vem do curso da Matemática, totalizando apenas 11,63%. A Tabela 7 a seguir mostra a distribuição dos tipos de cotas usufruídas pelas alunas cotistas:

Tabela 7: Sistema de cotas utilizado por alunas cotistas. Bacharelados do IE - UnB, 2011-2019

	Ciência da Computação	**Estatística**	**Matemática**
Cota			
Egresso de Escola Pública	-	33,33% (19)	20% (1)
Escola Pública Alta Renda - Não PPI	-	26,32% (15)	20% (1)
Escola Pública Alta Renda - PPI	-	26,32% (15)	-
Escola Pública Baixa Renda - Não PPI	-	1,75% (1)	-
Escola Pública Baixa Renda - PPI	-	5,26% (3)	-
Negro	-	7,2% (4)	60% (3)
Sem informação	100% (29)	-	-
Total	100% (29)	100% (57)	100% (5)

- indica frequência nula.

Não há informações disponíveis quanto ao tipo de cota utilizado pelas alunas da Ciência da Computação. Na Estatística, cotas de escola pública somam 92,8% do total; dessas, a maior parte é de escolas públicas de alta renda. Apenas 7,2% das cotas são destinadas às alunas negras. Já na Matemática o oposto é observado: a maioria das cotas é de negras, 60%, e o restante destina-se a estudantes de escola pública.

Tabela 8: Distribuição das estudantes segundo características acadêmicas. Bacharelados do IE - UnB, 2011-2019

	Ciência da Computação	**Estatística**	**Matemática**
Forma de ingresso			
PAS	38,64% (34)	19,12% (26)	2,33% (1)
SISU/ENEM	6,82% (6)	9,56% (13)	4,65% (2)
Vestibular	40,91% (36)	17,65% (24)	18,6% (8)
Mudança de curso	-	2,94% (4)	18,6% (8)
Duplo curso	-	-	32,56% (14)
Transferência	5,68% (5)	-	-
Convênio	2,27% (2)	-	-
Portadora de diploma	5,68% (5)	2,94% (4)	4,65% (2)
Sem informação	-	47,79% (65)	18,6% (8)
Forma de saída			
Ativo	48,86% (43)	47,79% (65)	18,6% (8)
Formatura	11,36% (10)	12,5% (17)	20,93% (9)
Novo vestibular	12,5% (11)	6,62% (9)	6,98% (3)
Mudança de curso	1,14% (1)	-	-
Desligamento - Abandono	9,09% (8)	11,03% (15)	32,56% (14)
Desligamento - Rendimento	10,23% (9)	18,38% (25)	11,63% (5)
Desligamento - Voluntário	4,55% (4)	3,68% (5)	9,3% (4)
Desligamento - Intercâmbio	2,27% (2)	-	-
Cursou Verão			
Não	80,68% (71)	86,03% (117)	83,72% (36)
Sim	19,32% (17)	13,97% (19)	16,28% (7)
Reprovou Cálculo 1			
Não	60,23% (53)	55,15% (75)	34,88% (15)
Sim	39,77% (35)	44,85% (61)	6,98% (3)
Sem informação	-	-	58,14% (25)
Currículo			
Novo	64,77% (57)	100% (136)	100% (43)
Antigo	35,23% (31)	-	-
Evasões anteriores			
Nenhuma	97,73% (86)	98,53% (134)	95,34% (41)
Uma	2,27% (2)	1,47% (2)	2,33% (1)
Duas	-	-	2,33% (1)
Total	100% (88)	100% (136)	100% (43)

- indica frequência nula.

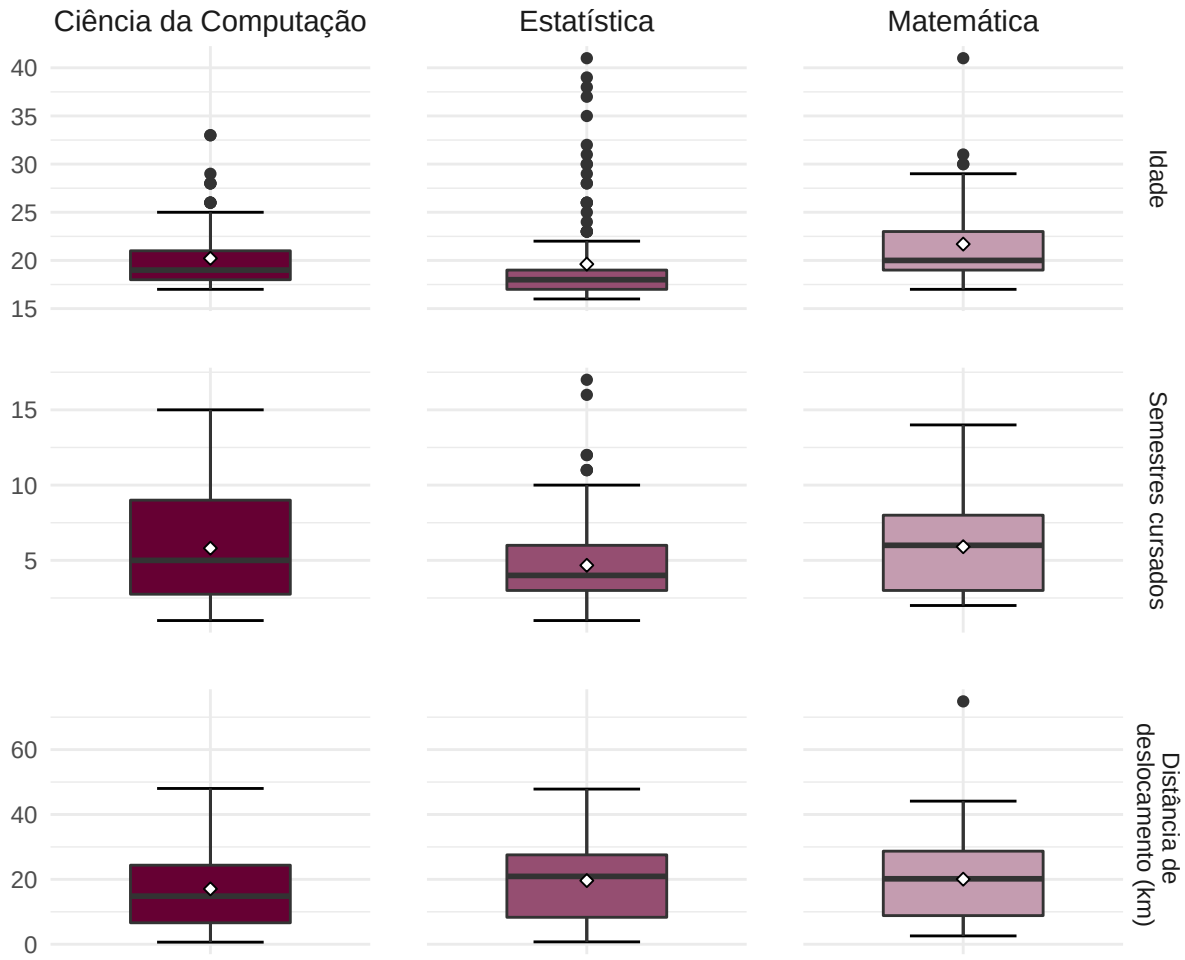
A Tabela 8 evidencia a diferença das formas de ingresso no curso de Matemática e nos cursos de Estatística e Ciência da Computação. Na Matemática, mudança de curso e duplo curso somam mais da metade dos ingressos. Na Estatística e na Ciência da Computação, essas formas de ingresso são de uma importância muito pequena, até mesmo nula, e aprovação no PAS ou vestibular são as formas mais frequentes. Apesar de também ser um programa avaliativo, o SISU/ENEM não é uma via de ingresso tão frequente quanto as outras.

É possível notar uma divergência entre a Matemática e os outros dois cursos também com relação à forma de saída. Na Estatística e Ciência da Computação, aproximadamente metade das alunas ainda estão ativas no curso; na Matemática, alunas ativas são apenas 18,6% do total. Percebe-se ainda que a maior proporção de alunas formadas são da Matemática, bem como a maior proporção de casos de desligamento por abandono. De modo geral, a maior parte das evasões no três cursos advém de algum tipo de desligamento institucional.

Mais de 80% das alunas de todos os cursos nunca cursou uma matéria de verão, e mais da metade reprovou Cálculo 1 alguma vez - com exceção da Matemática, em que 34,88% das alunas obteve reprovação. No caso específico do curso de Matemática, as lacunas no banco de dados impossibilitam determinar para 58,14% das alunas se houve ou não reprovação em Cálculo 1. A Ciência da Computação é a única graduação com mais de um tipo de currículo nesse estudo, sendo quase um terço currículos novos e o restante, antigos.

Por fim, vale destacar que algumas alunas possuem histórico de evasões anteriores, ou seja, evadiram o curso em um período anterior a 2011/1 e ingressaram novamente no mesmo curso em um período posterior. Foram identificadas seis alunas nessa situação, duas em cada curso. Particularmente na Matemática, uma das alunas chegou a evadir duas vezes antes de entrar mais uma vez no curso.

Gráfico 5: Distribuição das variáveis idade, semestres cursados e distância de deslocamento, por curso. Bacharelado do IE - UnB, 2011-2019



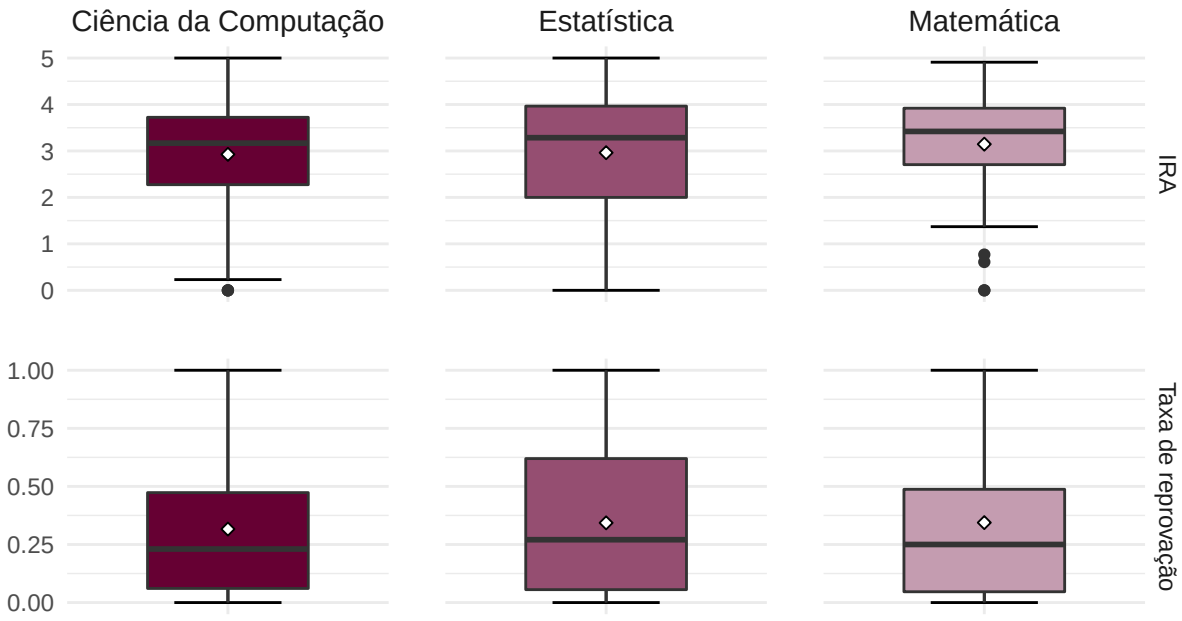
As idades médias de ingresso nos cursos de Ciência da Computação, Estatística e Matemática são 20,2, 19,6 e 21,7 anos, respectivamente, porém esses valores são distorcidos pela presença de valores discrepantes. Na Ciência da Computação, são discrepantes todas as idades acima de 25 anos; na Estatística, acima de 22 anos; na Matemática, acima de 29 anos. A idade mínima de ingresso observada é 17 anos, e a máxima, 41 anos. As alunas da Matemática são em geral um pouco mais velhas que as alunas dos outros cursos, porém é na Estatística que se observa a maior quantidade de alunas com mais de 30 anos.

É interessante notar que, apesar de ser o curso com mais estudantes acima dos 30 - nove ao todo -, a Estatística é também o curso com o perfil mais jovem. Com frequência, alunas já diplomadas ingressam na Estatística como uma segunda graduação, o que explica em partes a existência de alunas mais velhas no curso.

O tempo de permanência é em geral menor na Estatística do que nos outros dois cursos. As alunas que cursam mais de dez semestres na Estatística já são consideradas *outliers*. O número médio de semestres cursados é 5,8, 4,7 e 5,9 na ordem que aparecem no gráfico, e a variabilidade é bastante alta, ou seja, algumas alunas permanecem no curso por períodos bem mais longos do que outras. O tempo mínimo de permanência é 1 semestre, e o máximo 17 semestres.

As distribuições das distâncias de deslocamento são bastante similares entre os cursos, com médias entre 17,1 e 20,1 km. Assim como no tempo de permanência no curso, a variabilidade das distâncias é muito alta, indicando que algumas alunas percorrem distâncias muito maiores que outras desde o local de residência até a Universidade de Brasília. Na Matemática, a distância de 74,9 km é um *outlier* e corresponde a uma aluna que mora em Formosa, GO.

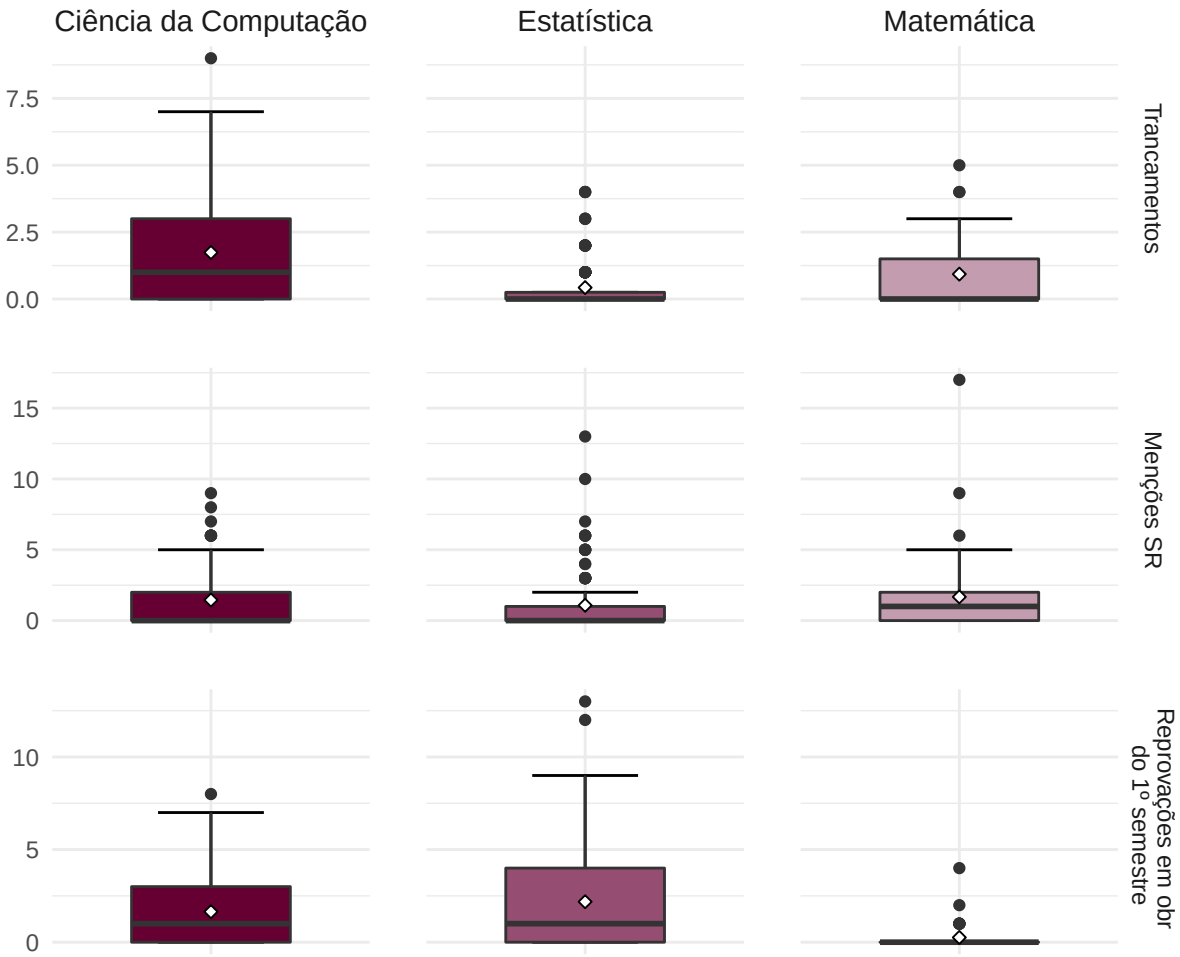
Gráfico 6: Distribuição das variáveis IRA e taxa de reprovação, por curso. Bacharelado do IE - UnB, 2011-2019



O IRA médio entre as alunas da Ciência da Computação e da Estatística é praticamente o mesmo 2,93 e 2,96, e um pouco superior na Matemática, 3,15. Nota-se também que IRAs muito baixos são menos comuns na Matemática do que nos outros dois cursos, com valores inferiores à 1,37 tidos como discrepantes. A metade central das alunas possui IRAs aproximadamente entre 2 e 4 na Ciência da Computação e na Estatística, e entre 2,5 e 4 na Matemática.

A distribuição das taxas de reprovação são muito semelhantes entre os cursos, principalmente entre a Ciência da Computação e a Matemática. A única diferença notável é a maior proporção de taxas acima de 0,5 na Estatística em comparação às outras graduações. As taxas médias são 0,32, 0,34 e 0,34, a mínima é 0 e a máxima é 1.

Gráfico 7: Distribuição das variáveis trancamentos, menções SR e reprovações em matérias obrigatórias do 1º semestre, por curso. Bacharelado do IE - UnB, 2011-2019



Há uma diferença notável no número de trancamentos de disciplinas realizados pelas alunas dos diferentes cursos. A maioria das alunas da Ciência da Computação trancam até três disciplinas, as da Matemática entre uma e duas disciplinas, e as da Estatística, menos de uma. O maior número de trancamentos, nove, é observado entre as estudantes da Ciência da Computação e é um *outlier*. O máximo de disciplinas trancadas na Estatística foram quatro e na Matemática, cinco.

Na Ciência da Computação e na Matemática, a maioria das alunas possui de zero a duas menções SR. Na Estatística, uma ou menos. É possível observar valores discrepantes em todos os cursos, entre eles o maior é de 17 menções SR, correspondente a uma aluna da Matemática.

No que diz respeito ao total de reprovações em matérias obrigatórias do 1º semestre, a Matemática destaca-se como o curso com menos reprovações. Vale mencionar, contudo, que apenas duas matérias são obrigatórias nesse curso, enquanto na Estatística são cinco e na Ciência da Computação, quatro ou seis a depender do currículo. Há na Estatística dois *outliers* que chamam atenção, um é de uma aluna com 10 reprovações e o outro é de uma aluna com 13 reprovações - o número máximo observado entre os três cursos.

4.1.2 Variáveis Explicativas e Evasão

Dando prosseguimento ao estudo do perfil acadêmico e social das alunas, serão avaliadas as relações existentes entre as variáveis explicativas e a evasão de curso. As Tabelas 9 e 10 a seguir mostram o percentual de evasão por categoria de cada variável.

Tabela 9: % de evasão em cada categoria das variáveis pessoais e de sistemas de cotas. Bacharelados do IE - UnB, 2011-2019

	Ciência da Computação	**Estatística**	**Matemática**
Raça			
Amarela	..	55,56% (5)	..
Branca	..	33,33% (14)	52,63% (10)
Negra	..	35% (14)	56,25% (9)
Sem informação	39,77% (35)	46,67% (21)	87,5% (7)
Local por renda			
Renda Alta	44,9% (22)	30,19% (16)	68,42% (13)
Renda Média-Alta	37,5% (6)	54,55% (18)	50% (6)
Renda Média-Baixa	36,36% (4)	50% (13)	80% (4)
Renda Baixa	-	30,77% (4)	66,67% (2)
Sem informação	42,86% (3)	27,27% (3)	25% (1)
Escola			
Particular	50% (29)	63,89% (23)	72,73% (16)
Pública	20% (6)	87,88% (29)	75% (9)
Sem informação	..	2,99% (2)	11,11% (1)
Sistema de cotas			
Não	47,46% (28)	35,44% (28)	65,79% (25)
Sim	24,14% (7)	45,61% (26)	20% (1)
Total	100% (35)	100% (54)	100% (26)

.. indica categoria inexistente; - indica frequência nula.

Tanto no curso de Estatística quanto no de Matemática, alunas brancas e negras evadiram dos cursos em proporções muito semelhantes. Na Estatística alunas de raça amarela evadiram com maior frequência que o restante, mas, como a quantidade de alunas é muito pequena, é preciso tomar cuidado com esse resultado. No geral, a relação entre raça e evasão parece ser fraca.

Já no caso do tipo de escola, há alguns indícios de relação com a evasão. No curso de Ciência da Computação, a proporção de evasão entre alunas advindas de escolas particulares é mais do que o dobro em comparação às alunas de escola pública. Na Estatística a relação é contrária: alunas de ensino público evadiram com maior frequência que as alunas de ensino particular. No curso de Matemática não se observa grande diferença entre os tipos de escola, mas a evasão de alunas sem informação de ensino foi proporcionalmente bem inferior às outras.

Na Ciência da Computação e principalmente na Matemática, alunas não-cotistas evadiram mais do que as cotistas. Aqui novamente é necessário cautela, pois o número de alunas cotistas da Matemática é muito pequeno para se ter uma noção adequada do relacionamento entre cota e evasão. O cenário é inverso na Estatística, em que alunas cotistas evadiram com mais frequência comparativamente às não-cotistas, apesar da diferença não ser tão grande.

Quanto à renda média do local de residência, os resultados são bem variados entre os três cursos. Na Ciência da Computação, as alunas residentes em locais de renda alta são as que evadiram em maior proporção; na Estatística, são as residentes em locais de renda média-alta; na Matemática, as que moram em locais de renda média-baixa. Somente partir dos dados da Tabela 9, é difícil identificar um padrão quanto à renda média da RA de residência e a taxa de evasão.

Tabela 10: % de evasão em cada categoria das variáveis acadêmicas. Bacharelados do IE - UnB, 2011-2019

	Ciência da Computação	Estatística	Matemática
Forma de ingresso			
PAS	41,18% (14)	80,77% (21)	-
SISU/ENEM	33,33% (2)	84,62% (11)	100% (2)
Vestibular	41,67% (15)	79,17% (19)	75% (6)
Mudança de curso	..	-	62,5% (5)
Duplo curso	..	..	85,71% (12)
Transferência	20% (1)	..	..
Convênio	100% (2)	..	..
Portadora de diploma	20% (1)	75% (3)	50% (1)
Sem informação	..	-	-
Cursou Verão			
Não	42,25% (30)	41,03% (48)	58,33% (21)
Sim	29,41% (5)	31,58% (6)	71,43% (5)
Reprovou Cálculo 1			
Não	33,96% (18)	21,33% (16)	46,67% (7)
Sim	48,57% (17)	62,3% (38)	66,67% (2)
Sem informação	..	..	68% (17)
Currículo			
Novo	24,56% (15)	39,71% (54)	60,47% (26)
Antigo	67,74% (21)	..	..
Evasões anteriores			
Nenhuma	40,7% (35)	39,55% (53)	60,98% (25)
Uma	-	50% (1)	100% (1)
Duas	..	..	-
Total	100% (35)	100% (54)	100% (26)

.. indica categoria inexistente; - indica frequência nula.

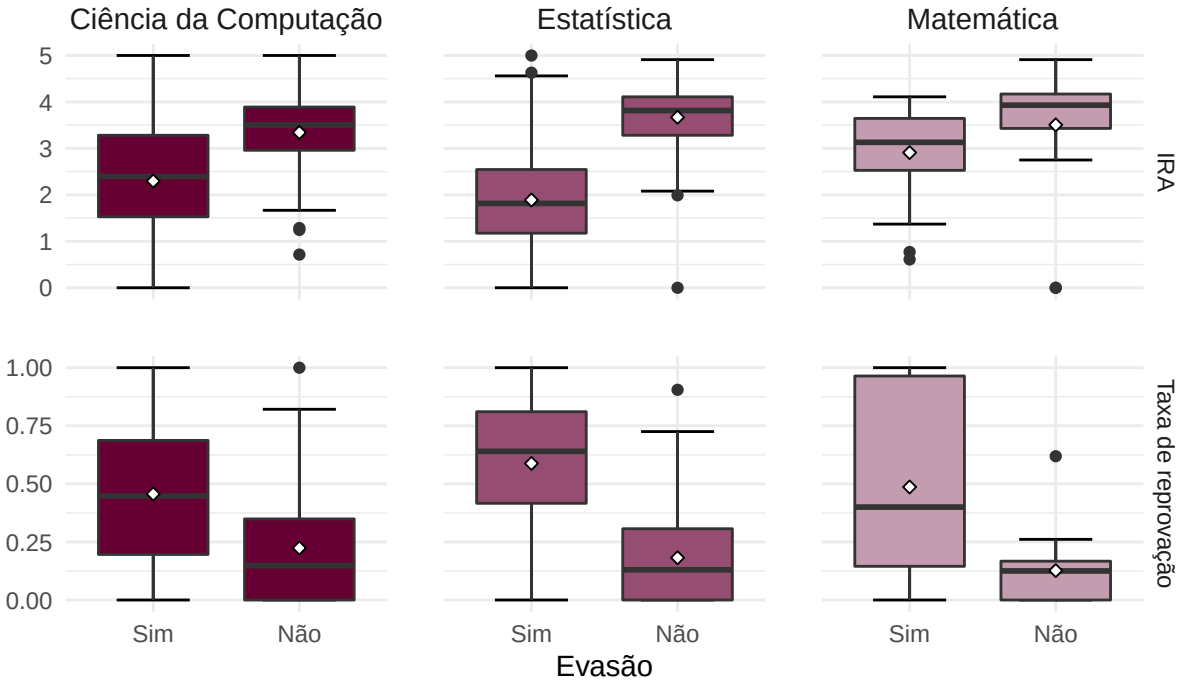
Na Ciência da Computação e na Estatística, não existe diferença notável de evasão entre as alunas que entraram pelo PAS ou vestibular, principais formas de ingresso desses cursos. Percebe-se também que a proporção de evasões de alunas que entraram pelo ENEM/SISU não destoa muito dessas duas formas de ingresso. Em contrapartida, na Matemática há uma maior disparidade de evasões entre as principais formas de ingresso, mudança de curso e duplo curso. Todas as alunas cuja forma de ingresso não foi especificada são ativas - portanto não evadiram-, e por isso a taxa de evasão é nula.

Aqui cabe um comentário sobre a Matemática e a Estatística: como grande parcela das alunas que não evadiram estão acumuladas na categoria "sem informação", a taxa de evasão das outras formas de ingresso torna-se bastante elevada - principalmente em comparação à Ciência da Computação, em que isso não ocorre.

A proporção de evasão é maior entre alunas que não cursaram disciplinas de verão, contudo a diferença não é muito grande em relação às que cursaram. Apenas na Matemática acontece o contrário: a evasão é maior entre alunas com matérias de verão. Nos três cursos, estudantes que possuem reprovação em Cálculo 1 evadiram em maiores proporções do que as que foram aprovadas, principalmente na Estatística.

No curso da Ciência da Computação, parece haver relação entre evasão e currículo vigente, uma vez que alunas do currículo antigo evadiram com frequência bem superior às alunas do currículo novo. Por último, os resultados revelam que nem todas as alunas com histórico de evasão acabam evadindo o curso nos ingressos seguintes. Na Ciência da Computação, as duas alunas com histórico de evasão não haviam saído do curso até o final do período de analise, 2019/2. Na Estatística, uma delas evadiu e a outra não. A aluna que evadiu duas vezes anteriormente do curso de Matemática conseguiu se formar.

Gráfico 8: Distribuição das variáveis IRA e taxa da reprovação, por curso e evasão. Bacharelado do IE - UnB, 2011-2019



Em todos os três cursos, o IRA das alunas evadidas é menor que das alunas não-evadidas. O mesmo também pode ser dito sobre a taxa de reprovação, porém o evento observado é contrário: alunas evadidas possuem taxa de reprovação em geral mais elevadas do que as alunas que não evadiram os cursos. Sabendo que IRA e taxa de reprovação são dois dos principais medidores de performance acadêmica na Universidade de Brasília e, portanto, do sucesso das alunas nos cursos, é esperado que essas sejam variáveis bastante relevantes no estudo de evasões.

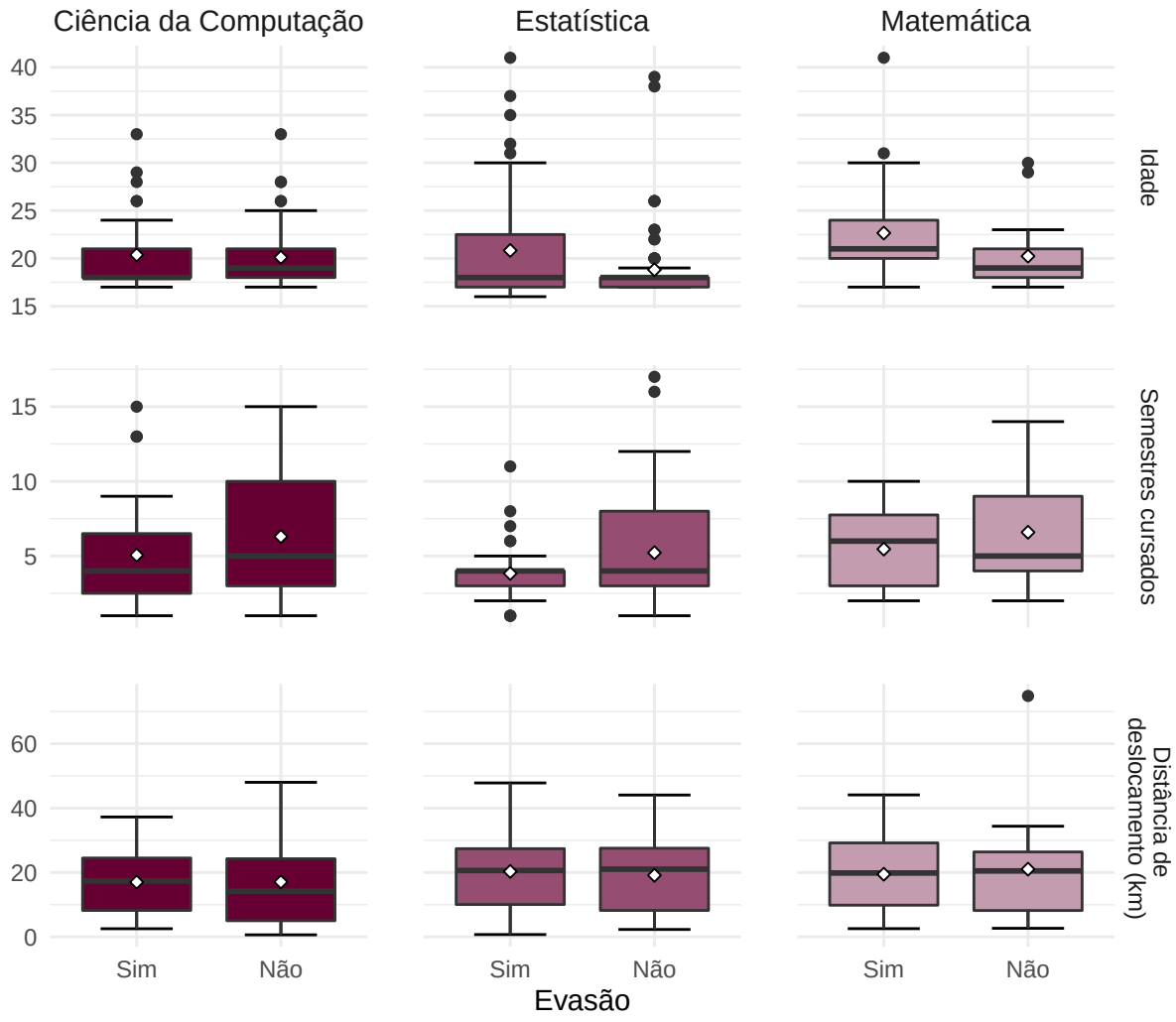
Com respeito ao índice de rendimento acadêmico, desperta certa suspeita a presença de alunas evadidas com IRAs muito elevados, em especial aquelas com IRA máximo. Para tentar compreender esse fenômeno, foi feita uma investigação de todas as alunas evadidas com IRA ≥ 4. São ao todo seis alunas, uma da Matemática, duas da Ciência da Computação e três da Estatística. Verificou-se que quatro delas saíram ou por desligamento voluntário ou novo vestibular, e evadiram logo no início do curso, passados um ano ou menos desde seu ingresso. Todas as quatro têm taxa de reprovação nula.

As outras duas alunas são da Estatística e correspondem exatamente aos dois valores discrepantes de IRA observados no gráfico - 4,63 e 5. Ambas foram desligadas por abandono e cursaram efetivamente apenas um semestre antes de abandonarem o curso.

No que se refere à taxa de reprovação, chama atenção a aluna não evadida da Ciência da Computação com taxa de 100%, considerada como *outlier*. Esse dado foi conferido no banco original e diz respeito a uma estudante ainda ativa no curso que reprovou todas as matérias cursadas desde o seu ingresso no curso, 2018/2, até o semestre de 2019/2.

Na Estatística e na Matemática, as duas taxas de reprovação discrepantes observadas entre as alunas que não evadiram são de 90,5% e 61,9%, respectivamente. A estudante de Estatística ainda está ativa no curso, mas a da Matemática se formou - mesmo com uma taxa de reprovação tão alta. Possivelmente isso é reflexo das lacunas nos dados da Matemática, pois o registros de muitas disciplinas não consta no banco e, consequentemente, a taxa de reprovação é afetada. Assim, essa provavelmente não é a verdadeira taxa de reprovação dessa aluna formada.

Gráfico 9: Distribuição das variáveis idade, semestres cursados e distância de deslocamento, por curso e evasão. Bacharelado do IE - UnB, 2011-2019



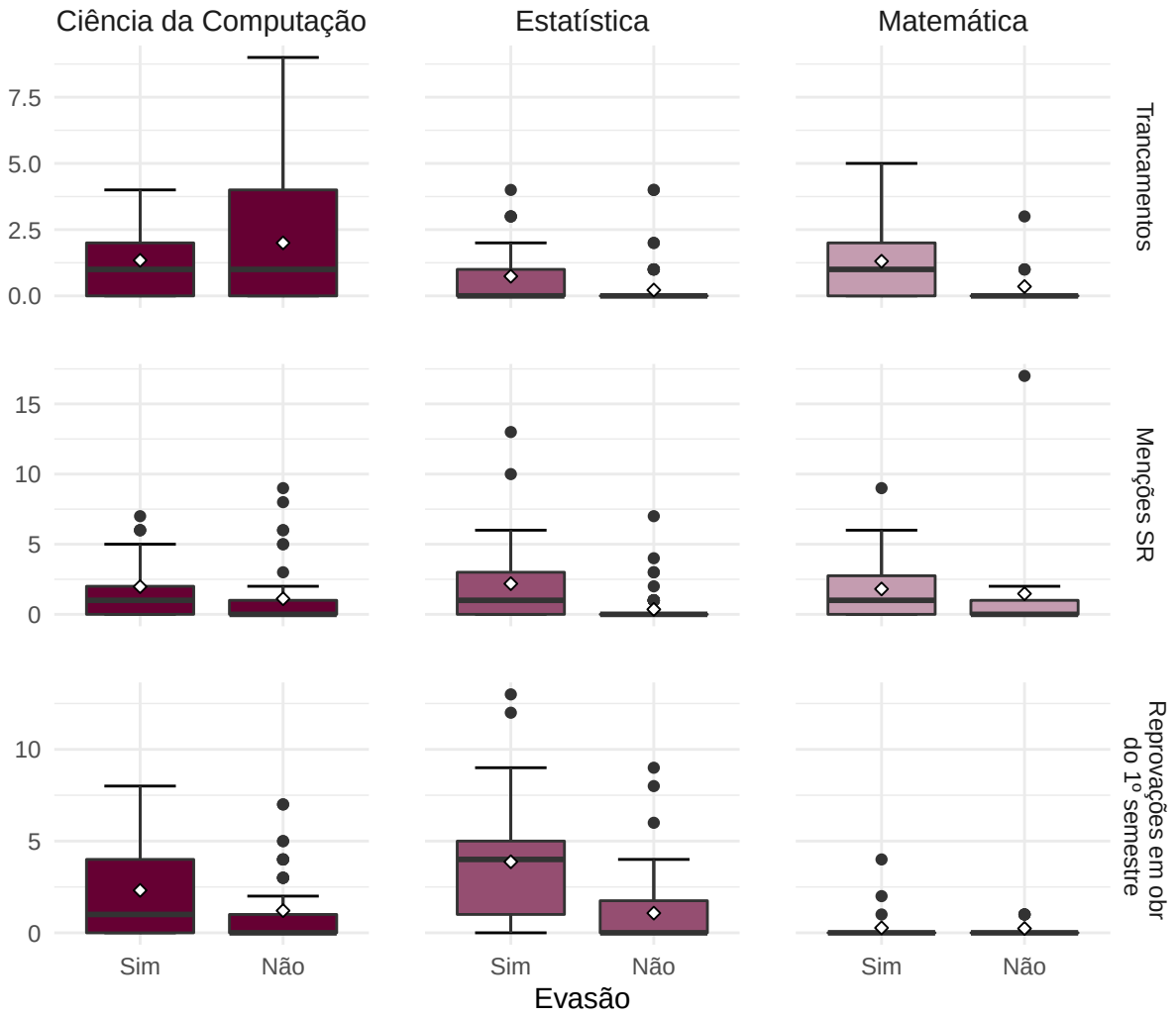
Não há grande diferença de idade entre as alunas que evadiram e não evadiram o curso de Ciência da Computação. Na Estatística e na Matemática, por outro lado, há indícios de que alunas mais velhas têm maior tendência à evasão do que alunas mais jovens. Dito isso, duas alunas da Estatística com idades de 38 e 39 anos ainda estavam ativas no semestre de 2019/2, e mais outras duas da Matemática de 29 e 30 anos conseguiram concluir a graduação.

Naturalmente, alunas evadidas cursam menos semestres do que as alunas não evadidas, dado que costumam sair do curso antes do tempo mínimo de formatura. Contudo, os gráficos revelam que algumas alunas permanecem um longo período de tempo no curso e mesmo assim evadem. Especificamente na Estatística, a grande maioria das alunas evadidas cursaram no máximo 4 semestres, revelando que as evasões ocorrem ma-

joritariamente no dois anos iniciais de curso. Além disso, na Matemática o tempo mínimo de permanência entre alunas evadidas e não evadidas foi de dois semestres, ou seja, todas as alunas ficaram pelo menos um ano no curso antes de sair.

Em todos os cursos as distâncias percorridas entre o local de residência e a universidade são muito similares entre alunas que evadiram e não evadiram, revelando que essa variável não é importante, à primeira vista, para o estudo das evasões.

Gráfico 10: Distribuição das variáveis trancamentos, menções SR e reprovações em matérias obrigatórias do 1º semestre, por curso e evasão. Bacharelado do IE - UnB, 2011-2019



Em relação ao número de trancamentos realizados, ocorre um fenômeno interessante: na Ciência da Computação alunas que não evadiram tendem a trancar mais disciplinas do que as alunas que evadiram, mas na Matemática e Estatística o oposto é observado. O trancamento é um recurso utilizado por alunos para interromper de forma voluntária a matrícula em determinada disciplina, qualquer que seja o motivo. Visto

por um lado, o trancamento pode evitar que a aluna reprove ou passe com uma menção muito baixa em uma disciplina, dando oportunidade para que curse a matéria em outro momento e em condições mais favoráveis ao seu aprendizado. Por outro, trancar uma disciplina pode indicar dificuldades em acompanhar a disciplina, e nesse caso, muitos trancamentos podem sinalizar que o rendimento da aluna é ruim.

Levando em conta essas duas interpretações, uma possível explicação para os resultados observados é: no curso de Ciência da Computação, as alunas são orientadas a trancar disciplinas quando reconhecem algum impedimento para a sua aprovação e, assim, o trancamento é um recurso utilizado à favor das estudantes; já na Matemática e na Estatística, as alunas não têm o costume de trancar disciplinas e só o fazem quando realmente estão enfrentando muita dificuldade com na matéria, ou com risco de serem reprovadas.

Quanto ao número de menções SR, não é surpresa que alunas evadidas abandonem disciplinas com mais frequência que as alunas não evadidas. Há, é claro, exceções e algumas alunas que se formaram ou ainda estão ativas acumulam um número alto de menções SR. É a situação, por exemplo, de uma aluna não evadida da Ciência da Computação com 9 menções SR, de uma aluna na Estatística com 7 e outra na Matemática com impressionantes 17 menções SR. Essa aluna é justamente aquela que evadiu o curso duas vezes anteriormente, e por isso tem um longo histórico de disciplinas abandonadas.

Por fim, o total de reprovações em matérias obrigatórias do 1º semestre segue o mesmo padrão das outras variáveis: alunas evadidas em geral possuem mais reprovações comparativamente às não evadidas. Na Estatística é possível identificar dois *outliers* entre as alunas que evadiram, equivalentes a 12 e 13 reprovações - os maiores números dentre os três cursos. O que mais chama atenção, porém, é a quantidade baixíssima de reprovações na Matemática, tanto entre alunas evadidas quanto não evadidas. Isso se justifica por dois motivos principais, já mencionados anteriormente: muitas informações dos históricos das alunas estão faltantes do banco da Matemática; e apenas duas matérias são obrigatórias no 1º semestre de curso, então espera-se que o total de reprovações seja de fato menor.

4.2 Análise de Associação

Previamente à modelagem, é interessante avaliar a relação entre as possíveis variáveis explicativas e a variável resposta para identificar quais fatores são mais relevantes para o estudo. A Tabela 11 a seguir contém os resultados dos testes de associação de cada variável com a evasão. Para as variáveis quantitativas ou numéricas, foram ajustados modelos de regressão logística simples e para as qualitativas ou categóricas, testes de associação Qui-quadrado de Pearson e teste exato de Fisher, nos casos em que o número de observações fosse inferior a 5.

Tabela 11: Associação entre variáveis explicativas e evasão

	Ciência da Computação		**Estatística**		**Matemática**	
Associação com Evasão	Estatística	p-valor	Estatística	p-valor	Estatística	p-valor
IRA	-3,6361	**0,0003**	-6,1270	≤ **0,0001**	-1,5594	0,1189
Taxa de reprovação	3,3082	**0,0009**	6,3179	≤ **0,0001**	2,6714	**0,0076**
Sistema de Cotas	3,4942	0,0616	1,0374	0,3084	..	0,0707
Evasões anteriores	-0,0095	0,9924	0,2976	0,7660	-0,7149	0,4747
Idade	0,3416	0,7326	2,2418	**0,0250**	1,5527	0,1205
Trancamentos	-1,4567	0,1452	3,0336	**0,0024**	2,0766	**0,0378**
Menções SR	1,7648	0,0776	3,8226	≤ **0,0001**	0,3560	0,7276
Cursou Verão	0,4843	0,4865	0,2786	0,5976	..	0,6845
Semestres cursados	-1,4419	0,1493	-2,5433	**0,0110**	-1,1846	0,2362
Reprovações em obg.	2,1840	**0,0290**	4,8934	≤ **0,0001**	0,1512	0,8798
Distância	-0,0258	0,9794	0,5640	0,5728	-0,3351	0,7376
Local por Renda	..	0,4449	..	0,1231	..	0,4588
Reprovou Cálculo 1	1,1378	0,2510	21,8965	≤ **0,0001**	..	0,3983
Raça	..	..	..	0,4086	..	0,2387
Escola	6,2294	**0,0126**	..	≤ **0,0001**	..	**0,0033**
Forma de ingresso	..	0,5394	..	≤ **0,0001**	..	**0,0005**
Currículo	13,8791	**0,0002**	..	..	..	..

Valores indicados como .. não se aplicam ou estão indisponíveis. No caso dos testes exatos de Fisher, as estatísticas estão indisponíveis porque o software utilizado retorna apenas os p-valores de teste.

Tomando um nível de significância de 5%, observa-se que poucas variáveis da Ciência da Computação e Matemática possuem associação significativa com a variável resposta evasão. A Estatística possui o maior conjunto de variáveis significativamente associadas à evasão, dez ao todo.

Na Ciência da Computação, são significativas as associações apenas com as variáveis IRA, taxa de reprovação, reprovação em matérias obrigatórias do 1º semestre, escola e currículo. Já na Matemática, as variáveis taxa de reprovação, trancamentos, escola e

forma de ingresso são as únicas que, nessa análise inicial, obtiveram associações significantes com a evasão de alunas.

Vale notar que os resultados observados nas variáveis escola e forma de ingresso nos dados da Estatística e da Matemática devem-se, em grande parte, à concentração de alunas ativas em uma única categoria - "Sem informação". Por isso, a significância da relação dessas variáveis com a evasão é questionável e deve ser interpretada com cautela.

Além das análises de associação com a variável resposta, é necessário também avaliar a correlação entre as variáveis explicativas, uma vez que variáveis muito correlacionadas entre si podem causar problemas de multicolinearidade no ajuste dos modelos. A Tabela 12 lista todos os pares de variáveis quantitativas cujo coeficiente de correlação de Pearson, r_{xy}, é igual ou superior à 0,5.

Tabela 12: Correlações entre variáveis explicativas quantitativas ($|r_{xy}| \geq 0,5$)

Variáveis explicativas	**Correlação**
Ciência da Computação	
IRA e Taxa de reprovação	-0,94
Menções SR e Taxa de reprovação	0,70
IRA e Menções SR	-0,68
Taxa de reprovação e Reprovações em obg.	0,62
IRA e Reprovações em obg.	-0,55
Trancamentos e Semestres cursados	0,53
Evasões anteriores e Menções SR	0,50
Menções SR e Reprovações em obg.	0,50
Estatística	
IRA e Taxa de reprovação	-0,95
Taxa de reprovação e Reprovações em obg.	0,77
IRA e Reprovações em obg	-0,68
IRA e Menções SR	-0,66
Menções SR e Taxa de reprovação	0,64
Menções SR e Reprovações em obg.	0,63
Matemática	
Evasões anteriores e Menções SR	0,76
IRA e Taxa de reprovação	-0,50

As variáveis IRA e taxa de reprovação são altamente correlacionadas nos dados da Ciência da Computação e Estatística. Curiosamente, esse resultado não é observado na Matemática, cuja principal correlação de variáveis ocorre entre o número de evasões

anteriores e o total de menções SR.

Nota-se que, de todas as variáveis numéricas consideradas nesse estudo, apenas idade e distância de deslocamento não possuem correlações iguais ou superiores à 0,5 com outras variáveis.

4.3 Modelagem

Pretende-se identificar os principais fatores que levam à evasão de estudantes do gênero feminino nos cursos de bacharelado do Instituo de Ciências Exatas (IE) a partir do ajuste de modelos de regressão logística. Inicialmente, será ajustado o mesmo modelo nos dados dos três cursos e, posteriormente, serão selecionados os modelos mais adequados para cada curso individualmente.

Algumas variáveis explicativas não poderão ser utilizadas nesse primeiro modelo comum entre os cursos. É o caso da variável raça, que está incompleta no banco da Ciência da Computação, da variável currículo, que só possui uma única categoria nos bancos da Matemática e da Estatística, e das variáveis forma de ingresso no curso e escola. Essas últimas precisaram ser retiradas porque provocam discriminação quase perfeita entre alunas evadidas e não evadidas, isto é, são capazes de separar a variável resposta com muita precisão em 0 e 1, provocando um erro no ajuste. Isso ocorre nos bancos da Matemática e da Estatística, nos quais as alunas ainda ativas - portanto, não evadidas - estão todas agrupadas em uma mesma categoria, "Sem informação".

Anteriormente foi verificado que as variáveis IRA e taxa de reprovação são altamente correlacionadas, assim, para evitar o efeito de multicolinearidade serão ajustados dois modelos diferentes, um com cada variável.

Tabela 13: Modelo com a variável IRA. Bacharelado do IE - UnB, 2011-2019

	Ciência da Computação		**Estatística**		**Matemática**	
Parâmetro	Estimativa	p-valor	Estimativa	p-valor	Estimativa	p-valor
Intercepto	7,85631	**0,0281**	2,13524	0,451350	17,2212	0,2718
IRA	-1,06734	**0,0363**	-1,49703	**0,000613**	-4,7263	**0,0275**
Sistema de Cotas - Sim	-2,40047	**0,0138**	-0,42052	0,606223	-13,5641	0,1349
Evasões anteriores	-16,38541	0,9951	-1,62556	0,435759	-9,6048	0,2465
Idade	-0,20576	0,0548	0,11530	0,141071	0,4232	0,1968
Trancamentos	-0,24854	0,2579	1,25464	**0,003945**	4,3995	0,0803
Menções SR	0,12224	0,7170	-0,04535	0,868268	-0,1123	0,9053
Cursou Verão - Sim	-0,65490	0,4627	-0,37344	0,745409	7,0897	0,0713
Semestres cursados	-0,13020	0,1805	-0,52740	**0,003932**	-2,3717	**0,0273**
Reprovações em obg.	0,16145	0,4990	0,15631	0,443025	0,8054	0,7973
Distância	0,04100	0,3277	-0,03620	0,406882	0,1089	0,4049
Local - Renda Alta	0,31606	0,7865	1,49955	0,280624	2,6477	0,7222
Local - Renda Média-Alta	-0,59944	0,6419	2,12218	0,136962	-1,4151	0,8293
Local - Renda Média-Baixa	-2,22317	0,1765	2,13710	0,127644	6,7645	0,7852
Local - Renda Baixa	-15,81293	0,9924	0,01694	0,990463	4,7976	0,5886
Reprovou Cálculo 1 - Sim	0,41507	0,6904	0,83154	0,347504	5,1347	0,6043
Reprovou Cálculo 1 - Sem info	..	..	..	..	-3,5527	0,3773

.. indica parâmetro inexistente.

Tabela 14: Modelo com a variável Taxa de Reprovação. Bacharelado do IE - UnB, 2011-2019

	Ciência da Computação		**Estatística**		**Matemática**	
Parâmetro	Estimativa	p-valor	Estimativa	p-valor	Estimativa	p-valor
Intercepto	3,448	0,1910	-4,59114	**0,03021**	-23,82833	0,1416
Taxa de Reprovação	0,0299	0,1451	0,05397	**0,00868**	0,23155	0,0702
Sistema de Cotas - Sim	-2,234	**0,0158**	-0,34114	0,65099	-5,88991	0,3374
Evasões anteriores	-17,720	0,9946	-1,22534	0,54162	1,26665	0,8827
Idade	-0,1921	0,0571	0,13211	0,06007	0,23282	0,4555
Trancamentos	-0,2374	0,2723	1,43842	**0,00111**	2,87573	0,1282
Menções SR	0,2762	0,3905	0,04276	0,86230	-1,34231	0,2945
Cursou Verão - Sim	-0,7403	0,4189	1,13229	0,77869	0,63241	0,7832
Semestres cursados	-0,1224	0,1983	-0,49777	**0,00484**	-0,45461	0,3455
Reprovações em obg.	0,08826	0,7156	0,03006	0,89231	0,04972	0,9855
Distância	0,04595	0,2560	-0,02534	0,53999	0,19213	0,1822
Local - Renda Alta	0,00526	0,9965	1,20321	0,36009	14,84359	0,1730
Local - Renda Média-Alta	-0,8355	0,5062	2,00939	0,12806	9,57312	0,2887
Local - Renda Média-Baixa	-2,710	0,1017	1,60592	0,21890	11,71352	0,2396
Local - Renda Baixa	-16,610	0,9920	-0,54639	0,69049	3,04804	0,9708
Reprovou Cálculo 1 - Sim	0,7248	0,4761	0,97589	0,23655	-1,19357	0,8856
Reprovou Cálculo 1 - Sem info	..	..	..	..	4,93803	0,2011

.. indica parâmetro inexistente.

De um modo geral, as estimativas dos parâmetros são bem semelhantes entre os modelos com as variáveis IRA e taxa de reprovação. Somente na Matemática nota-se maior divergência entre os valores estimados. Em ambos os modelos também é possível observar alguns coeficientes muito elevados, como nos parâmetros "Evasões anteriores"e "Local - Renda Baixa"nos modelos da Ciência da Computação, e "Sistema de Cotas"e "Local - Renda Média-Baixa"nos modelos da Matemática. Além disso, algumas variáveis que deram significativas individualmente na análise de correlação, nos modelos completos não apresentaram significância, e vice-versa.

Uma vez ajustados os modelos em comum, serão selecionados os modelos mais adequados para cada um dos três cursos separadamente. Os dados foram divididos em duas amostras, uma de construção do modelo com aproximadamente 80% das observações, e outra de validação com 20% das observações. A princípio, serão candidatas aos modelos finais todas as variáveis que em uma análise preliminar obtiveram p-valor $\leq$ 0,25, como sugerido por Hosmer e Lemeshow (2000). A adoção do nível de confiança mais tradicional de 5% nessa etapa inicial pode acabar deixando de fora variáveis que se revelarão importantes para a análise na etapa de modelagem. A seleção será feita com auxílio de métodos *Stepwise* e dos resultados de testes de significância dos parâmetros, objetivando encontrar os modelos mais parcimoniosos.

4.3.1 Ciência da Computação

A partir das análises anteriores, foram selecionadas as variáveis que, à priori, demonstraram relevância no estudo da evasão das alunas de Ciência da Computação: IRA, taxa de reprovação, currículo, escola, sistema de cotas, semestres cursados, trancamentos, menções SR, reprovou Cálculo 1, forma de ingresso no curso e local por renda. No caso dessas duas últimas, foram testados tanto o agrupamento original quanto diferentes agrupamentos de categorias.

A forma de ingresso foi agrupada em três classes: "PAS", "Vestibular"e "Outra", contendo todas as categorias restantes. Já o local por renda foi agrupado em duas categorias: "Renda Alta ou Média-Alta"e "Renda Baixa, Média-Baixa ou Sem informação". Os p-valores do teste Qui-quadrado de associação entre a evasão e essas variáveis agrupadas são, respectivamente, 0,82 e 0,41, indicando que não são significativamente relacionadas. Apesar disso, a escolha de agrupar essas variáveis justifica-se pela necessidade de se ter

mais observações em cada categoria.

Foram testados modelos com o IRA e a taxa de reprovação separadamente. Utilizou-se o método *Stepwise* para reduzir o conjunto de variáveis explicativas e, a partir daí, outras variáveis foram inseridas ou retiradas dos modelos com base no teste de razão de verossimilhança, à um nível de 5%. As interações entre variáveis também foram testadas. Com base no *AIC* e na significância dos parâmetros, concluiu-se que a taxa de reprovação não representava vantagem em relação à variável IRA. Assim, o modelo final escolhido foi:

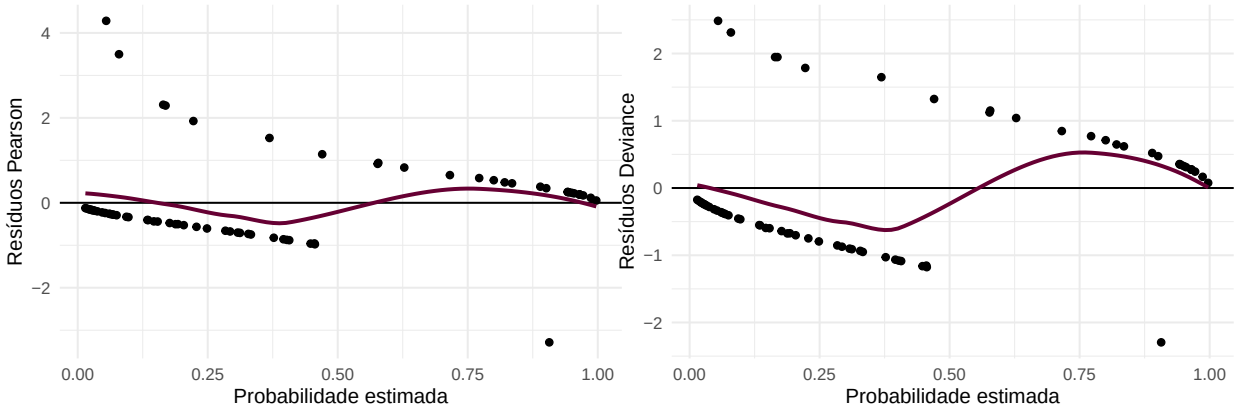
Tabela 15: Modelo do Bacharelado em Ciência da Computação, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	6,2882	1,6997	3,700	0,000216
IRA	-1,1532	0,3527	-3,270	0,001076
Currículo - Novo	-2,5363	0,9086	-2,791	0,005248
Escola - Pública	-1,8216	0,9390	-1,940	0,052379
Semestres cursados	-0,2178	0,1076	-2,025	0,042866

O modelo selecionado contém as variáveis IRA, currículo, escola e total de semestres cursados. Observa-se que a variável escola não é significativa, porém foi mantida no modelo devido à proximidade do p-valor obtido, 0,052, com o nível de significância adotado de 0,05. O *AIC* do modelo é de 63,975.

A qualidade do ajuste pode ser analisada por meio dos gráficos dos resíduos *versus* probabilidades estimadas pelo modelo, a partir da suavização de Lowess. Um modelo adequado possui suavização aproximadamente sem inclinação e com intercepto zero. Não é o caso dos resíduos desse modelo, como observado nos gráfico abaixo. Contudo, deve-se levar em conta que o número pequeno de observações torna esse resultado pouco confiável.

Gráfico 11: Resíduos - Modelo Ciência da Computação



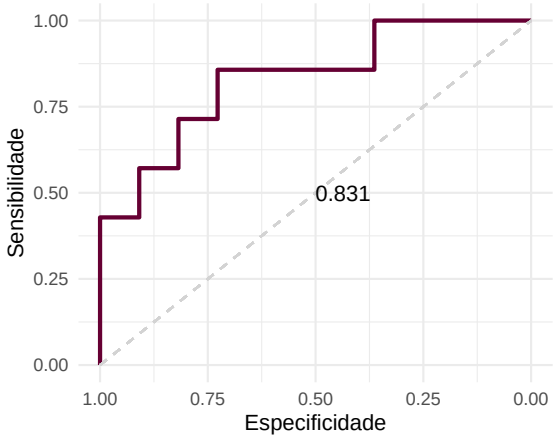
(a) Resíduos *Pearson* (b) Resíduos *Deviance*

Tabela 16: Testes de adequabilidade de ajuste - Modelo Ciência da Computação

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	8,1688	0,4171
Resíduos *Pearson*	69,930	0,316
Resíduos *Deviance*	53,975	0,834

Outra forma de verificar o ajuste dos modelos são os testes de adequabilidade do ajuste. Com base nos resultados da Tabela 16, nenhum teste rejeita a hipótese nula de que o modelo está bem ajustado aos dados, a um nível de significância de 5%.

Gráfico 12: Curva ROC - Modelo Ciência da Computação



Em relação ao poder preditivo do modelo, a área abaixo da curva ROC (0,831)

indica que a discriminação do modelo é excelente. O melhor ponto de corte das probabilidades preditas é 0,6, ponto em que se obtém as seguintes classificações:

Tabela 17: Matriz de Confusão - Modelo Ciência da Computação

		Observado	
		$Y = 1$	$Y = 0$
Predito	$\hat{Y} = 1$	6	3
	$\hat{Y} = 0$	1	8

A acurácia do modelo no ponto de corte 0,6 é de 77,78%, a sensibilidade é de 85,71% e a especificidade é de 72,73%. Uma vez verificados o poder de teste e a adequabilidade do modelo, resta interpretar os parâmetros. Para isso, o modelo será ajustado aos dados completos a fim de se obter estimativas mais precisas dos coeficientes.

Tabela 18: Modelo Final do Bacharelado em Ciência da Computação, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	6,67677	1,65931	4,024	$\leq$ 0,0001
IRA	-1,19420	0,33892	-3,524	0,000426
Currículo - Novo	-2,62267	0,83924	-3,125	0,001778
Escola - Pública	-1,98458	0,84438	-2,350	0,018756
Semestres cursados	-0,22903	0,09761	-2,346	0,018953

Percebe-se que, com os dados completos, os parâmetros do modelo são bem similares aos estimados com a amostra de construção. Assim, pode-se concluir que o modelo escolhido é válido.

Tabela 19: Razão de chances e IC 95% - Modelo Final Ciência da Computação

Variável Explicativa	**Razão de Chances**	**IC 95%**
IRA	0,303	0,141 - 0,546
Currículo - Novo	0,073	0,011 - 0,328
Escola - Pública	0,137	0,021 - 0,627
Semestres cursados	0,795	0,643 - 0,948

A cada acréscimo de uma unidade no IRA da aluna, suas chances de evasão são reduzidas em 70%, controlando os outros fatores. Além disso, a cada semestre cursado as chances de evasão diminuem em 20,5%.

As alunas que cursaram Ciência da Computação com o currículo novo em vigor possuem suas chances de evadir o curso reduzidas em 92,7% com relação às alunas que cursaram com o currículo antigo. Assim, pode-se concluir que a troca de currículo é um fator de grande relevância na evasão do curso. Contudo, é preciso lembrar que todas as alunas ativas seguem o currículo novo e, apesar de não terem evadido o curso até o momento final analisado nesse estudo, 2019/2, não significa que elas não evadirão em um momento futuro.

Alunas advindas de escolas públicas têm 86,3% menos chances de evadir o curso em comparação às alunas advindas de escola particular, mantendo-se os outros fatores constantes.

4.3.2 Estatística

As variáveis inicialmente selecionadas para o estudo da evasão das alunas de Estatística foram: IRA, taxa de reprovação, idade, semestres cursados, trancamentos, menções SR, reprovações em matérias obrigatórias, reprovou Cálculo 1 e local por renda. As variáveis escola e forma de ingresso, apesar de significativas na análise de correlação, não foram incluídas devido ao problema mencionado no início da Seção 4.3 de modelagem.

Em relação à variável raça, alunas que se identificam como amarelas foram agrupadas junto com as alunas sem informação, na categoria "Amarela ou sem informação". O local por renda foi agrupado em duas categorias: "Renda Alta ou Média-Alta"e "Renda Baixa, Média-Baixa ou Sem informação". Os testes Qui-quadrado de associação das variáveis agrupadas com a variável resposta resultaram em p-valores iguais a 0,261 e 1, respectivamente, indicando que local por renda agrupado e raça não são significativamente relacionados com a evasão. Assim como ocorreu no banco da Ciência da Computação, a escolha de agrupar essas variáveis justifica-se pela necessidade de se ter mais observações em cada categoria. Ambas as variáveis foram testadas nos modelos em sua forma original e com o novo agrupamento proposto.

Modelos com o IRA e a taxa de reprovação foram ajustados separadamente. A seleção das variáveis se deu da mesma forma como descrita anteriormente. Curiosamente, os modelos testados com a variável IRA quase sempre tinham algum problema de ajuste, com base nos resultados dos testes de adequabilidade e análise de resíduos. Dessa forma, apesar de ser muito significativa, essa variável foi desconsiderada e o modelo final escolhido foi:

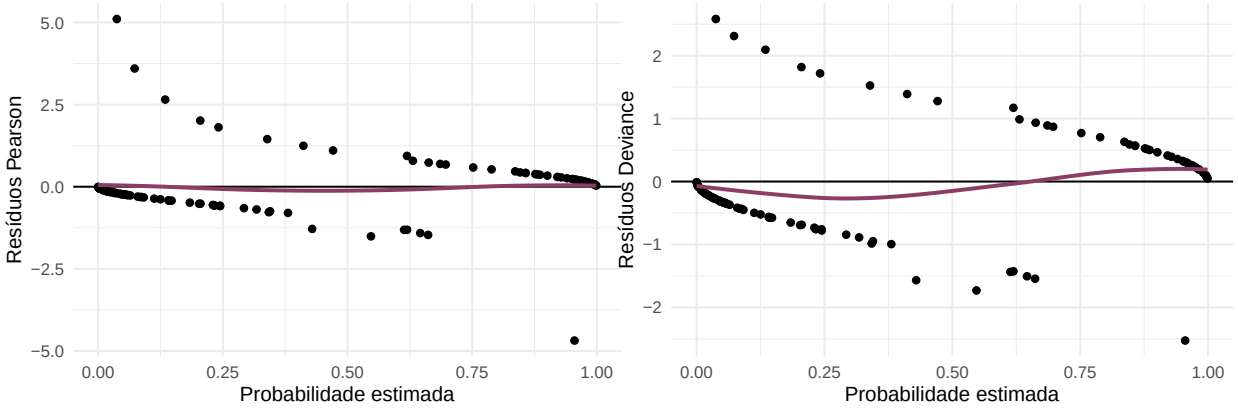
Tabela 20: Modelo do Bacharelado em Estatística, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	-4,60759	1,68969	-2,727	0,00639
Taxa de Reprovação	0,06613	0,01330	4,971	$\leq$ 0,0001
Semestres cursados	-0,47444	0,19948	-2,378	0,01739
Trancamentos	1,57363	0,51953	3,029	0,00245
Idade	0,15482	0,07103	2,180	0,02929

De todas as variáveis explicativas analisadas, apenas quatro ficaram no modelo final: taxa de reprovação, semestres cursados, trancamentos e idade. A taxa de reprovação precisou ser redimensionada, pois o coeficiente do modelo estava muito elevado com os valores originais. Sendo assim, a taxa de reprovação foi multiplicada por 100, e portanto deve ser interpretada em termos percentuais. O *AIC* do modelo é de 75,24.

As suavizações de Lowess nos gráficos de resíduos abaixo são aproximadamente horizontais com intercepto zero, indicando que o modelo está ajustado adequadamente.

Gráfico 13: Resíduos - Modelo Estatística



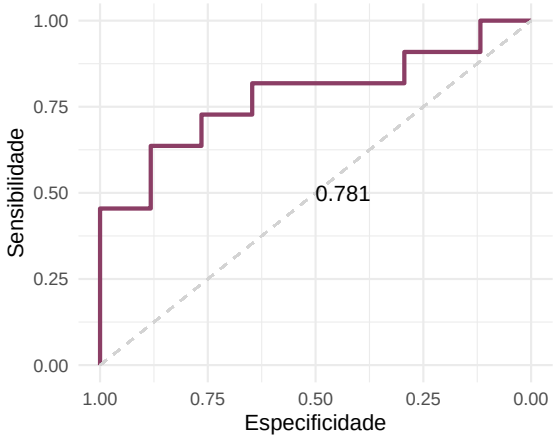
(a) Resíduos *Pearson* (b) Resíduos *Deviance*

Tabela 21: Testes de adequabilidade de ajuste - Modelo Estatística

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	5,3761	0,717
Resíduos *Pearson*	98,594	0,604
Resíduos *Deviance*	65,240	0,998

Os testes de adequabilidade corroboram o resultado observado nas gráficos anteriores, uma vez que a hipótese de que o modelo está bem ajustado não é rejeitada em nenhum dos testes.

Gráfico 14: Curva ROC - Modelo Estatística



A área abaixo da curva ROC (0,781) indica que a discriminação do modelo é aceitável, ou seja, o poder preditivo do modelo não é ideal mas é bom o suficiente. O melhor ponto de corte das probabilidades preditas é 0,6, ponto em que se obtém as seguintes classificações:

Tabela 22: Matriz de Confusão - Modelo Estatística

		Observado	
		$Y = 1$	$Y = 0$
Predito	$\hat{Y} = 1$	5	0
	$\hat{Y} = 0$	6	17

A acurácia do modelo no ponto de corte 0,6 é de 78,67%, a sensibilidade é de 45,45% e a especificidade é de 100%. O modelo prediz muitos falsos negativos, ou seja, classifica como não evadidas as alunas que de fato evadiram o curso.

Antes da interpretação dos parâmetros, o modelo será ajustado aos dados completos para se obter estimativas mais precisas dos coeficientes.

Tabela 23: Modelo Final do Bacharelado em Estatística, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	-3,63398	1,33065	-2,731	0,006315
Taxa de Reprovação	0,05868	0,01078	5,443	$\leq$ 0,0001
Semestres cursados	-0,45165	0,15435	-2,926	0,003432
Trancamentos	1,44175	0,39292	3,669	0,000243
Idade	0,12179	0,06002	2,029	0,042446

As estimativas dos parâmetros do modelo ajustado com os dados completos são muito semelhantes às da amostra de construção, portanto o modelo escolhido é válido.

Tabela 24: Razão de chances e IC 95% - Modelo Final Estatística

Variável Explicativa	**Razão de Chances**	**IC 95%**
Taxa de Reprovação	1,060	1,040 - 1,086
Semestres cursados	0,637	0,454 - 0,834
Trancamentos	4,228	2,095 - 9,931
Idade	1,130	1,001 - 1,276

A cada 1% de acréscimo na taxa de reprovação, a chance de evasão aumenta em 6%. Isto é, uma aluna que possui 10% de taxa de reprovação a mais que outra aluna, tem chance de evasão 60% maior - dado que os outros fatores são iguais entre ambas.

A cada semestre a mais de permanência no curso, a chances de evasão diminuem em 36,3%. Quanto ao número de trancamentos, cada disciplina trancada pela aluna resulta em um aumento impressionante de 322,8% nas chances de evasão. Deve-se considerar, contudo, que pouquíssimas alunas trancam mais de uma disciplina ao longo de toda sua permanência no curso.

Por fim, a aumento de um ano na idade de ingresso no curso aumenta em 13% as chances de evasão, indicando que alunas mais velhas enfrentam maiores dificuldades para permanecer no curso, mantendo-se constantes os outros fatores.

4.3.3 Matemática

As variáveis inicialmente selecionadas para o estudo da evasão das alunas de Matemática foram: IRA, taxa de reprovação, idade, semestres cursados, trancamentos, menções SR, raça e sistema de cotas. Assim como ocorreu na modelagem da Estatística, as variáveis escola e forma de ingresso não foram incluídas na análise.

O IRA e a taxa de reprovação não são fortemente correlacionados, então foram ajustados tanto modelos com as variáveis separadas, quanto modelos com as variáveis juntas. O IRA, contudo, não se mostrou significativo em nenhum dos modelos testados. A seleção das variáveis se deu da mesma forma como descrita anteriormente, e o modelo final obtido foi:

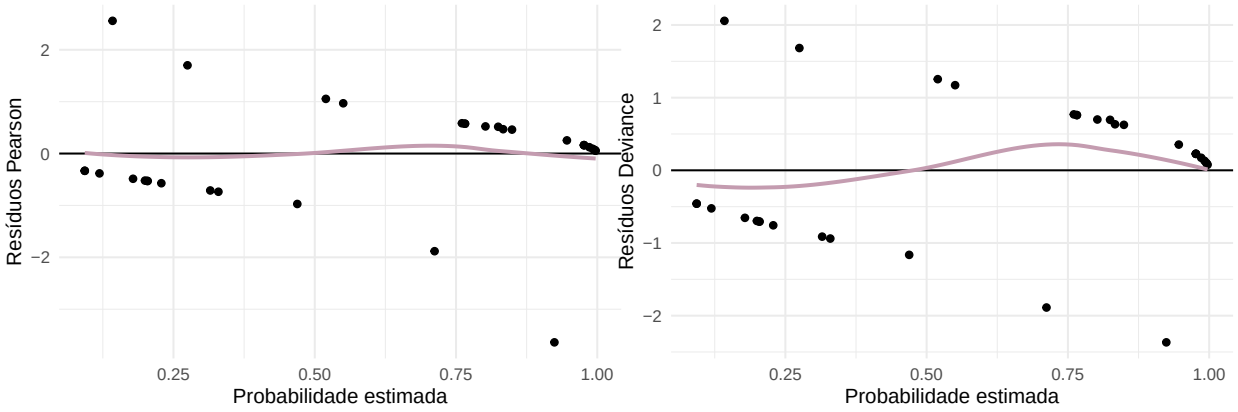
Tabela 25: Modelo do Bacharelado em Matemática, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	-2,27744	0,92513	-2,462	0,0138
Taxa de Reprovação	0,06004	0,02385	2,517	0,0118
Trancamentos	1,06102	0,51329	2,067	0,0387

Com apenas duas variáveis explicativas, o modelo final da Matemática leva em conta a taxa de reprovação das alunas e total de trancamentos ao longo da permanência no curso. Nenhuma das outras variáveis, incluindo as interações entre elas, gerou um ganho significativo para o modelo. À semelhança da Estatística, a variável taxa de reprovação do modelo foi redimensionada para termos percentuais. O AIC é igual a 31,827.

Quanto ao ajuste do modelo escolhido, a curva na suavização de Lowess dos resíduos *deviance* é um indício de que o modelo não está muito bem ajustado, mas o número muito pequeno de observações torna esse resultado questionável.

Gráfico 15: Resíduos - Modelo Matemática



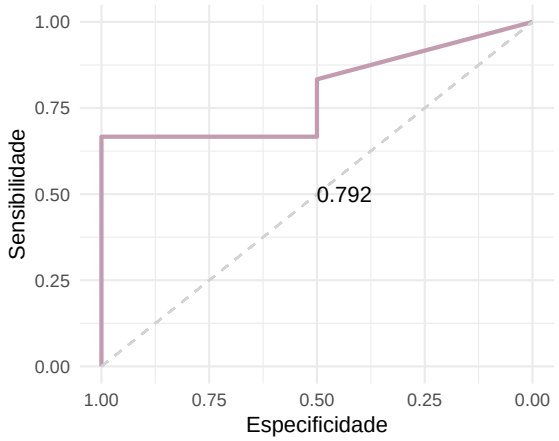
(a) Resíduos *Pearson* (b) Resíduos *Deviance*

Tabela 26: Testes de adequabilidade de ajuste - Modelo Matemática

Teste	**Estatística**	**p-valor**
Hosmer e Lemeshow	8,4569	0,390
Resíduos *Pearson*	30,449	0,443
Resíduos *Deviance*	25,927	0,684

Nenhum dos testes de adequabilidade rejeitam a hipótese de que o modelo está bem ajustado, portanto pode-se concluir que o modelo é adequado.

Gráfico 16: Curva ROC - Modelo Matemática



A área abaixo da curva ROC (0,792) indica que a discriminação do modelo é aceitável, porém, a julgar pela quantidade limitadíssima de observações usadas para ajustar o modelo, esse poder preditivo pode ser considerado bem satisfatório. O melhor

ponto de corte das probabilidades preditas é 0,55, ponto em que se obtém as seguintes classificações:

Tabela 27: Matriz de Confusão - Modelo Matemática

		Observado	
		$Y = 1$	$Y = 0$
Predito	$\hat{Y} = 1$	4	0
	$\hat{Y} = 0$	2	4

A acurácia do modelo no ponto de corte 0,55 é de 80%, a sensibilidade é de 66,67% e a especificidade é de 100%. Isto é, os únicos equívocos de classificação do modelo são falsos negativos, ou seja, classificar como não evadidas as alunas que de fato evadiram o curso.

A seguir, serão estimados os coeficientes do modelo com base nos dados completos.

Tabela 28: Modelo Final do Bacharelado em Matemática, 2011-2019

Parâmetro	**Estimativa**	**Erro padrão**	**Estatística**	**p-valor**
Intercepto	-1,68825	0,65645	-2,572	0,01012
Taxa de Reprovação	0,04967	0,01812	2,741	0,00613
Trancamentos	0,99354	0,42714	2,326	0,02002

Verifica-se que os parâmetros dos modelos ajustados com os dados completos e com a amostra de construção são bem parecidos, então pode-se concluir que o modelo selecionado é válido.

Tabela 29: Razão de chances e IC 95% - Modelo Final Matemática

Variável Explicativa	**Razão de Chances**	**IC 95%**
Taxa de Reprovação	1,051	1,021 - 1,100
Trancamentos	2,701	1,292 - 7,241

A cada 1% a mais na taxa de reprovação, a chance de evasão cresce 5,1%. Em relação ao número de trancamentos, cada disciplina trancada provoca um aumento de 170% nas chances de evasão.

5 Conclusão

Os cursos de Bacharelado do Instituo de Ciências Exatas (IE) são predominantemente masculinos, porém os dados revelaram que não há diferença na taxa de evasão entre os gêneros, desafiado o proposto por Chen (2013) e Ellis, Fosdick e Rasmussen (2016). Entre as estudantes mulheres ingressantes nos cursos de Ciência da Computação, Estatística e Matemática no período de 2011 a 2019, as taxas de evasão foram de 39,77%, 39,71% e 60,47%, respectivamente. Se desconsideradas as alunas ainda ativas nos cursos, as taxas sobem para 77,78%, 76,06% e 74,29%.

Na análise dos perfis sociais e acadêmicos das alunas, muitos resultados interessantes foram encontrados. Verificou-se que há uma grande diferença entre o curso de Matemática e os cursos de Ciência da Computação e Estatística quanto à forma de ingresso. Nos dois últimos, as principais vias de ingresso no curso são o PAS e o Vestibular da UnB, ambos sistemas avaliativos, enquanto na Matemática as alunas ingressam principalmente por mudança de curso ou duplo curso. Também foi identificado que o maior percentual de alunas cotistas é do curso de Estatística, 41,91%, e o menor é da Matemática, com apenas 11,63%.

Outro resultado que se destaca é a dicotomia a respeito das idades das alunas da Estatística - apesar de ser o curso com o perfil mais jovem no IE, é ao mesmo tempo o curso com maior número de alunas com mais de 30 anos de idade. Isso indica que a Estatística atrai tanto alunas recém-saídas do Ensino Médio, como também alunas já formadas em busca de uma segunda graduação.

Quanto à evasão, os resultados mostraram que em todos os três cursos de bacharelado a principal forma de saída entre alunas evadias é o desligamento institucional, seja por abandono, por rendimento ou voluntário. Um achado bastante notável do estudo foi a percepção de que um histórico de evasões nem sempre é fator determinante para evasões futuras. Isto é, só porque uma aluna evadiu o curso anteriormente, não significa que irá evadir novamente em uma nova tentativa - é o caso, por exemplo, de uma estudante da Matemática que chegou a evadir o curso duas vezes antes de conseguir se formar. Nos cursos da Ciência da Computação e da Matemática, alunas cotistas evadem menos que as alunas não cotistas - resultado também bastante relevante, principalmente à luz das discussões a respeito da implementação de programas de cotas dentro das universidades.

Outra descoberta curiosa diz respeito ao total de disciplinas trancadas pelas alunas que evadiram os cursos. Na Estatística e na Matemática, as alunas evadidas acumulam

uma quantidade maior de trancamentos em comparação às não evadias, porém na Ciência da Computação o oposto é observado. Ainda com relação aos principais achados, a análise descritiva evidenciou que as informações acadêmicas das alunas - IRA, taxa de reprovação, menções SR, etc. - possuem o melhor potencial de discriminação entre alunas evadidas e não evadidas, dentre as variáveis explicativas consideradas.

Isso se refletiu na modelagem, em que parâmetros de variáveis acadêmicas foram predominantes. Na Ciência da Computação, as variáveis explicativas currículo, escola, IRA e semestres cursados se mostraram as mais relevantes para explicar a evasão do curso. Destacam-se os efeitos do currículo novo e do egresso de escola pública, que reduzem em 92,7% e 86,3% as chances de evasão, respectivamente, em relação ao currículo antigo e o egresso de escola particular. O poder preditivo desse modelo é excelente, com acurácia de 77,78%.

No curso de Estatística, as variáveis mais significativas no estudo da evasão de alunas foram a taxa de reprovação, o total de semestres cursados, a quantidade de trancamentos e a idade. Em especial, foi estimado que cada disciplina trancada pela aluna resulta em um aumento de 322,8% nas chances de evasão. Deve-se levar em conta, porém, que esse resultado pode ter sido afetado pelo número pequeno de observações na amostra. O poder preditivo obtido por esse modelo é aceitável, com acurácia de 78,67%.

Na Matemática apenas duas variáveis contribuem significativamente para a modelagem das evasões no curso - taxa de reprovação e total de trancamentos. Assim como no modelo da Estatística, o impacto do trancamento de disciplinas é o que mais se sobressai. Cada matéria trancada aumenta em 170% as chances de evasão. O modelo obteve bom poder preditivo, com acurácia de 80%.

Os achados finais destacam o impacto do desempenho acadêmico na evasão de estudantes mulheres dos cursos de Bacharelado do IE, contudo isso não significa que fatores pessoais, culturais e sociais não têm um papel importante nesse fenômeno. Os bancos de dados utilizados nesse trabalho privilegiam informações a respeito da vida acadêmica das alunas, e os resultados descobertos ecoam as restrições dos dados disponíveis.

Antes de concluir, é importante pontuar as limitações encontradas ao longo desse estudo. Em primeiro lugar, a quantidade de observações nas bancos finais não é ideal para o ajuste de modelos de regressão logística, uma vez que essa técnica depende de muitos dados amostrais para ter estimativas mais precisas. Dificuldades relacionadas ao tamanho da amostra ocorreram princialmente durante a modelagem dos dados da Matemática, cujo banco possui apenas 45 observações ao todo.

Em segundo lugar, as bases de dados originais extraídas dos sistemas utilizados pela Universidade de Brasília apresentaram algumas inconsistências, com divergências de padronização, muitos dados faltantes ou equivocados e informações duplicadas. Lidar com as especificidades dos dados de cada sistema foi bastante desafiador e demandou muito tempo de dedicação.

Como sugestão para estudos futuros, eis algumas propostas interessantes de serem exploradas:

- Ampliar o corte temporal para obter mais observações;
- Fazer um estudo comparativo com os alunos do gênero masculino;
- Criar uma variável que indique o avanço da aluna no curso, com base no fluxograma proposto pelos departamentos;
- Adicionar variáveis explicativas com informações socioeconômicas das alunas, viabilizando o estudo dos fatores individuais relacionados à evasão;
- Explorar técnicas estatísticas que lidem bem com quantidades limitadas de observações.

Referências

AGRESTI, A. *An Introduction to Categorical Data Analysis, 3rd edition*. [S.l.]: John Wiley and Sons Inc, New York, 2019.

ANDIFES; ABRUEM; SESU/MEC. Diplomação, retenção e evasão nos cursos de graduação em instituições de ensino superior públicas. *Comissão Especial de Estudo sobre a Evasão nas Universidades Públicas Brasileiras*, 1996.

CHEN, X. Stem attrition: college students' paths into and out of stem fields. *National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC*, 2013.

CORRELL, S. J. Gender and the career choice process: The role of biased self-assessments. *American Journal of Sociology*, 2001.

CÔRTES, G. L. Estudo da evasão no curso de engenharia da computação da universidade de brasília. *Trabalho de Conclusão de Curso (Bacharelado em Estatística), Universidade de Brasília, Brasília*, 2023.

ELLIS, J.; FOSDICK, B.; RASMUSSEN, C. Women 1.5 times more likely to leave stem pipeline after calculus compared to men: Lack of mathematical confidence a potential culprit. *PLoS One*, 2016.

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

KAHLE, D.; WICKHAM, H. ggmap: Spatial visualization with ggplot2. *The R Journal*, v. 5, n. 1, p. 144–161, 2013. Disponível em: ⟨https://journal.r-project.org/archive/2013-1/kahle-wickham.pdf⟩.

NETER, J. et al. *Applied Linear Statistical Models*. [S.l.]: The McGraw-Hill Irwin Inc, New York, 2004.

PINTO, L. G. Estudo sobre evasão acadêmica no curso de licenciatura em computação da universidade de brasília : uma aplicação de regressão logística. *Trabalho de Conclusão de Curso (Bacharelado em Estatística), Universidade de Brasília, Brasília*, 2022. Disponível em: ⟨https://bdm.unb.br/handle/10483/34701⟩.

SEMESP, I. Mapa do ensino superior no brasil. 10ª edição. 2020.

UNESCO. Decifrar o código: educação de meninas e mulheres em ciências, tecnologia, engenharia e matemática (stem). *Brasília*, 2018.