Universidade de Brasília
Departamento de Estatística

Estudo sobre Evasão Acadêmica no Curso de Licenciatura em Computação da Universidade de Brasília: uma aplicação de Regressão Logística

Larissa Gomes Pinto

Brasília
2022

Larissa Gomes Pinto

Estudo sobre Evasão Acadêmica no Curso de Licenciatura em Computação 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
2022

À Deus, pois o Seu jugo é suave e Seu fardo é leve.

Resumo

A evasão acadêmica é um dos problemas que mais atingem as instituições de ensino superior no Brasil e no mundo. O conceito de evasão, utilizado nesse trabalho, consiste no aluno que sai do curso que ingressou por alguma forma diferente da formatura. Esse trabalho busca então identificar quais são as características dos estudantes que estão associadas a evasão do curso de Licenciatura em Computação da Universidade de Brasília utilizando o método de análise de regressão logística. O conjunto de dados é composto por 727 observações fornecidas pelos sistemas de informação da Universidade de Brasília com as características dos alunos que se matricularam no curso durante o período de 2012/2 a 2019/2. Fatores ligados ao desempenho acadêmico e a características socioeconômicas foram avaliados na construção de dois modelos para descreverem o fenômeno da evasão, alguns fatores foram significativos, entre eles o currículo vigente no ingresso do aluno, a forma de ingresso na universidade, além do desempenho acadêmico.

Palavras-chaves: evasão, regressão logística, Licenciatura em Computação.

Lista de Tabelas

Lista de Gráficos

Sumário

1 Introdução

A evasão acadêmica é um dos problemas que mais atingem as instituições de ensino superior no Brasil e no mundo. É fácil de ser observado que, costumeiramente, o número de formandos é consideravelmente menor que o número de ingressantes em um curso superior.

O conceito de evasão, em sua forma mais simples, consiste no aluno que sai do curso que ingressou por alguma forma diferente da formatura. Esse conceito engloba tanto a evasão do curso em si, como a evasão da instituição de ensino e a evasão do ensino superior em geral. Esses diferentes modos de evadir podem dificultar a mensuração. O Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP, 1998) ainda diferencia o conceito de abandono e evasão. O primeiro refere-se ao aluno que sai da instituição de ensino por um ano e retorna no seguinte. Já o segundo, sai e não retorna mais ao sistema. Apesar das várias definições de evasão universitária, o conceito de evasão abordado nesse trabalho é o de evasão de curso.

De acordo com o Mapa do Ensino Superior do Brasil do Instituto Semesp (2020), a taxa de evasão do ensino superior foi de 26,5% em 2018. Para as instituições de públicas, essa taxa ficou em 18,5%. A evasão acadêmica é um problema que afeta diversas perspectivas: quando um aluno inicia um curso e não conclui, desperdícios pessoais, sociais e econômicos são gerados.

É muito importante que as instituições de ensino superior tal como a Universidade de Brasília (UnB) busquem compreender o fenômeno de evasão. Entender quais fatores estão influenciando na evasão dentro de uma universidade dá a possibilidade de que as instituições identifiquem os aspectos que possam ser trabalhados no sentido de evitar a evasão e colaborar para os estudantes não apenas ingressem na universidade, mas se formem. Assim, contribuir para que a universidade cumpra com sua missão na formação de profissionais e pesquisadores para a sociedade.

Outro aspecto relacionado a evasão acadêmica é que ela também gera um impacto orçamentário, uma vez que parte expressiva da matriz orçamentária de uma instituição de ensino está relacionada a quantidade de alunos matriculados. Nas instituições públicas, entender o fenômeno pode contribuir significativamente na diminuição de desperdícios de recursos públicos.

Os fatores que levam a evasão são muito diversos. De acordo com Barroso e Falcão (2004) as causas que influenciam na evasão são divididas em três questões: econômica

- onde questões socioeconômicas impossibilitam a permanência; vocacional – não identificação com o curso; e institucional - o fracasso nas disciplinas iniciais, seja por deficiências prévias de conteúdo, seja por inadequação nos métodos de estudo, seja por dificuldades de relacionamento com os demais colegas e com a instituição. Segundo Braga, Carmo e Bogutchi (2003) os cursos com maiores índices de evasão são geralmente aqueles que possuem reprovações mais elevadas nos anos iniciais.

Nesse trabalho, há o interesse de estudar a evasão acadêmica no curso de Licenciatura em Computação da Universidade de Brasília. O curso de Licenciatura em Computação é um curso noturno, presencial, que confere o grau de licenciado. O curso visa atender à necessidade imediata de informatização e absorção dos avanços dessa área nos diversos setores da sociedade, formando um profissional educador de Licenciatura em Computação tanto para escolas quanto para grandes corporações. Para isso, curso apresenta uma proposta de formação que integra as áreas de educação e computação. Para graduar-se nesse curso, o aluno deve cursar 180 créditos distribuídos em 9 semestres.

Nesse cenário, busca-se então identificar quais são as características dos estudantes estão associadas a evasão acadêmica no curso de Licenciatura em Computação da Universidade de Brasília. Assim, como a característica de interesse é saber se o aluno evadiu ou não, uma variável resposta qualitativa binária, um modelo estatístico adequado para modelar esse fenômeno é a regressão logística. Estudos utilizando o método de análise de regressão logística são de grande utilidade para compreender a relação entre a evasão e as características dos estudantes e, assim, traçar o perfil dos alunos evadidos do curso Licenciatura em Computação da UnB e obter um modelo para descrever essa relação.

2 Referencial Teórico

2.1 Evasão

Apesar do processo de expansão ocorrido no sistema de educação superior no Brasil, ainda é possível observar altos índices de evasão nessa modalidade de ensino. A evasão de alunos de cursos de ensino superior é um problema a ser enfrentado no Brasil e no mundo.

O conceito de evasão pode parecer à primeira vista bem sucinto e autoexplicativo, sendo explicado como o fenômeno de saída do aluno do curso ingressado antes da formatura. Porém, essa definição não é consensual entre autores, podendo provocar dificuldades de mensuração e interpretação do fenômeno.

Quando se fala em evasão, o modelo teórico mais citado na literatura é o desenvolvido por Tinto (1975). Nesse é descrito um conjunto de fatores que podem influenciar no processo de evasão do estudante, sendo esses atributos individuais e familiares, habilidades prévias a entrada no ensino superior, integração social e acadêmica dentro da universidade, comprometimento individual, comprometimento da instituição, fatores sociais e familiares externos ao aprendizado acadêmico.

Por não haver uma definição consensual a quanto aos tipos de evasão no ensino superior, a Comissão Especial de Estudos sobre a Evasão no Brasil (1997) distingue a evasão em três tipos: a de curso, a de instituição e a do ensino superior.

2.1.1 Evasão de Curso

A evasão de curso é aquela onde ocorre a saída definitiva do aluno do seu curso de origem sem concluí-lo, por qualquer motivo. Ela aborda o fator de mudança de curso, que não significa necessariamente uma mudança de instituição de ensino ou até mesmo a desistência do ensino superior. A evasão de curso engloba diversas situações como o abandono (onde o aluno deixa se matricular-se), a desistência, a transferência interna (mudança de curso) e a exclusão por norma institucional.

Esse tipo de evasão ocorre muitas vezes pela não identificação do estudante com o curso. Estudantes que ingressam no curso realmente desejado são menos propensos a evadir (MCMILLAN, 2005). Um fenômeno recorrente é o de alunos que ingressam em um determinado curso almejando outro. Os alunos ingressam em um curso de menor

concorrência com o objetivo de se transferir futuramente para o curso desejado (DIAS; THEÓPHILO; LOPES, 2010).

2.1.2 Evasão de Instituição

Esse tipo de evasão ocorre quando o estudante se desliga da instituição na qual está matriculado, mas permanece no sistema de ensino superior. Aponta-se aqui diferença entre a evasão de curso e a evasão de instituição. Os alunos que mudam de curso, e, portanto, se encaixam no conceito de evasão de curso, mas continuam na mesma instituição de ensino são desconsiderados na contagem de ingressantes.

Problemas relacionados com a evasão da instituição provavelmente serão mais graves que os relacionados com a evasão de curso, pois na evasão de curso ainda haveria algum curso dentro da instituição onde o aluno se adaptaria; na evasão da instituição, isso não ocorreu e aluno deixa a instituição de ensino. Questões geográficas relacionadas com o deslocamento e a distância até a instituição de ensino podem estar relacionadas em algum grau com a motivação para a evasão de uma instituição de ensino superior.

2.1.3 Evasão do Ensino Superior

A evasão do ensino superior ocorre quando o aluno abandona de forma definitiva o sistema de ensino superior, deixando assim de estar vinculado a qualquer instituição de ensino superior.

Comparado aos outros tipos de evasão, a evasão do ensino superior é o tipo mais difícil de rastrear, pois o estudante deixa a instituição sem pedir transferência como também não passa por outro processo seletivo, gerando assim menos informações sobre o fenômeno.

2.1.4 Fatores Relacionados a Evasão

Assim como Barroso e Falcão (2004) discutem os fatores relacionados a evasão em três questões: socioeconômica, vocacional e institucional, a Comissão Especial de Estudos sobre a Evasão no Brasil (1997) classifica esses fatores associados a evasão em três formas, relacionando-as ao próprio estudante, ao curso ou instituição e os fatores socioeconômicos externos.

Dos fatores associados ao estudante em si, percebe-se que estão muito ligados

principalmente às dificuldades escolares que o acompanham durante a vida e aqueles relacionados aos seus desejos pessoais. Muitas vezes tais fatores o levam a descobrir que escolheram o curso errado e acabam por tentar realizar mudança para outro.

Em relação aos fatores institucionais relacionados a evasão, currículos extensos e rígidos, os critérios de avaliação e a formação pedagógica e docente contribuem para o desinteresse pelo curso. Já os fatores externos podem ser citados aqueles relativos à preocupação com mercado de trabalho, o reconhecimento da carreira e até a desvalorização da profissão, como no caso das licenciaturas. É muito comum a preocupação de como se encontra o mercado de trabalho voltado a um curso. Muitas vezes, mesmo o aluno se identificando com o curso, por ordem financeira, ele acaba sendo levado a evadir e mudar de curso.

2.2 Regressão Logística

Os modelos de regressão são uma das formas de analisar a relação entre uma variável de interesse, denominada variável resposta, com outros fatores, denominados variáveis explicativas ou independentes. O modelo de regressão logística é utilizado quando a variável resposta é qualitativa, isto é, possui dois ou mais resultados possíveis. Com a técnica de regressão logística é possível estimar a probabilidade associada à ocorrência de um determinado evento a partir de um conjunto de variáveis explicativas.

O modelo de regressão logística binária é um caso particular dos modelos lineares generalizados, onde a variável resposta é qualitativa e binária, ou seja, apresenta dois resultados possíveis: sucesso ou fracasso. Em uma regressão linear existem pressupostos a serem atendidos. Dado que a variável resposta estabelecida é dicotômica, alguns problemas surgem:

- Não há normalidade do erro;
- Variância do erro não é constante;
- Há uma limitação na função resposta, isto é, dado que a variável Y é binária e a esperança ser uma probabilidade, a resposta média deve seguir $0 \leq E(Y) = \pi \leq 1$.

Sendo assim, os pressupostos do modelo de regressão linear não são atendidos e neste caso o modelo de regressão logística se mostra mais adequado. Na regressão logística, a variável resposta Y tem distribuição Bernoulli, sendo π e $1 - \pi$ as probabilidades de sucesso e fracasso, respectivamente. O modelo de regressão logística é definido por:

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}{1 + exp(\beta_0 + \beta_1 X_1 + ... + \beta_p X_p)}, \tag{2.2.1}$$

onde $\pi(x) = E(Y|X)$, p é o número de variáveis explicativas (X_i) incluídas no modelo, $\beta_0 + \beta_1 X_1 + ... + \beta_p X_p$ o preditor linear e os β_i's os parâmetros do modelo.

É importante observar que ao aplicar a seguinte tranformação em $\pi(x)$, designada transformação logito, tem-se que:

$$logito(\pi(x)) = ln\left[\frac{\pi}{1-\pi}\right] = \beta_0 + \beta_1 X_1 + ... + \beta_p X_p, \tag{2.2.2}$$

sendo o logito de $\pi(x)$ linear nos parâmetros.

2.2.1 Estimação dos Parâmetros do Modelo

Em regressão logística, as estimativas para os parâmetros β_k do modelo podem ser obtidas através do método de máxima verossimilhança. Dado que, em uma amostra aleatória de tamanho n, cada Y_i segue uma distribuição de Bernoulli com probabilidade de sucesso π_i e fracasso $1 - \pi_i$ com independência das observações, a função de máxima verossimilhança é dada por

$$g(Y_1, ..., Y_n) = \prod_{i=1}^{n} f_i(Y_i) = \prod_{i=1}^{n} \pi_i^{Y_i} \, (1 - \pi_i)^{1-Y_i}, \tag{2.2.3}$$

aplicando o logaritmo dessa expressão, uma vez que é mais fácil de trabalhá-la matematicamente, tem-se o log da verossimilhança como

$$ln\ g(Y_1, ..., Y_n) = ln \prod_{i=1}^{n} \pi_i^{Y_i} (1 - \pi_i)^{1-Y_i} = \sum_{i=1}^{n} \left[Y_i \ ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} ln\ (1 - \pi_i), \tag{2.2.4}$$

e, dado que $1 - \pi_i = [1 + exp(\beta_0 + \beta_1 X_i)]^{-1}$, obtém-se que a função de log-máxima verossimilhança é definida como

$$ln\ L(\beta) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_{i+...+\beta_p X_i}) - \sum_{i=1}^{n} ln[1 + exp(\beta_0 + \beta_1 X_i + ... + \beta_p X_i)]. \tag{2.2.5}$$

Derivando a expressão acima em relação a cada parâmetro do modelo, serão obtidas p+1 equações de máxima verossimilhança para os $p + 1$ parâmetros. Não há uma fórmula fechada para os valores de β que maximizem a função de máxima veros-

similhança. Para encontrar estimativas de máxima verossimilhança é necessário utilizar métodos numéricos de solução, como o método de Newton-Raphson.

Uma vez obtido as estimativas de máxima verossimilhança para β_{p+1} parâmetros, tem-se a função resposta ajustada denotada por

$$\hat{\pi_i} = \frac{exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \hat{\beta}_p X_i)}{1 + exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + ... + \beta_p \hat{X}_i)}. \tag{2.2.6}$$

2.2.2 Interpretação dos Parâmetros

A interpretação das estimativas dos parâmetros do modelo de regressão logística não é semelhante a dos coeficientes em um modelo de regressão linear. Uma interpretação de β_k pode ser encontrada a partir do $logito(\pi_i)$ definido em (3.2.2) que é uma função da $odds$[1] $= \frac{\pi(x)}{1-\pi(x)}$ no ponto x, ou seja,

$$logito(\pi(x)) = \frac{\pi(x)}{[1 - \pi(x)]} = odds = e^{\beta_0 + \beta_1 x_1 + ... + \beta_p x_p}. \tag{2.2.7}$$

Assim, tomando dois valores distintos de uma variável explicativa x_j e x_{j+1}, e considerando $odds_1 = \pi_1/(1 - \pi_1)$ e $odds_2 = \pi_2/(1 - \pi_2)$, tem-se que

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

Aplicando o exponencial de ambos os lados, obtém-se a razão estimada de duas chances, definida como *odds ratio*[2] ou razão de chances (OR), denotado como

$$OR = \frac{odds_2}{odds_1} = e^{\beta_1}. \tag{2.2.8}$$

Desta maneira, tomando como referência os valores de uma variável explicativa X_i, pode-se obter a chance de sucesso de um grupo em relação a outro utilizando a função *odds ratio*, em que cada um desses dois grupos possuem valores diferentes da mesma variável X_i.

Dessa forma, quando um $\beta_k > 1$, então OR > 1, logo as chances de sucesso dos indivíduos x_{j+1} são maiores que as dos indivíduos x_j. Por outro lado, quando OR < 1, as chances de sucesso dos indivíduos x_{j+1} são menores que as dos x_j.

[1] *Odds* ou chance é definida como a razão a probabilidade de sucesso e a probabilidade de insucesso.
[2] *Odds ratio* é a razão entre duas chances (*odds).*

Para variáveis explicativas quantitativas, β_k representa o efeito médio de uma unidade a mais da variável correspondente ao parâmetro na variável resposta.

2.2.3 Intervalo de Confiança para os Parâmetros do Modelo

A partir da estimativa pontual do parâmetro pode-se contruir uma estimativa intervalar para o parâmetro com confiança de $1-\alpha$. O intervalo de confiança de $1-\alpha$ para um parâmetro β_k é dado por

$$\hat{\beta}_k \mp z_{1-\frac{\alpha}{2}} s\{\hat{\beta}_k\}, \tag{2.2.9}$$

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

2.3 Testes de Significância do Modelo

Após estimar os parâmetros do modelo, é importante avaliar a significância das variáveis do modelo. Para isso, utiliza-se alguns testes para verificar se as variáveis explicativas são significativamente relacionadas com a variável dependente, isto é, a variável resposta. Para avaliar essa relação no modelo logístico, os testes de Razão de Verossimilhança e o de Wald são os mais utilizados.

2.3.1 Teste de Razão de Verossimilhança

O teste da razão de verossimilhança é usado para testar a significância dos p parâmetros estimados das variáveis independentes do modelo. Esse teste busca avaliar se todos os parâmetros β's associados as variáveis independentes são nulos, o que indicaria a não existência de regressão. As hipóteses a serem testadas são:

$$\begin{cases} H_0 : \beta_1 = ... = \beta_p = 0 \\ H_1 : \beta_j \neq 0 \; para \; algum \; j. \end{cases}$$

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

$$G^2 = -2ln\left[\frac{L(R)}{L(F)}\right] = -2[ln\ L(R) - ln\ L(F)], \tag{2.3.1}$$

onde L(R) e L(F) são os valores da função de máxima verossimilhança para os modelos

reduzido e completo, respectivamente. Considera-se como modelo reduzido quando a hipótese nula é verdadeira.

A estatística G^2, para n grande, segue uma distribuição Qui-quadrado com graus de liberdade correspondente a diferença no número de parâmetros entre o modelo reduzido e o completo, $p - q$.

Assim, a regra de decisão é dada por

$$Se\ G^2 \leq \chi^2(1-\alpha; p-q),\ aceita-se\ H_0;$$
$$Se\ G^2 > \chi^2(1-\alpha; p-q),\ rejeita-se\ H_0.$$

O teste de Razão de Verossimilhança também pode ser utilizado para verificar se vários β's do modelo são nulos. Nesse caso, as hipóteses testadas são:

$$\begin{cases} H_0 : \beta_q = \beta_{q+1} = ... = \beta_{p-1} = 0 \\ H_1 : Nem\ todos\ \beta_k\ em\ H_0\ são\ iguais\ a\ 0. \end{cases}$$

Por conveniência, o modelo é ordenado de forma que os últimos $p-q$ coeficientes são aqueles a serem testados. Nesse caso, p representa o número de parâmetros no modelo completo e q do modelo reduzido.

2.3.2 Teste de Wald

O teste de Wald também é utilizado para testar a significância dos coeficientes do modelo, testando se cada coeficiente é diferente de zero. As hipóteses do teste são dadas por

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

E a estatística do teste obtida por

$$Z_k = \frac{\hat{\beta}_k}{s(\hat{\beta}_k)}. \tag{2.3.2}$$

O teste de Wald considera a razão da estimativa de máxima verossimilhança do parâmetro β_k com a estimativa do seu erro padrão dado por $s(\hat{\beta}_k)$. Sob a hipótese nula, Z_k segue uma distribuição aproximadamente Normal Padrão para n grande.

Se H_0 não for rejeitada, o teste de Wald indica que a variável X_k não influencia

significativamente na variável resposta.

2.4 Seleção do Modelo

Assim como toda regressão, em regressão logística também é necessário selecionar um modelo. A seleção se torna mais difícil quanto mais possíveis variáveis independentes o modelo possa ter. Na construção de modelos estatísticos tem-se como objetivo selecionar as variáveis que resultem no "melhor" modelo possível dentro do contexto do problema e que seja de relativamente fácil interpretação mas que ainda seja complexo o suficiente para ajustar os dados.

Existem alguns critérios que são utilizados em métodos de seleção para selecionar modelo como o Critério de Informação Akaike (AIC) e o Critério de Informação Bayesiano (BIC). Esses critérios fornecem o quão próximo os valores ajustados do modelo se aproximam dos verdadeiros valores médios, no termo de valor esperado, aplicando uma penalidade a modelos com maior número de parâmetros.

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

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

onde log L(b) é a log-verossimilhança e p o número de parâmetros.

2.4.1 Métodos de Seleção das Variáveis para o Modelo

Existem vários procedimentos que podem ser seguidos para a seleção de variáveis. A abordagem estatística é buscar por um modelo parcimonioso que descreva o fenômeno. (HOSMER; LEMESHOW, 2000).

- *Forward*

  Esse procedimento parte da inclusão de uma variável independente, começando por aquela que possui maior correlação com a variável resposta. A ideia do método é adicionar uma variável de cada vez de acordo com o incremento de cada uma a razão de verossimilhança do modelo até que a adição de uma nova variável não melhore o modelo. Uma vez que a variável foi selecionada e incluída por ser significativa, ela não deve mais ser excluída.

- *Backward*

  O *Backward* realiza o procedimento inverso do *Forward*. Aqui, começa inicialmente um ajuste com todas as variáveis explicativas e a cada passo, é testado se cada variável deve ser eliminada do modelo.

- *Stepwise*

  A seleção *Stepwise* é uma modificação do *Forward* onde a cada passo, todas as variáveis do modelo são verificadas para identificar quais devem permanecer no modelo e quais podem ser excluídas dele. A cada passo, tenta-se incluir uma nova variável. Caso entre, tenta-se a eliminação das que já estão no modelo. O procedimento acaba quando não se consegue nem adicionar, nem eliminar variáveis.

2.5 Técnicas de Diagnóstico

2.5.1 Testes de Adequabilidade de Ajuste

Verificar a adequabilidade do ajuste de um modelo estatístico serve analisar o quão bem ele se ajusta a um conjunto de observações. Medidas de adequabilidade de ajuste servem para medir a distância entre os valores observados e os valores esperados.

1. **Teste de Qui-Quadrado de Pearson**

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

   As hipóteses testadas são:

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

   Com as estatística do teste definida por

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

   Se a função resposta do modelo logístico é apropriada, X^2 segue aproximadamente uma distribuição Qui-Quadrado com $c - p$ graus de liberdade.

Adotando um nível de significância, a regra de decisão é dada por

$$Se\ X^2 \leq \chi^2(1-\alpha; c-p),\ aceita-se\ H_0;$$
$$Se\ X^2 > \chi^2(1-\alpha; c-p),\ rejeita-se\ H_0.$$

onde c é o número de combinações distintas dos preditores das variáveis.

2. **Teste Deviance de Adequabilidade**

Assim como no teste Qui-Quadrado, assume-se no teste *deviance* c distintas combinações dos preditores denotados por $X_1, ..., X_c$. Se a função resposta do modelo logístico é correta e n for grande, a *deviance* seguirá aproximadamente uma Qui-Quadrado com $c-p$ graus de liberdade. A estatística do teste *deviance* é dada por

$$G^2 = -2[ln\ L(R) - ln\ L(F)]$$
$$= -2\sum_{i=1}^{n}[Y_i ln(\hat{\pi}_i) + (1-Y_i)ln(1-\hat{\pi}_i)] = DEV(X_0, ..., X_{p-1}). \quad (2.5.2)$$

Valores grandes de *deviance* indicam que o modelo logístico ajustado não está correto. As hipóteses a serem testadas são as mesmas apresentadas no teste Qui-Quadrado de Pearson, com a regra de decisão dada por

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

3. **Teste de Hosmer-Lemeshow para Variáveis Explicativas Contínuas**

Hosmer e Lemeshow (2000) propõem utilizar os valores das probabilidades estimadas para agrupá-los. A finalidade desse teste é verificar se existem diferenças significativas entre os valores preditos pelo modelo e os observados.

As hipóteses do teste são:

$$\begin{cases} H_0 : E(Y) = \frac{exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}, & \text{ou seja, o modelo está bem ajustado} \\ H_1 : E(Y) \neq \frac{exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}{1+exp(\beta_0+\beta_1X_1+...+\beta_pX_p)}, & \text{ou seja, o modelo não está bem ajustado} \end{cases}$$

Para verificar o ajuste do modelo através desse teste, as probabilidades estimadas são ordenadas e a partir delas são criados g grupos que, por recomendação dos autores, costumam ser g = 10. Os grupos são formados de maneira que o grupo 1

tenha probabilidade predita entre 0,0 e 0,1 e o grupo 2 entre 0,1 e 0,2 e assim em diante até o grupo 10 com probabilidades preditas entre 0,9 e 1,0.

Depois da separação em grupos, são calculados os valores esperados para cada grupo e esses comparados com os valores observados utilizando a estatística qui-quadrado de Pearson. Através de simulação, Hosmer e Lemeshow mostraram que, para amostras grandes, a estatística do teste segue, aproximadamente, uma distribuição Qui-Quadrado com g-2 graus de liberdade, onde é o numero de grupos.

2.5.2 Análise de Resíduos

Em regressão logística é necessário realizar uma análise dos resíduos para avaliar a adequabilidade do modelo, analisando as distâncias entre os valores observados e estimados. Um modelo de regressão com boa adequabilidade apresenta pequena discrepância residual.

Em regressão logística, a variável resposta Y assume apenas dois valores, 0 ou 1. Assim, consequentemente, os resíduos também assumiram apenas dois valores, o que dificulta a análise. Com isso, a distribuição dos resíduos não pode ser considerada normal e, sob a suposição que o modelo ajustado é correto, ela é desconhecida.

Existem alguns tipos de resíduos que podem ser utilizados para verificar se há diferenças significativas entre os valores observados e estimativas e avaliar a qualidade do modelo como os resíduos de Pearson e os resíduos Deviance.

1. **Resíduos de Pearson**

   Os resíduos de Pearson calculam a razão entre a diferença de um valor observado e o seu valor estimado e o erro padrão estimado de Y_i.

   $$r_{Pi} = \frac{Y_i - \hat{\pi}_i}{\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)}} \tag{2.5.3}$$

2. **Resíduos Studentizados de Pearson**

   Os resíduos studentizados são a razão entre o resíduo de Pearson e a estimativa do seu desvio padrão. Entretanto, o desvio padrão não é constante, assim uma aproximação para o valor é dada por $\sqrt{\hat{\pi}_i(1-\hat{\pi}_i)(1-h_{ii})}$ onde h_{ii} é a i-ésima diagonal da matriz H chapéu.

   $$H = W^{1/2}X(X'WX)^{-1}X'W^{1/2}, \tag{2.5.4}$$

   onde W é uma matriz diagonal $n\ x\ n$ composta pelos elementos $\hat{\pi}_i(1-\hat{\pi}_i)$.

Assim, os resíduos studentizados de Peason são definidos como

$$r_{SPi} = \frac{r_{Pi}}{\sqrt{1 - h_{ii}}}. \tag{2.5.5}$$

3. **Resíduos *Deviance***

   Os resíduos Deviance medem a discordância entre a função de máxima verossimilhança observada e estimada. Para dados binários, o modelo *deviance* é dado por

$$DEV(X_0, ..., X_{p-1}) = -2\sum_{i=1}^{n}[Y_i ln(\hat{\pi}_i) + (1 - Y_i)ln(1 - \hat{\pi}_i)] \tag{2.5.6}$$

   O resíduo *deviance* para i, dev_i, é definido como a raiz da contribuição de cada i para o modelo DEV em (3.5.6).

$$dev_i = sign(Y_i - \hat{\pi}_i)\sqrt{-2\sum_{i=1}^{n}[Y_i log(\hat{\pi}_i) + (1 - Y_i)log(1 - \hat{\pi}_i)]}. \tag{2.5.7}$$

O sinal será positivo quando $Y_i \geq \hat{\pi}_i$ e negativo quando $Y_i \leq \hat{\pi}_i$. Assim, a soma dos resíduos *deviance* ao quadrado é igual ao modelo *deviance*.

$$\sum_{i=1}^{n}(dev_i)^2 = DEV(X_0, ..., X_{p-1}) \tag{2.5.8}$$

2.5.3 Identificação de Valores Influentes

É possível que uma única observação tenha grande influência no resultado de uma análise de regressão. Por isso, é importante detectar observações influentes e levá-las em consideração ao interpretar os resultados. A seguir são apresentadas algumas medidas utilizadas para avaliar a influência dos valores de uma amostra de dados.

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

   Seja X^2 e DEV as estatísticas Qui-Quadrado e *deviance* denotadas anteriormente e X_i^2 e DEV_i os valores das estatísticas do teste quando as observação de padrão i são deletadas. As mudanças nas estatísticas Qui-Quadrado e *Deviance* quando as observação de padrão i são excluídas, denotadas por delta Qui-quadrado e delta *deviance*, respectivamente, são definidas como

$$\Delta X_i^2 = X^2 - X_i^2,$$

$$\Delta dev_i = DEV - DEV_i.$$

Devido a recursos computacionais, comumente é utilizado as seguintes aproximações

$$\Delta X_i^2 = r_{SPi}^2, \tag{2.5.9}$$

$$\Delta dev_i = r_{SPi}^2 + dev_i^2. \tag{2.5.10}$$

A interpretação das estatísticas delta Qui-Quadrado e *deviance* não são simples. A decisão se um caso é ou não influente é geralmente feita através de uma avaliação visual através de gráficos. Geralmente, as estatísticas delta Qui-Quadrado e delta *deviance* são colocadas em um gráfico *versus* o padrão i ou *versus* $\hat{\pi}i$. Valores extremos aparecem como picos no gráfico *versus* o padrão i e como *outliers* nas partes superiores do gráfico *versus* $\hat{\pi}i$.

2. **Distância de Cook**

   A distância de Cook é utilizada para identificar observações influentes e medir o impacto da exclusão de análise de uma determinada observação. A medida é obtida como a diferença entre $\hat{\beta}$ e $\hat{\beta}_j$, que são as estimativas de máxima verossimilhança calculadas incluindo todos os j das variáveis e excluindo as observações com padrão j, padronizando essa diferença pela matriz de covariância de $\hat{\beta}$.

2.6 Qualidade do Ajuste

Assim como em regressão linear, as técnicas de qualidade do ajuste em regressão logística tentam verificar quão bem o modelo se ajusta aos dados. Geralmente são aplicadas depois que o modelo final foi selecionado.

2.6.1 A Curva ROC

A curva ROC (Receiver Operating Characteristic) é uma ferramenta de representação gráfica que ilustra o desempenho de um sistema de classificação binário, como o modelo de regressão logística. A curva fornece a probabilidade de detectar a sensibilidade e a especificidade para diferentes pontos de quebra.

A sensibilidade é definida como a probabilidade do teste fornecer um resultado positivo, dado que o indivíduo realmente sofreu o evento de interesse. Já a especificidade é

definida como a probabilidade do teste fornecer um resultado negativo quando o indivíduo não sofreu o evento de interesse.

A curva ROC é um gráfico de sensibilidade *versus* taxa de falsos positivos, ou seja, representa a sensibilidade (ordenadas) vs 1 – especificidade (abcissas) resultantes da variação de um valor de corte ao longo do eixo de decisão x.

A área embaixo da curva, que vai de zero a um, fornece uma medida para a habilidade que o modelo tem em discriminar em dois conjuntos quais experimentaram o evento de interesse e quais não.

Considerando R como o valor corresponde à área abaixo da curva, tem-se como regra geral:

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

2.6.2 Matriz de confusão e desempenho

Dada a AUC e um determinado ponto de quebra c, é possível construir uma matriz de confusão, utilizada em métricas para avaliar o desempenho de predição de um modelo estatístico. A matriz de confusão é uma tabela que mostra as frequências para cada classe predita e observada do modelo, sendo assim a matriz de confusão composta por:

- Verdadeiro positivo (VP): indivíduo classificado corretamente pelo modo como evento dado que é realmente um evento;
- Verdadeiro positivo (VP): indivíduo classificado corretamente pelo modo como evento dado que é realmente um evento;
- Verdadeiro negativo (FP): indivíduo classificado corretamente pelo modelo como não evento dado que é um não evento;
- Falso negativo (FN): indivíduo classificado incorretamente pelo modelo como não evento dado que é um evento.

Assim, a matriz de confusão é apresentada conforme a Tabela 1.

Tabela 1: Matriz de Confusão.

	Valores observados	
Valores Preditos	Evento	Não evento
Evento	VP	FP
Não evento	FN	VN

Existem algumas medidas úteis para avaliar o desempenho preditivo do modelo a partir da matriz de confusão. A acurácia é uma medida que fornece o quanto o modelo acertou em suas previsões, ou seja, é a razão entre a soma das previsões corretas sobre o somatório de todas as previsões.

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN}. \tag{2.6.1}$$

3 Metodologia

3.1 Banco de dados

Os dados utilizados nesse estudo foram extraídos do Sistema de Informações Acadêmicas de Graduação (SIGRA) e do Sistema Integrado de Gestão de Atividades Acadêmicas (SIGAA) contendo informações sociodemográficas e acadêmicas dos estudantes do Departamento de Ciências da Computação fornecidos pela Universidade de Brasília. O banco de dados fornecido contém com 24 variáveis sendo elas:

1. Índice de Rendimento Acadêmico - IRA[3];
2. Gênero;
3. Data de Nascimento;
4. CEP;
5. UF de Nascimento;
6. Sistema de Cotas: a UnB dispõe sistemas de cotas que podem ser utilizados para ingresso na universidade. A variável atribui *Sim* ao aluno que usou o sistema de cotas e *Não* ao aluno que não fez uso do sistema de cotas;
7. Tipo de cota: para os cotistas, identifica-se o sistema adotado da seguinte forma:
   - *Escola Pública Baixa Renda-Não PPI*
   - *Escola Pública Baixa Renda-PPI*
   - *Escola Pública Alta Renda-Não PPI*
   - *Escola Pública Alta Renda-PPI*
   - *Negro*;

   sendo PPI: Preto, Pardo ou Indígena;
8. Escola: refere-se se o aluno estudou em escola pública ou particular durante a educação básica. É uma informação fornecida pelo aluno ao ingressar na Universidade de Brasília e é indicado que responda tendo como base o ensino médio;
9. Chamada de Ingresso na UnB;
10. Período de Ingresso UnB;
11. Período de Ingresso no Curso;
12. Forma de Ingresso UnB;

[3]O cálculo do IRA está disponível em: https://www.boasvindas.unb.br/images/GUIA_CALOURO_2019.pdf

13. Período de Saída do Curso;
14. Forma de Saída do Curso;
15. Período que Cursou Disciplina;
16. Média Semestral do Aluno;
17. Mínimo de Créditos faltantes para Formatura;
18. Créditos no Período;
19. Total de Créditos Cursados;
20. Créditos Aprovados no Período;
21. Código da Disciplina;
22. Nome da Disciplina;
23. Créditos da Disciplina;
24. Menção na Disciplina.

Os dados do curso de Licenciatura em Computação têm 63.764 linhas e 24 variáveis, sendo que essas linhas não correspondem a um único aluno, visto que um aluno pode ter cursado várias disciplinas em diferentes semestres aparecendo assim diversas vezes nesse banco de dados. Para o curso de Licenciatura em Computação, há dados desde o primeiro semestre de 1997 até o segundo semestre de 2019.

Partindo dos dados do curso de Licenciatura em Computação, foi realizado uma limpeza e filtragem desse banco de dados a fim de construir a base para análise e modelagem. Para esse trabalho, foi delimitado como escopo de estudo o período de 2012/2 até o período de 2019/2. Esse recorte se faz necessário visto as mudanças no currículo do curso de Licenciatura em Computação e para que fosse mais representativo ao que é o curso atualmente. Para esse recorte de tempo, chegou-se a um banco com 727 alunos diferentes.

3.2 Criação de variáveis

Dadas as variáveis do banco de dados original, foram criadas algumas variáveis de interesse, assim como algumas variáveis que foram modificadas com agregação de categorias.

- **Evasão**
  Neste trabalho, tem-se como objetivo avaliar a evasão acadêmica no curso de Licenciatura em Computação da Universidade de Brasília e quais fatores contribuem para

esse fenômeno. Para a construção dessa variável foi utilizada como base a variável "forma de saída do curso".

Como dito anteriormente, foi considerado como evasão a saída definitiva do aluno do seu curso de origem sem concluí-lo, por qualquer motivo. Sendo assim, a variável evasão é dicotômica, recebendo dois valores: 0 para quando não houve evasão e 1 para quando houve evasão do curso.

A Tabela 2 apresenta como foram classificadas as formas de saída do curso de acordo com o conceito de evasão. As únicas formas classificadas como não evasão são os "ativos" que são os alunos que continuam matriculados no curso e os formados, uma vez que saíram do curso através da formatura.

Tabela 2: Formas de saída do curso de Licenciatura em Computação-UnB, 2012-2019.

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

- **Idade ao ingressar**

  A partir da variável data de nascimento foi criada a variável idade que leva em consideração a idade em anos do aluno quando ingressou no curso. Para isso, foi utilizada a variável período de ingresso no curso, tomando como referência o dia 1 de março para alunos que ingressaram no primeiro semestre de determinado ano e dia 1 de agosto para alunos que ingressaram no segundo semestre de terminado ano.

- **Número de Trancamentos**

  Essa variável surge a partir da contagem de quantas vezes um aluno i trancou alguma disciplina. Para isso, foi considerado como trancamento as categorias TJ e TR da variável Menção de disciplina. Para cada aluno, a criada variável Número

de trancamentos corresponde a soma de vezes que essas duas menções apareceram.

- **Taxa de Reprovação**

  A taxa de reprovação é uma variável construída para verificar a proporção de créditos com reprovação dentre o total de créditos cursados pelos estudantes. Para a construção da taxa de reprovação, foram utilizadas as variáveis "menção na disciplina"[4] e "créditos da disciplina" já contidas no banco original. Assim, para cada aluno foi considerado como taxa de reprovação:

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

  .

  A variável taxa de reprovação varia de 0 a 1, visto que os créditos reprovados fazem parte do total de créditos cursados.

- **Quantidade de menções SR**

  A variável "quantidade de menções SR" busca quantificar quantas vezes um aluno recebeu em alguma disciplina a menção SR. A percepção da quantidade de menções SR obtidas por um aluno é importante pois essa menção é atribuída quando há o abandono de uma disciplina, seja porque o aluno nunca compareceu, seja porque não foi alcançado o mínimo de presença. Assim, para cada aluno, a quantidade de menções SR a corresponde a soma das vezes que a menção SR foi atribuída.

- **Forma de Ingresso no Curso**

  A variável "forma de ingresso"já é original do banco de dados, porém, sofreu algumas alterações em suas categorias. Originalmente, os 727 alunos de Licenciatura em Computação se distribuíam-se conforme as formas de ingresso de acordo com a tabela a seguir.

[4]As menções atribuídas ao rendimento acadêmico do aluno em disciplina na UnB e sua equivalência numérica são: SS (9,0 a 10), MS (7,0 a 8,9), MM (5,0 a 6,9), MI (3,0 a 4,9), II (0,1 a 2,9) e SR zero. É aprovado na disciplina o aluno que obtiver menção igual ou superior a MM. Disponível em: https://www.boasvindas.unb.br/images/GUIA_CALOURO_2019.pdf

Tabela 3: Formas de ingresso em Licenciatura em Computação-UnB, 2012-2019.

Formas de ingresso	Percentual
Vestibular	43,33
Programa de Avaliação Seriada - PAS	24,76%
Sistema de Seleção Unificada - SISU	17,61%
Enem UnB	3,71%
Portador de Diploma de Curso Superior	8,39%
Transferência Facultativa	0,83%
Transferência Obrigatória	1,38%

Observa-se que há categorias com baixas frequências relativas e poderiam ser agrupadas com outra forma de ingresso semelhante para uma melhor análise descritiva, como as transferências facultativas e obrigatórias. A Universidade de Brasília também aceita alunos através de processos seletivos próprios que aparecem como "Enem UnB", que assim como o Sistema de Seleção Unificada (SISU), utilizam-se do mesmo critério, o Exame Nacional do Ensino Médio (ENEM), tendo assim sentido agrupar essas duas categorias.

Assim, a nova tabela para variável "forma de ingresso"dos estudantes de Licenciatura em Computação dispõe das seguintes categorias:

Tabela 4: Formas de ingresso em Licenciatura em Computação-UnB, 2012-2019. Forma agrupada.

Formas de ingresso	Percentual
Vestibular	43,33%
Programa de Avaliação Seriada - PAS	24,76%
ENEM	21,32%
Portador de Diploma de Curso Superior	8,39%
Transferência	2,2%

- **Currículo**

  A variável currículo foi criada para identificar qual currículo estava vigente no ingresso do aluno no curso. A variável possui duas categorias, sendo "velho" para os que ingressaram de 2012/2 até 2015/1 e "novo" para alunos que ingressaram a partir do período 2015/2.

- **Cursou Verão**

  A variável "cursou verão"foi criada para identificar se um aluno i cursou alguma disciplina em um semestre de verão. Para isso, foi utilizado da variável "período cursou disciplina" para identificar os alunos que cursaram alguma disciplina durante o verão. A variável é então binária, indicando "Sim" para os que cursaram e "Não" para os que não.

- **Semestres Cursados**

  A variável "semestres cursados" busca identificar quantos semestre o aluno cursou antes de saída do curso. Para isso foram utilizadas as variáveis "período de entrada no curso" e "período de saída do curso", sendo a diferença entre elas o número de semestres cursados. Para os alunos ainda ativos no curso, foi atribuído como período de saída o 2019/2, visto que é o último semestre a qual temos informações. Já aqueles que foram desligados durante o período de verão, foi atribuído como semestre de saída o primeiro semestre subsequente.

- **Local de Residência**

  A partir da variável CEP do banco de dados foi criada a variável "local de residência", cruzando o CEP do aluno com um banco de CEPs do Brasil, para assim identificar a região administrativa (RA) do Distrito Federal ou cidade brasileira que o aluno reside ao fazer o registro na UnB.

  Com a perspectiva de ter um indicativo socioeconômico, as RAs foram agrupadas em quatro grupos de RAs, classificadas segundo padrões de rendimento médio, com o intuito de explorar as heterogeneidades regionais existentes na Capital Federal, sendo esses grupos:

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

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

Fonte: Elaboração própria, dados da PED 2018 - DIEESE. [5]

As cidades de estado de Goiás foram classificadas segundo seu rendimento médio

[5]Disponível em: https://www.dieese.org.br/analiseped/2018/201804pedbsb.html

em 2018 conforme as categorias acima, sendo Planaltina, Novo Gama, Luziânia, Cidade Ocidental, Anápolis e Águas Lindas de Goiás classificadas como baixa renda, Valparaíso de Goiás e Formosa como média baixa renda e Goiânia como média alta renda.

4 Resultados

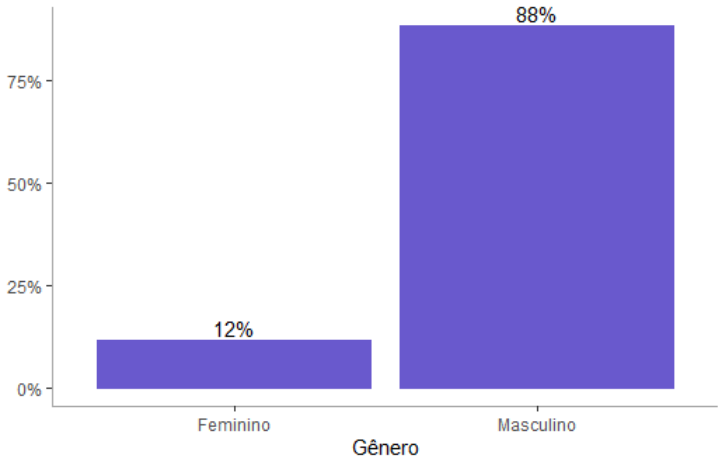
4.1 Análise Descritiva

Antes de buscar identificar quais os fatores que levam os estudantes do curso de Licenciatura em Computação a evadirem, é importante traçar e identificar o perfil do estudante através de suas características acadêmicas e sociais.

4.1.1 Dados Pessoais

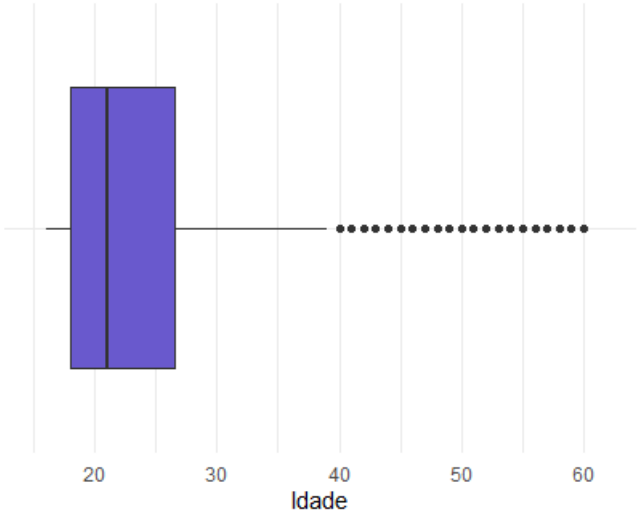
No curso de Licenciatura em Computação pode-se observar uma grande predominância de alunos do gênero masculino visto que apenas 12% dos 727 estudantes são do sexo feminino (Gráfico 1).

Gráfico 1: Distribuição dos alunos por gênero.
Licenciatura em Computação-UnB, 2012-2019.



Ao avaliar a idade ao ingressar dos estudantes, observa-se pelo Gráfico 2 uma grande concentração nas idades mais baixas, principalmente por volta dos 18 e 19 anos e, essa concentração diminui conforme o avanço das idades. A mediana da idade dos alunos de Licenciatura em Computação é de 21 anos, isto é, 50% dos estudantes tinham até 21 anos ao ingressarem no curso. A média para a idade é de 23,79 anos, com um desvio padrão de 8,18 e um coeficiente de variação de 34,42%. Apesar de conter alunos que ingressaram com mais de 50 anos, 75% dos alunos tem até 26,5 anos. As idades acima de 40 anos, apesar de ocorrem, são consideradas *outliers*, visto que a maior parte da distribuição dos dados está concentrada em idades mais baixas.

Gráfico 2: Idade e *outliers*. Licenciatura em Computação-UnB, 2012-2019



Dado a UF de nascimento dos estudantes de Licenciatura em Computação, eles se encontram advindos de 23 unidades da federação diferentes, sendo cerca de 69% deles nascidos no DF. Os demais alunos têm como UF de nascimento as outras 22 UFs, com Goiás e Rio de Janeiro representando cerca de 5% cada e Minas Gerais 4%. Para 2% dos estudantes não há informação dessa variável.

Do total de alunos, cerca de 93% residiam no Distrito Federal (DF) quando ingressaram na UnB. A Tabela 6 apresenta como eles estão distribuídos pelas regiões administrativas do DF.

Tabela 6: Distribuição dos alunos de Computação por RA.
Licenciatura em Computação-UnB, 2012-2019.

Região Administrativa	Percentual	Região Administrativa	Percentual
RA I - Plano Piloto	19,39%	RA XIV - São Sebastião	2,34%
RA II - Gama	3,03%	RA XV - Recanto das Emas	1,10%
RA III - Taguatinga	11,28%	RA XVI - Lago Sul	4,40%
RA IV - Brazlândia	0,55%	RA XVII - Riacho Fundo I	1,10%
RA V - Sobradinho	7,84%	RA XVIII - Lago Norte	2,06%
RA VI - Planaltina	1,93%	RA XIX - Candangolândia	0,69%
RA VII - Paranoá	2,34%	RA XX - Águas Claras	7,84%
RA VIII - Núcleo Bandeirante	1,24%	RA XXI - Riacho Fundo II	0,41%
RA IX - Ceilândia	4,95%	RA XXII - Sudoeste/Octogonal	3,44%
RA X - Guará	6,46%	RA XXIV - Park Way	0,96%
RA XI - Cruzeiro	4,13%	RA XXV - Estrutural/Scia	0,69%
RA XII - Samambaia	3,44%	RA XXVIII - Itapoã	0,14%
RA XIII - Santa Maria	1,79%	RA XXX - Vicente Pires	0,14%

Grande parte das regiões administrativas do DF são representadas na Tabela 6, porém, algumas delas se destacam mais por um maior quantitativos de alunos, como o Plano Piloto, onde residiam quase 20% dos estudantes. Taguatinga também é uma RA expressiva, com 11% dos alunos residentes, seguidas de Sobradinho e Águas Claras com 7,84%.

Vinte e dois alunos, cerca de 3%, residiam em Goiás no momento de ingresso na UnB, muitos em cidades consideradas como entorno do DF e pertencentes a Área Metropolitana de Brasília.

Tabela 7: Distribuição dos alunos por cidade. Licenciatura em Computação-UnB, 2012-2019.

Cidade	Frequência
Águas Lindas de Goiás	3
Anápolis	1
Cidade Ocidental	5
Formosa	2
Goiânia	1
Luziânia	3
Novo Gama	2
Planaltina	1
Rialma	1
Valparaíso de Goiás	3

Além de estudados residentes no Distrito Federal e no estado de Goiás, há a ocorrência de um aluno da cidade de Ribeirão Preto, São Paulo e mais 24 alunos onde não foi possível identificar o local de residência por problemas na informação da variável CEP.

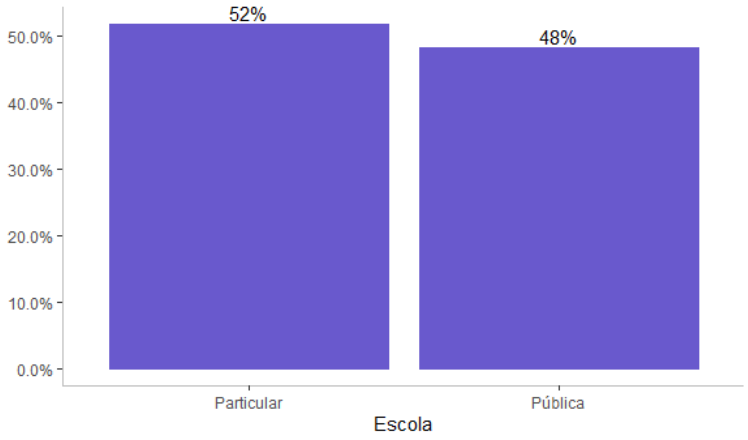
Ao olhar-se as regiões administrativas e cidades por uma visão socioeconômica conforme descrita na subseção 3.2, tem-se a distribuição dos alunos conforme a Tabela 8. Observa-se que mais de 70% dos estudantes são provenientes de RAs e cidades de alta e média alta renda, enquanto apenas 6% residiam em regiões de baixa renda ao ingressarem na Universidade de Brasília.

Tabela 8: RAs e cidades por renda. Licenciatura em Computação-UnB, 2012-2019.

Renda	Percentual
Baixa	6,33%
Média Baixa	17,19%
Média Alta Renda	42,78%
Alta	30,26%
Sem informação	3,44%

Outra característica prévia dos estudantes é referente a escola que cursou a educação básica. Percebe-se que a distribuição dos alunos conforme onde estudaram no ensino médio é parecida, sendo para particular ligeiramente maior, com 52% dos alunos (Gráfico 3).

Gráfico 3: Distribuição dos alunos por escola. Licenciatura em Computação-UnB, 2012-2019



4.1.2 Ingresso na Universidade

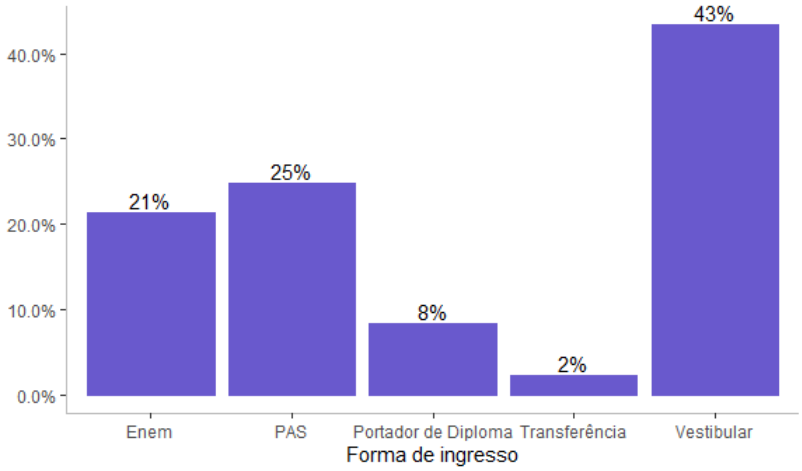
Após analisar as características prévias, isto é, aquilo que é nato dos alunos antes do curso de graduação, é interessante analisar o ingresso dos estudantes na universidade. Do total de alunos, 98% possuem a data de ingresso no curso de Licenciatura em Computação igual a de ingresso na Universidade de Brasília, sendo assim Licenciatura em Computação o primeiro curso na UnB. Os 727 alunos se encontram distribuídos de acordo com o período de ingresso no curso conforme o Gráfico 9.

Tabela 9: Distribuição dos alunos por período de ingresso. Licenciatura em Computação-UnB, 2012-2019.

Período	Frequência	Período	Frequência	Período	Frequência
2012/2	44	2015/1	53	2017/2	49
2013/1	35	2015/2	50	2018/1	48
2013/2	49	2016/1	58	2018/2	49
2014/1	50	2016/2	47	2019/1	39
2014/2	55	2017/1	54	2019/2	47

Uma característica interessante é verificar a forma com que os estudantes ingressam no curso. Percebe-se que a forma de ingresso mais usual de ingresso é o vestibular da UnB, seguido pelo PAS (Gráfico 4). Uma característica que deve ser observada é que 8% dos ingressantes são portadores de diploma de ensino superior, isto é, pessoas formadas em algum outro curso.

Gráfico 4: Distribuição dos alunos por forma de ingresso.
Licenciatura em Computação, 2012-2019.



A Universidade de Brasília dispõe sistemas de cotas que podem ser utilizados para ingresso na universidade. Pelo gráfico 5, observa-se que 75% dos estudantes não utilizaram o sistema de cotas e, dos 25% que utilizaram, eles se encontram distribuídos nos seguintes tipos de cota detalhados na tabela a seguir.

Gráfico 5: Distribuição dos alunos por sistema de cotas. Licenciatura em Computação-UnB, 2012-2019.

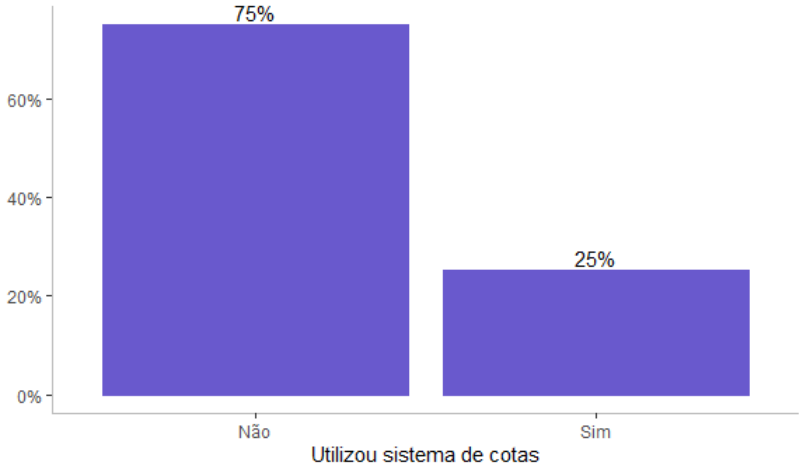
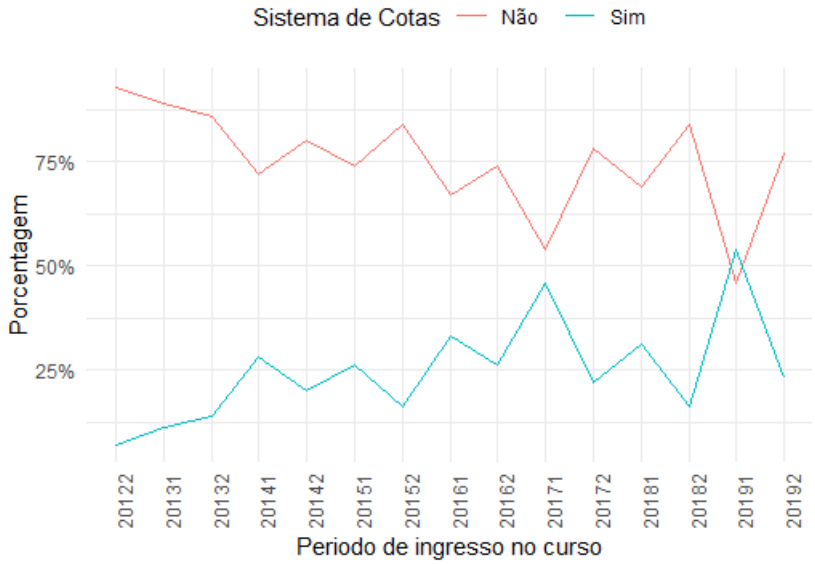


Tabela 10: Distribuição dos alunos cotistas por tipo de cotas.
Licenciatura em Computação-UnB, 2012-2019.

Cota	Percentual
Escola Púb Baixa Renda Não-PPI	2.3%
Escola Púb Baixa Renda PPI	5.2%
Escola Púb Alta Renda Não-PPI	6.6%
Escola Púb Alta Renda PPI	5.9%
Negro	5.1 %

O Gráfico 6 mostra o uso do sistema de cotas segundo o período de ingresso dos alunos de Licenciatura em Computação. Observa-se que a porcentagem de uso do sistema de cotas cresce no decorrer dos períodos, muito em virtude da implementação desses sistemas. O percentual de ingressantes que utilizam do sistema de cotas tende a ser menor, menos no período 2019/1, que teve um comportamento atípico. A tabela com os percentuais se encontra no apêndice A.

Gráfico 6: Distribuição do uso do sistema de cotas por período.
Licenciatura em Computação-UnB, 2012-2019.



Dentro do sistema de cotas, a maioria dos tipos de cota são destinados a alunos de escola pública. Assim, é interessante analisar como os estudantes então utilizando dessas cotas conforme a variável "escola". A Tabela 11 fornece o percentual de quantos alunos são advindos da escola pública ou particular conforme os sistemas de cotas. O percentual refere-se ao total da linha.

Tabela 11: Distribuição do sistema de cotas por tipo de escola. Licenciatura em Computação-UnB, 2012-2019.

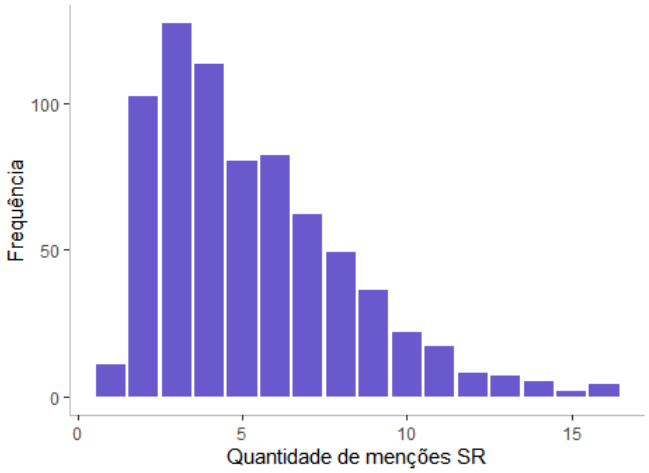
		Escola	
Variável		Particular	Pública
Sistema de cotas	Não	64,34%	35,66%
	Sim	14,21%	85,79%
Cotas	Esc. Pública Baixa Renda-Não PPI	0%	100%
	Esc. Pública Baixa Renda- PPI	2,63%	97,37%
	Esc. Pública Alta Renda-Não PPI	4.17%	95,83%
	Esc. Pública Alta Renda- PPI	6,98%	96,02%
	Negro	54,05%	45,95%
	Sem cota	64,34%	35,66%

Com relação a variável sistema de cotas, nota-se que a maioria dos alunos que utilizaram o sistema, 85,79%, são advindos de escola pública, o que é coerente, visto que grande parte das cotas são para escola pública. Já ao observar isso cota a cota, para as cotas de escola pública, em sua maioria, foram preenchidas por alunos que também possuem escola pública como categoria da variável escola. Mas, observando as cotas de escola pública, se nota algumas inconsistências, visto que não deveria ocorrer de alunos de escola particular adentrarem a universidade com cotas de escola pública. Essas inconsistências podem ser advindas de um erro no momento da coleta da informação da variável escola, visto que essa é uma informação fornecida pelo aluno no momento da matrícula.

4.1.3 Vida Acadêmica

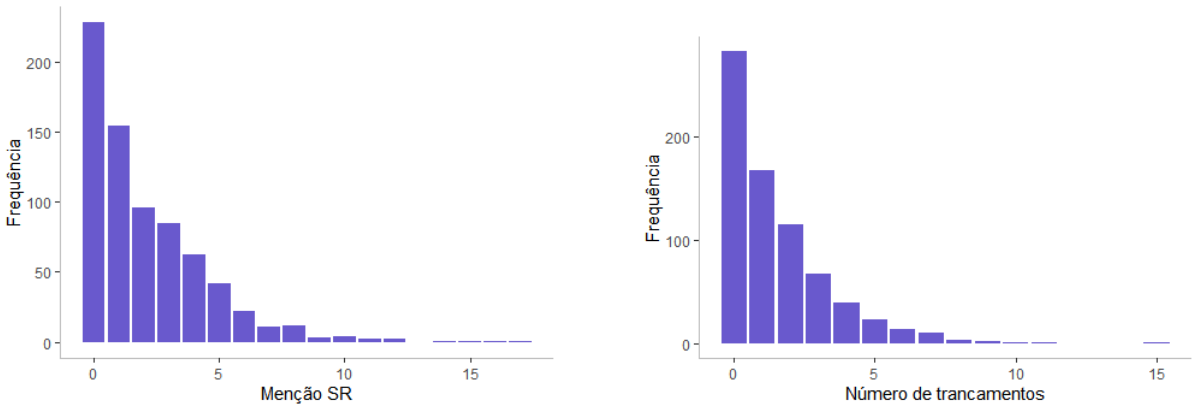
Observa-se no Gráfico 7 que há uma concentração maior em poucos semestres cursados pelos estudantes, com maior frequência para 3 semestres cursados. Os alunos atribuídos a 1 semestre cursado são aqueles que evadiram do curso no mesmo semestre que entraram. É possível notar que poucos alunos alcançam a marca de 9 semestres, que é o tempo mínimo para formatura. Mais adiante, veremos como essa variável se comporta em relação a evasão.

Gráfico 7: Distribuição dos alunos por semestres cursados. Licenciatura em Computação-UnB, 2012-2019



Identificar a quantidade de menções SR é importante pois fornece uma noção se os alunos de Licenciatura em Computação costumam abandonar as disciplinas. Pelo gráfico, nota-se uma alta concentração de alunos que não receberam menção SR e também aqueles que essa menção apareceu uma única vez em seu histórico, sendo também a mediana dos dados. Em geral, 75% os estudantes tiveram até 3 SRs, mas há algumas observações discrepantes com muitas menções SRs, que geralmente são de alunos do currículo antigo, desligados ou ainda ativos.

Gráfico 8: Distribuição das variáveis quantidade de menções SR e trancamentos. Licenciatura em Computação-UnB, 2012-2019



(a) Quantidade de SR

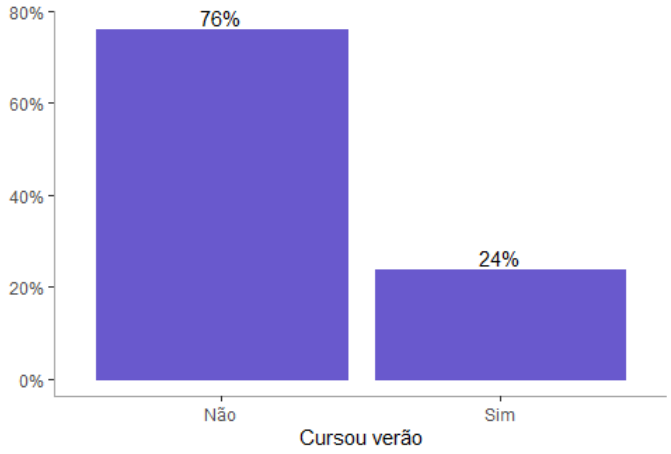
(b) Número de trancamentos

Para o número de trancamentos, o Gráfico 8 indica que há uma grande concentração para valores mais baixos de trancamentos de disciplinas realizadas pelos alunos durante o curso. Em sua grande maioria, os alunos realizam 0 ou 1 trancamento de

disciplina durante o curso, visto que a mediana é de 1 trancamento.

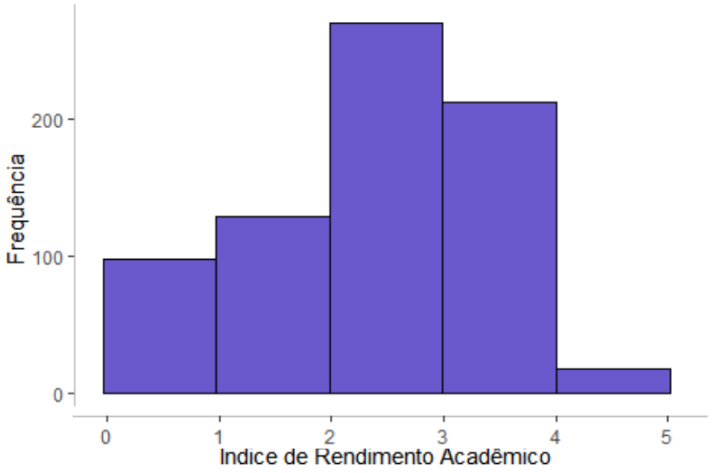
Outra característica interessante é observar se o aluno cursou ou não alguma vez uma matéria durante um semestre de verão. O Gráfico 9 indica que 76% dos alunos não cursaram verão nenhuma vez durante o curso enquanto 24% cursou alguma disciplina durante o verão no decorrer do curso. A realização de disciplinas durante o verão é importante pois o aluno pode encontrar aqui possibilidade de cursar disciplinas atrasadas, visto que as disciplinas que costumam ser ofertadas no verão são aquelas de maior retenção.

Gráfico 9: Distribuição dos alunos por cursou verão. Licenciatura em Computação-UnB, 2012-2019



O Índice de Rendimento Acadêmico (IRA) é utilizado pela Universidade de Brasília como uma nota atribuída ao desempenho acadêmico dos estudantes, sendo atribuído 5 para melhor rendimento e 0 para o pior. Abaixo, tem-se a distribuição da variável IRA para os alunos do curso de Licenciatura em Computação:

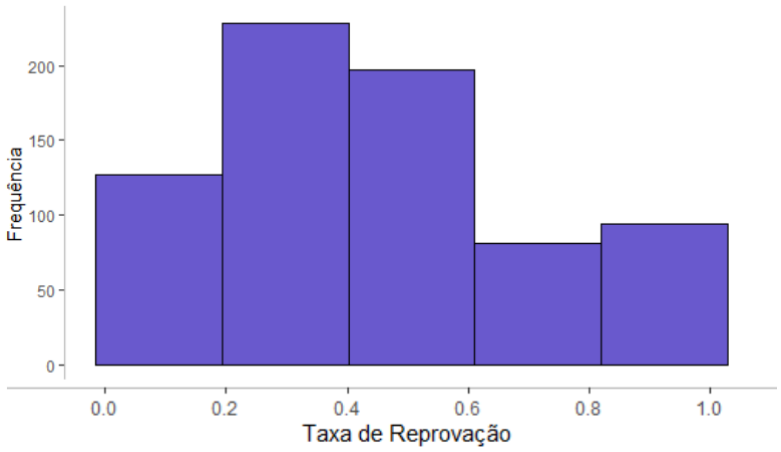
Gráfico 10: Distribuição dos alunos por IRA. Licenciatura em Computação-UnB, 2012-2019



Nota-se que há uma maior frequência de alunos com IRA entre 2 e 3, seguidos por alunos com IRA entre 3 e 4. É de se esperar que não haja muitos alunos perto dos extremos, mas nota-se que a frequência de IRA entre 0 e 1 é expressiva. Isso pode dar-se por causa de um número expressivo de menções SR, onde é atribuído o valor zero ao IRA.

A taxa de reprovação é uma variável construída conforme indicado na seção 3.2. Essa variável busca medir a proporção de créditos que um aluno cursou na Universidade de Brasília e recebeu dela uma menção de reprovação.

Gráfico 11: Distribuição dos alunos por taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019



O Gráfico 11 mostra que há uma maior concentração na taxa de reprovação para valores entre 0,2 e 0,4. Ainda assim, há uma quantidade relativamente expressiva na faixa

de 0,8 a 1, indicando que há um número de alunos que reprovou mais de 80% dos créditos cursados. De modo geral, a mediana para essa variável é de 0,38 com uma média de 0,42.

4.1.4 Saída do Curso

Um dos objetivos desse trabalho é identificar a evasão acadêmica e os seus fatores. A variável forma de saída do curso dá uma visão de como os estudantes estão saindo do curso através da tabela a seguir.

Tabela 12: Distribuição dos alunos por forma de saída. Licenciatura em Computação-UnB, 2012-2019.

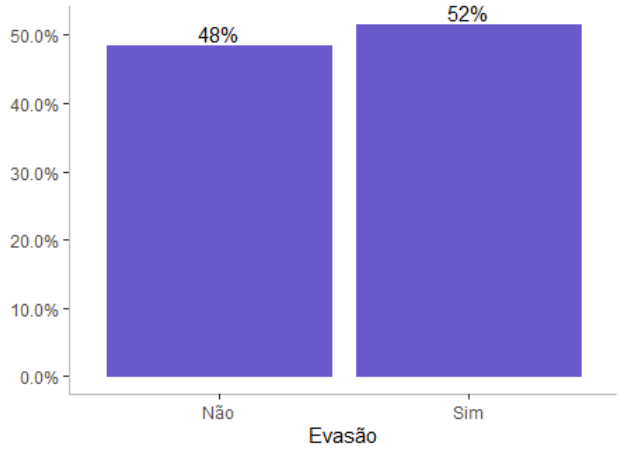
Formas de saída	Frequência	Percentual
Alunos ativos	304	41,8%
Formatura	48	6,6%
Reprovar 3x a mesma disciplina obrigatória	26	3,6%
Desligamento - Não cumpriu condição	184	25,3%
Desligamento - Abandono	108	14,9%
Desligamento - Decisão Judicial	1	0,1%
Desligamento - Voluntário	18	2,5%
Mudança de curso	6	0,8 %
Novo vestibular	30	4,1%
Transferência	2	0,3%

Os alunos ativos representaram 41,8% dos estudantes, permanecendo ainda no curso. Mas, o que também chama atenção é que, de todos os alunos que ingressaram no curso a partir do período 2012/2, apenas 6,6% alcançaram a formatura. Nota-se muito facilmente como desligamento do curso é um fator de peso na saída dos alunos do curso. De todos, 25,3% deixaram o curso por desligamento onde não houve o cumprimento de condição.

A variável Evasão criada conforme detalhado na seção 3.2 busca medir a proporção alunos que evadem o curso de Licenciatura em Computação conforme a definição de evasão adotada nesse trabalho.

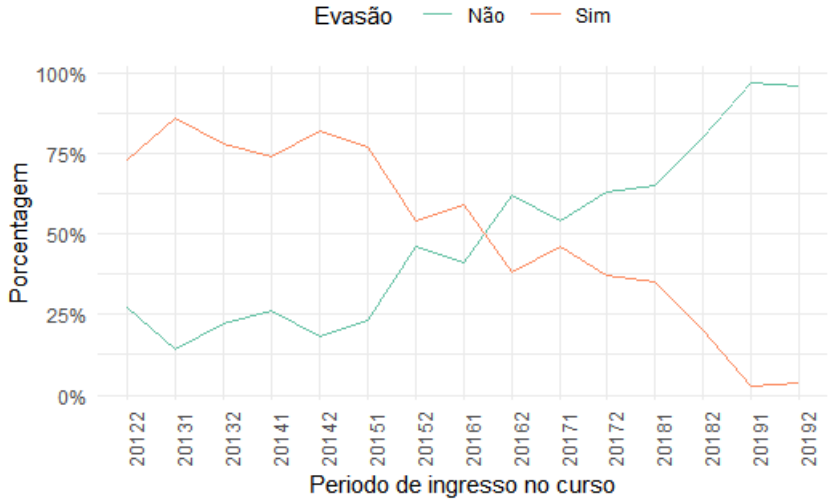
Percebe-se pelo gráfico abaixo que a evasão no curso de Licenciatura em Computação é muito alta, em torno dos 52%. A categoria Não-Evasão engloba os alunos que se formaram e os que continuavam ativos no curso até o semestre 2019/2 e, ainda assim, tem-se um maior número de alunos que evadiram do curso segundo as formas de saída do curso.

Gráfico 12: Distribuição dos alunos por evasão. Licenciatura em Computação-UnB, 2012-2019



Outra maneira de avaliar a evasão é observar o percentual de evadidos conforme o período de ingresso. No Gráfico 13, observa-se que o percentual de evasão para os períodos mais antigos chega a alcançar 85% e permanece alto até o período 2016/2 onde há uma inversão, sendo o percentual de não evadidos maior. Isso decorre por dois fatores, um que a partir de 2016 grande parte dos alunos continuam ativos visto que ainda não ficaram expostos ao risco de evadir por muito tempo e segundo, pois realmente aparenta ter ocorrido uma diminuição no percentual de evasão.

Gráfico 13: Distribuição dos alunos por evasão e período de ingresso. Licenciatura em Computação-UnB, 2012-2019



4.2 Análise Bivariada

Como mencionado, um dos objetivos desse trabalho é avaliar que fatores influenciam na evasão ou na não evasão dos estudantes. Para isso, uma análise bivariada pode fornecer uma ideia inicial de como se comporta a evasão conforme outros fatores. A tabela a seguir dispõe da informação da evasão conforme algumas características dos estudantes.

Tabela 13: Análise bivariada por evasão. Licenciatura em Computação-UnB, 2019-2019.

		Evasão	
Variável		Sim	Não
Gênero	Feminino	54%	46%
	Masculino	51%	49%
Local de residência	Baixa renda	33%	67%
	Média baixa renda	47%	53%
	Média alta renda	54%	46%
	Alta renda	53%	47%
Escola	Particular	49%	51%
	Pública	54%	46%
Forma de ingresso	Enem	57%	43%
	PAS	40%	60%
	Portador de Diploma	69%	31%
	Transferência	69%	31%
	Vestibular	51%	49%
Sistema de cotas	Não	53%	47%
	Sim	49%	51%
Currículo	Velho	78%	22%
	Novo	39%	61%
Cursou verão	Não	60%	40%
	Sim	24%	76%

Percebe-se que existe alguma diferença entre as categorias das variáveis apresentadas em relação a evasão, como um maior percentual de evadidos para o sexo feminino

do que para o masculino, assim como para a variável escola, onde a categoria de escola pública aparenta sofrer mais com a evasão que a escola privada. Para o local de residência observa-se que os locais de média baixa e baixa renda os alunos evadem menos que os que residem em locais de renda mais alta. A categoria de baixa renda é a que mais se destacam, onde apenas 33% evadem.

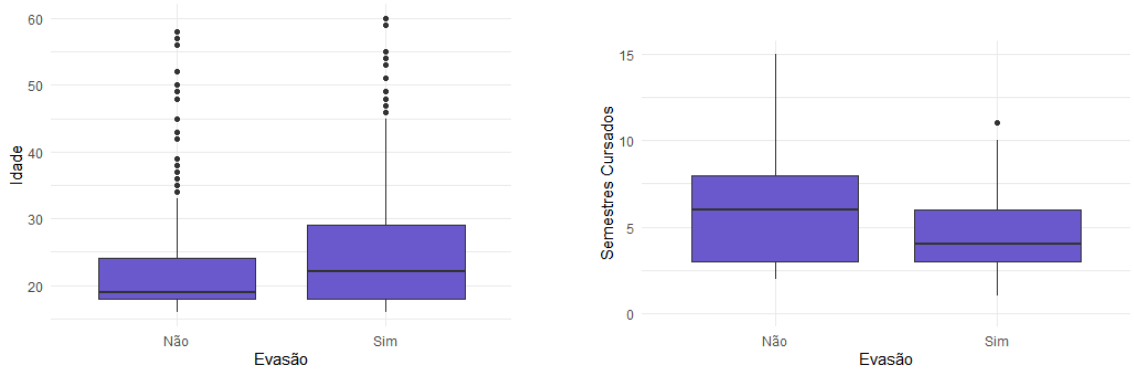
A variável forma de ingresso apresenta resultados diferentes em relação a evasão conforme suas categorias. Para a modelagem, as categorias Portador de Diploma e Transferência foram agrupadas, dados que elas possuem comportamentos parecidos e que as transferências correspondem menos de 2% dos alunos.

A variável currículo também apresenta grande diferença em suas categorias em relação a evasão. No geral, 68% dos alunos são do currículo novo e 32% do velho. A variável cursou verão apresenta grandes mudanças em suas categorias em relação a evasão, os alunos que cursaram disciplinas durante o verão aparentam evadir menos. Essas diferenças podem indicar que esses fatores influenciam na evasão caso sejam significativas.

Já para as variáveis quantitativas, uma maneira de perceber como elas se comportam com relação a variável evasão é através de gráficos *bloxplot*. O Gráfico 14 apresenta o gráfico para as variáveis idade ao ingressar e semestres cursados pela variável evasão.

Ao analisar a idade ao ingressar pela evasão, vê-se que os alunos que evadem costumam ser mais velhos, visto que a sua distribuição é mais dispersa. A mediana para os que evadiram é de 22 anos, enquanto para os que não é de 19 anos.

Gráfico 14: Distribuição de alunos por idade e semestres cursados em relação a evasão. Licenciatura em Computação-UnB, 2012-2019



(a) Idade ao ingressar

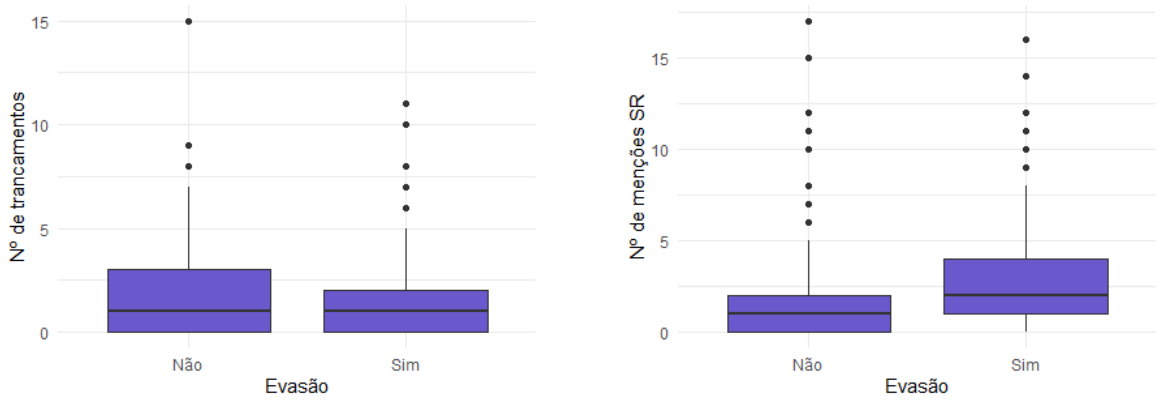
(b) Semestres cursados

Nota-se que a distribuição dos semestres cursados por aqueles alunos que evadiram é mais concentrada em valores mais baixos, o que faz sentido, uma vez que saíram do curso

antes da formatura (Gráfico 14). A mediana dos semestres cursados para esse grupo é de 4 semestres, tendo uma alta concentração por volta do segundo semestre. Observa-se que para os que evadiram, 75% deles não passaram de 6 semestres no curso. Para o grupo dos não evadidos, os semestres são mais distribuídos, visto que a categoria engloba alunos ainda ativos. Mas, é possível observar que os alunos que não evadem tendem a ultrapassar a marca de 9 semestres, que é o tempo que um aluno no fluxo do curso leva para se formar.

Ao olhar a evasão pelas variáveis Número de trancamentos e quantidade de menções SR no Gráfico 15, para o número de trancamentos, observa-se que apesar de terem medianas parecidas, os alunos que não evadiram costumam usar mais do trancamento. Já para a variável quantidade de menções SR, verifica-se que os alunos que evadiram possuem mais menções SR, visto que seu primeiro quartil é superior ao segundo quartil dos que não evadiram, o que é esperado.

Gráfico 15: Distribuição dos alunos por Número de trancamentos e quantidade de menções SR em relação a evasão. Licenciatura em Computação-UnB, 2012-2019

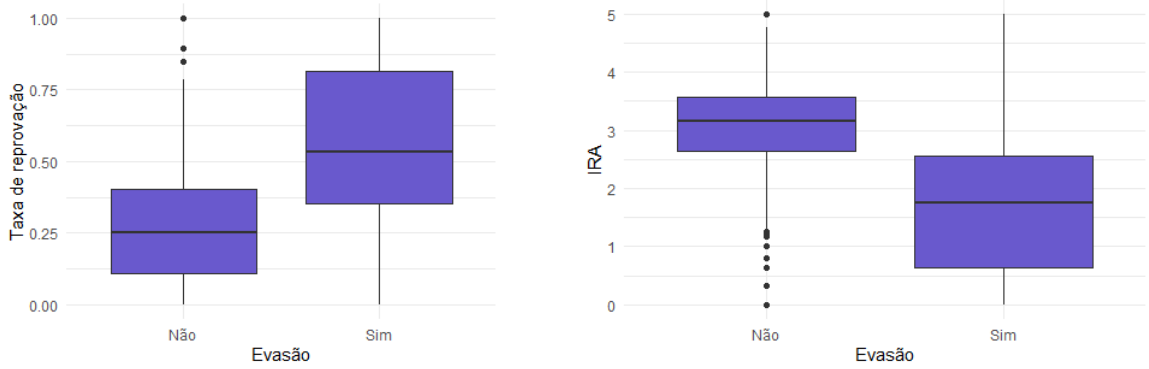


(a) Número de trancamentos

(b) Quantidade de menções SR

O Gráfico 16a apresenta as variáveis taxa de reprovação e IRA. A variável taxa de reprovação apresenta um comportamento diferentes para evasão e não evasão. A mediana para quem não evadiu é entorno de 0,25, já para os que evadiram essa supera 0,50. O Gráfico 16b, apresenta o *boxplot* da variável IRA pela evasão. Assim como a taxa de reprovação, o IRA também funciona como uma medida de desempenho, mas essa é inversa a taxa de reprovação. Observa-se que os alunos não evadidos apresentam um IRA maior que os evadidos, um resultado esperado. Os valores identificados como *outliers* para os não evadidos são de alunos que possuem IRA baixos decorrentes e muitos SRs mas que não evadiram. A diferença da variável IRA aparenta ser significativa visto que a mediana para não evadidos é de 3,15 e para evadidos de 1,76.

Gráfico 16: Distribuição dos alunos por taxa de reprovação e IRA com relação a evasão. Licenciatura em Computação-UnB, 2012-2019



(a) Taxa de reprovação

(b) IRA

Após a análise descritiva das variáveis, é preciso elencar possíveis variáveis explicativas para o modelo. Para isso, primeiramente, foi testado a associação e o efeito de cada variável em relação a evasão, a variável de interesse desse estudo. Na tabela a seguir encontram-se o resultado de testes para cada variável de maneira isolada. Para as variáveis qualitativas, foi realizado o teste Qui-Quadrado de Pearson para associação e para as quantitativas, uma regressão logística simples.

Tabela 14: Associação das variáveis com evasão. Licenciatura em Computação-UnB, 2012-2019.

Variável	Estatística do teste	P-valor
Gênero	0,073933	0,7857
Idade	-4,452	$< 0,0001$
Local de Residência agrupado	6,136	0,1051
Escola	1,2234	0,2687
Forma de ingresso	20,517	0,0001
Sistema de cotas	0,70056	0,4026
Currículo	95,076	$< 0,0001$
Cursou verão	66,342	$< 0,0001$
Semestres cursados	7,953	$< 0,0001$
Número de trancamentos	3,658	0,0002
Número de SR	-6,985	$< 0,0001$
Taxa de reprovação	-11,543	$< 0,0001$
IRA	12,6	$< 0,0001$

Da Tabela 14, observa-se que as variáveis sexo, sistema de cotas e escola não apresentaram uma associação significativa com a evasão, considerando um nível de significância de 5%. A variável local de residência em sua forma agrupada apresentou um p-valor perto dos 10%, mas não suficiente baixo para ser significativa a associação. Apesar de não apresentarem uma associação forte, essas e as demais variáveis serão testadas nos modelos.

4.2.1 Correlação entre variáveis

Os problemas de multicolinearidade nos modelos de regressão, ou seja, as relações entre as variáveis presentes no modelo, podem prejudicar a capacidade preditiva do mesmo. Nesta seção, será apresentado uma análise de correlação entre algumas variáveis visto que a presença de variáveis muito relacionadas no modelo não é ideal.

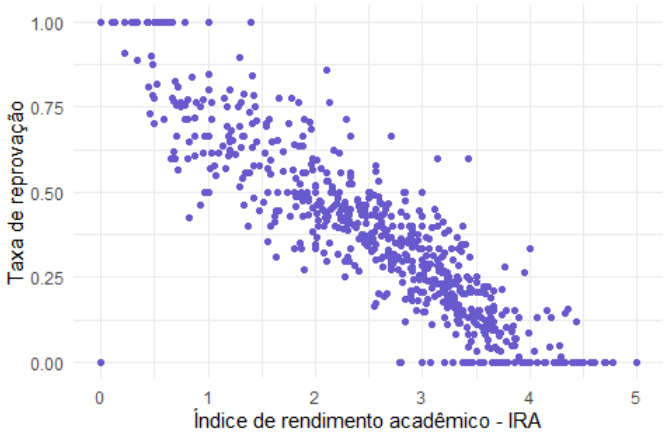
- **IRA e Taxa de reprovação**

  No banco de dados tem-se duas variáveis quantitativas referentes de alguma forma ao desempenho do aluno durante o curso, o índice de rendimento acadêmico e a taxa de reprovação.

  Tendo em vista que são duas variáveis quantitativas contínuas, foi utilizado o coeficiente de correlação de Pearson (ρ) para identificar o grau de associação entre essas variáveis, obtendo um ρ de -0.9161842. Esse valor para o coeficiente de correlação de Pearson indica uma correlação linear entre as variáveis IRA e taxa de reprovação, sendo uma correlação forte e negativa, ou seja, inversamente proporcional.

  Isso pode ser visualizado através do gráfico de dispersão das variáveis:

Gráfico 17: IRA e taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019



- **Sistema de Cotas e Escola**

  Para as variáveis sistema de cotas e escola, há uma suspeita de que elas sejam correlacionadas visto que grande parte dos sistemas de cota são destinados a alunos de escola pública. As variáveis são qualitativas com duas categorias cada. Assim, para investigar se há associação entre elas foi realizado o teste de Qui-quadrado de Pearson na tabela de contingência de sistema de cotas por escola.

Tabela 15: Sistema de cotas e escola. Licenciatura em Computação-UnB, 2012-2019.

		Escola	
Variável		Particular	Pública
Sistema de cotas	Não	350	194
	Sim	26	157

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

Dado que o p-valor para o teste é $< 0,0001$, há evidências para rejeitar a hipótese de não associação das variáveis, isto é, elas são associadas.

4.3 Modelagem

Dado os resultados encontrados na seção anterior, a abordagem para o construção do modelo seguiu-se em construir dois modelos para o banco de dados, um com a variável IRA e outro com a taxa de reprovação, visto que as variáveis são correlacionadas e muito significativas. O mesmo foi realizado com as variáveis Escola e Sistema de cotas.

O banco foi dividido em duas amostras, uma utilizada para a construção do modelo e outra para a validação dele. A ideia é usar o mesmo modelo para o banco validação e observar se os valores das estimativas são semelhantes.

Utilizando a base de construção, foram aplicados dois modelos iniciais, um com IRA e o outro com taxa de reprovação com todas as demais variáveis presentes na Tabela 13.

Tabela 16: Teste inicial com todas variáveis para o modelo com IRA.
Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa	Desvio Padrão	Estatística do teste	P-valor
Intercepto	4,4705	1,4547	3,293	$< 0,0001$
IRA	-1,4337	0,2428	-5,905	$< 0,0001$
Gênero - Masc	-0,6396	0,5129	-1,247	0,2123
Escola - Pública	0,3089	0,3607	0,857	0,3916
Forma Ingresso - Enem	-0,94183	0,6924	-1,360	0,1738
Forma Ingresso - PAS	-0,82516	0,7701	-1,071	0,0,2839
Forma Ingresso - Vestibular	-0,9899	0,6745	-1,468	0,1421
Currículo - Velho	2,4196	0,4264	5,674	$< 0,0001$
Trancamentos	0,08115	0,12038	0,674	0,5002
SR	0,0250	0,1013	0,247	0,8048
Semestres cursados	-0,2324	0,0766	-3,033	0,0024
Verão - SIM	-1,7331	0,4630	-3,7431	$< 0,0001$
Local Res - Baixa Renda	-0,3647	0,7507	-0,486	0,6271
Local Res - Média Baixa Renda	0,4483	0,4990	0,898	0,3690
Local Res - Média Alta Renda	0,6066	0,4057	1,495	0,1348

Observando os resultados dos modelos iniciais com todas as possíveis variáveis explicativas, nota-se que mesmo algumas variáveis que sozinhas deram significativas, no modelo com as demais não apresentaram significância. Para a seleção das variáveis de cada modelo e suas interações serão aplicadas técnicas de seleção de variáveis.

Tabela 17: Teste inicial com todas variáveis para o modelo com taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa	Desvio Padrão	Estatística do teste	P-valor
Intercepto	-0,3641	1,1549	-0,315	0,7525
Taxa de reprovação	3,7312	0,8206	4,547	$< 0,0001$
Gênero - Masc	-0,4908	0,4775	-1,028	0,3040
Idade	0,0082	0,0239	0,346	0,7293
Escola - Pública	0,2137	0,2401	0,628	0,5297
Forma Ingresso - Enem	-0,8391	0,6293	-1,333	0,1824
Forma Ingresso - PAS	-0,87335	0,7118	-1,227	0,2198
Forma Ingresso - Vestibular	-1,0318	0,6151	-1,667	0,0934
Currículo - Velho	2,4011	0,4044	5,938	$< 0,0001$
Trancamentos	0,0985	0,1132	0,870	0,3843
SR	0,2019	0,0938	2,151	0,0315
Semestres cursados	-0,28631	0,0741	-3,862	0,0001
Verão - SIM	-1,8752	0,4337	-4,324	$< 0,0001$
Local Res - Baixa Renda	-0,5526	0,7202	-0,767	0,4429
Local Res - Média Baixa Renda	0,3771	0,4715	0,799	0,4241
Local Res - Média Alta Renda	0,6508	0,3881	1,677	0,0936

4.3.1 Modelo com a variável IRA

Para o modelo incluindo a variável IRA, a seleção de variáveis seguiu o critério apresentado na seção 2.4, utilizando o método *stepwise* para a seleção das variáveis explicativas e suas interações. Após a obtenção dessas variáveis, também foi realizado testes de modelos encaixados para verificar se o acréscimo de determinada variável era realmente significativo no modelo utilizando um nível de significância de 5%.

Ao final do processo, o modelo ficou com 5 variáveis: IRA, forma de ingresso, currículo, semestres cursados e cursou verão, além da interação entre IRA e semestres cursados. A tabela abaixo contém as estimativas para essas variáveis para a amostra de construção, validação e o banco completo.

Tabela 18: Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com IRA. Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa-Const.	Estimativa-Valid.	Estimativa-Geral
Intercepto	3,69	3,69	3,61
IRA	-0,80	-0,70	-0,74
Forma Ingresso - Enem	-1,14	-1,60	-1,33
Forma Ingresso - PAS	-1,35	-1,61	-1,44
Forma Ingresso - Vestibular	-1,64	¨-1,63	-1,61
Currículo - Velho	2,81	2,91	2,79
Semestres cursados	0,33	0,03	0,19
Verão - SIM	-1,35	-0,95	-1,15
IRA*Semestres cursados	-0,19	-0,10	-0,15

As estimativas dos parâmetros são bem parecidas entre os bancos de construção e validação e também para o banco geral. Assim, pode-se concluir que o modelo proposto é válido. Os resultados finais para o modelo com IRA com o banco completo se encontram abaixo. O modelo final com IRA conta com um AIC de 588,7.

Tabela 19: Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com IRA. Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa	Desvio Padrão	Estatística do teste	P-valor
Intercepto	3,6164	0,6182	5,850	$< 0,0001$
IRA	-0,7405	0,1863	-3,974	$< 0,0001$
Forma Ingresso - Enem	-1,3397	0,4382	-3,057	0,0022
Forma Ingresso - PAS	-1,4432	0,4237	-3,406	0,0006
Forma Ingresso - Vestibular	-1,6126	0,4198	-3,841	0,0001
Currículo - Velho	2,7949	0,3086	9,057	$< 0,0001$
Semestres cursados	0,1959	0,1460	1,342	0,1796
Verão - SIM	-1,1525	0,2674	-4,311	$< 0,0001$
IRA*Semestres cursados	-0,1502	0,0510	-2,945	0,0032

4.3.2 Modelo com a variável Taxa de Reprovação

Para o modelo com taxa de reprovação, o processo seguido foi o mesmo descrito para o modelo com IRA. Ao final, as variáveis que permaneceram no modelo foram: forma de ingresso, currículo, cursou verão, taxa de reprovação, semestres cursados, SR, além das interações taxa de reprovação e semestres cursados e semestres cursados e SR. A tabela abaixo contém as estimativas para os parâmetros para a amostra de construção, validação e o banco completo.

Tabela 20: Estimativas dos parâmetros para as bases de construção, validação e geral para o modelo com taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa-Const.	Estimativa-Valid.	Estimativa-Geral
Intercepto	0,67	1,12	0,90
Forma Ingresso - Enem	-0,97	-1,79	-1,38
Forma Ingresso - PAS	-1,10	-1,84	-1,50
Forma Ingresso - Vestibular	-1,68	-1,83	-1,76
Currículo - Velho	3,12	3,26	3,15
Verão - Sim	-1,68	-1,15	-1,37
Taxa de reprovação	-0,95	-0,44	-0,65
Semestres cursados	-0,47	-0,42	-0,45
SR	0,87	0,69	0,75
Taxa Reprov*Semestres cursados	1,30	1,16	1,26
SR*Semestres cursados	-0,11	-0,10	-0,10

As estimativas para os parâmetros tanto entre as amostras de construção e validação como com o banco geral são relativamente parecidas, assim, podemos concluir que o modelo proposto construído em cima da amostra de construção é valido. Os resultados das estimativas para o modelo com taxa de reprovação aplicados no banco completo se encontram na Tabela 21. O modelo final conta com um AIC de 583,35.

Tabela 21: Estimativas dos parâmetros, desvio padrão, estatística e p-valor com os dados completos para o modelo com taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019.

Parâmetro	Estimativa	Desvio Padrão	Estatística do teste	P-valor
Forma Ingresso - Enem	-1,3832	0,4107	-3,367	0,0007
Forma Ingresso - PAS	-1,5071	0,4009	-3,759	0,0001
Forma Ingresso - Vestibular	-1,7650	0,3967	-4,449	8,62 e^{-06}
Currículo - Velho	3,1544	0,3301	9,556	$\leq 2e^{-16}$
Verão - Sim	-1,3774	0,2755	-4,999	$5{,}76e^{-07}$
Taxa de reprovação	-0,6595	0,8073	-0,817	0,4140
Semestres cursados	-0,4542	0,0829	-5,474	$4{,}4e^{-08}$
SR	0,7562	0,1487	5,083	$3{,}72e^{-07}$
Taxa Reprov*Semestres curs.	1,2676	0,2496	5,077	$3{,}83e^{-07}$
SR*Semestres cursados	-0,1087	0,0222	-4,879	$1{,}07e^{-06}$

4.4 Interpretação dos parâmetros

Em regressão logística, é possível interpretar os parâmetros estimados do modelo através da Razão de chances (Odds ratio). As Tabelas 22 e 23, apresentaram os valores estimados da razão de chances para cada parâmetro estimado e seus respectivos intervalos de confiança para os modelos IRA e taxa de reprovação, respectivamente.

4.4.1 Modelo IRA

Tabela 22: Razão de chance e IC de 95% para o modelo IRA. Licenciatura em Computação-UnB, 2012-2019.

Variável Explicativa	Razão de chances	IC 95%
IRA	0.4768	0,3257 - 0,6777
Forma Ingresso - Enem	0,2619	0,1091 - 0,6107
Forma Ingresso - PAS	0,2361	0,1012 - 0,5352
Forma Ingresso - Vestibular	0,1993	0,0860 - 0,4477
Currículo - Velho	16,3613	9,1488 - 30,7794
Semestres cursados	1,2163	0,9147 - 1,6229
Verão - SIM	0,3158	0,1846 - 0,5279
IRA*Semestres cursados	0,8605	0,7774 - 0,9498

De acordo com as razões de chance estimadas apresentados na tabela acima, a variável forma de ingresso é uma variável que apresentou um impacto significativo na evasão. Como forma de ingresso é uma variável categórica, com 4 categorias, as comparações das chances de um aluno evadir são comparadas com a forma de ingresso "Portador de Diploma e Transferências". Assim, um aluno que ingressou na Universidade de Brasília por meio do ENEM, PAS ou vestibular tem respectivamente 73%, 76% e 80% menos chances de evadir o curso comparado aos alunos que adentraram como portador de diploma e transferências.

A variável currículo apresentou uma razão de chances de 16,3613 quando o currículo velho é comparado com o novo. Assim, a chance de um aluno oriundo do currículo antigo evadir é cerca 16 vezes a chance de um aluno do currículo novo evadir. Isso mostra que a mudança de currículo em 2015/2 vem surgindo efeito.

A variável cursou verão, que leva em conta aqueles que cursaram verão alguma vez durante o curso obteve razão de chances de 0.3158, sendo assim, ao comparar os alunos que cursaram verão com o que não cursaram, esses possuem 59% menos chances de evadir comparado aos que não cursaram.

O impacto das variáveis IRA e semestres cursados nesse modelo é dado pela interação entre elas. Olhando para a interação IRA e semestres cursados, a informação dada por essa estimativa é que a cada unidade acrescida na interação IRA e semestres

cursados, a chances de evadir diminuem em cerca de 14% mantidas constantes as demais variáveis.

4.4.2 Modelo Taxa de Reprovação

Tabela 23: Razão de chance e IC de 95% para o modelo taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019.

Variável Explicativa	Razão de chances	IC 95%
Forma Ingresso - Enem	0,2507	0,1104 - 0,5548
Forma Ingresso - PAS	0,2215	0,0994 - 0,4806
Forma Ingresso - Vestibular	0,1711	0,0773 - 0,3676
Currículo - Velho	23,4403	12,6278 - 46,2271
Verão - Sim	0,2522	0,1447 - 0,4275
Taxa de reprovação	0,5171	0,1065 - 2,5536
Semestres cursados	0,6349	0,5366 - 0,7433
SR	2,1303	1,6087 - 2,8846
Taxa Reprov*Semestres cursados	3,5526	2,1977 - 5,8610
SR*Semestres cursados	0,8969	0,8570 - 0,9350

Para a interpretar os parâmetros do modelo taxa de reprovação, também será utilizado as estimativas das razões de chances para cada um. Assim como descrito no modelo IRA, a variável forma de ingresso é categórica com 4 categorias, sendo a "Portador de Diploma e Transferência" a levada em comparação. Sendo assim, um aluno que ingressou na UnB pelo ENEM, PAS ou Vestibular diminui sua chance de evadir em 74%, 77% ou 82% respectivamente, comparado ao que ingressou como portador de diploma e transferência.

O parâmetro que representa o currículo pode ser interpretado como a chance de um aluno que entrou ainda no currículo velho é 23,44 vezes maior que a chance de um aluno do currículo novo em evadir. Isso dá a percepção que a troca de currículo foi um fator importante na evasão do curso.

Já a variável cursou verão mostra que um aluno que cursou alguma disciplina de verão durante o curso tem suas chances de evadir diminuídas em cerca de 74% comparado aos que não cursaram verão mantendo os demais fatores constantes.

O impacto das variáveis taxa de reprovação, semestres cursados e SR é obtido analisando as interações entre elas. Para a interação taxa de reprovação e semestres cursados temos que para cada unidade acrescida as chances de o aluno evadir aumentam em 255%, isto é, se ao passar de cada semestre a taxa de reprovação do aluno for muito alta, as chances dele de evadir aumentam em 255%. Já para a interação semestres cursados e SR, tem-se que para cada unidade acrescida, as chances de evadir diminuem em cerca de 10%. Isso mostra que apesar das menções SR, se o aluno continuar insistindo em cursar o curso, as chances dele em evadir diminuem.

4.5 Teste de ajuste e diagnóstico dos modelos

Após ajustar um modelo estatístico, é necessário verificar a adequabilidade desse ajuste para analisar o quão bem ele se ajusta ao conjunto de dados. Os testes apresentados na subseção 2.5 serão aplicados aqui.

4.5.1 Modelo IRA

- **Testes de adequabilidade e resíduos**
  A adequabilidade do modelo pode ser avaliada através dos testes de Hosmer-Lemeshow, dos resíduos Pearson e *Deviance*. Para o teste de Hosmer-Lemeshow a estatística χ^2 obtida para esse modelo foi de 5,7645 com 8 graus de liberdade e um p-valor de 0,6736. Com isso, não há evidências para rejeitar a hipótese que o modelo esteja bem ajustado.

  Ao verificar a adequabilidade do ajuste através dos testes de resíduos de Pearson e *Deviance*, os resultados obtidos para este modelo se encontram na tabela abaixo. Em nenhum deles há evidência para rejeitar a hipótese nula de que o modelo se encontra bem ajustado, a um nível de significância de 5%.

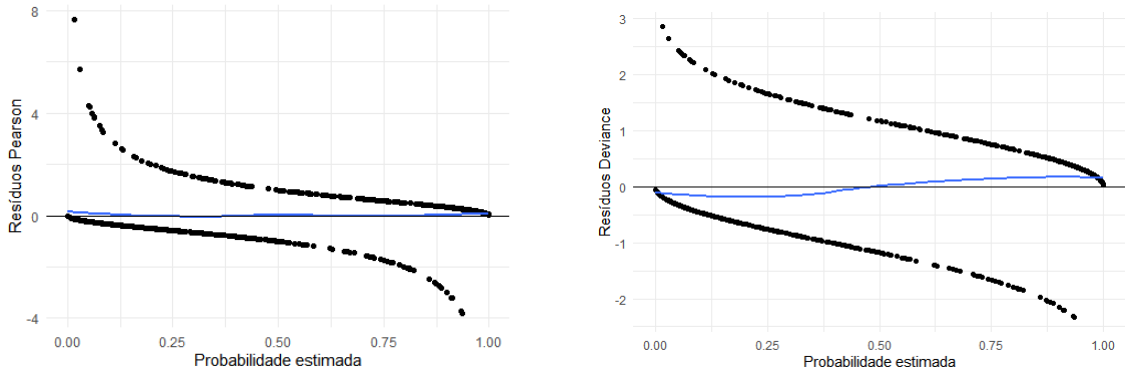
Tabela 24: Testes de adequabilidade modelo IRA.

Critério	Estatística	p-valor
Pearson	691,1567	0,757956
Deviance	570,683	0,999985

Os Gráficos 18a e 18b apresentam os resíduos de Pearson e *Deviance versus* a probabilidade estimada através do modelo de regressão logística. É possível observar um comportamento similar em ambos os gráficos. Se o modelo proposto está correto, a suavização de Lowess deve resultar em uma linha aproximadamente horizontal com

intercepto zero. Como as duas suavizações são aproximadamente uma linha com inclinação zero, não há evidências para concluir que o modelo esteja inadequado.

Gráfico 18: Resíduos modelo IRA. Licenciatura em Computação-UnB, 2012-2019



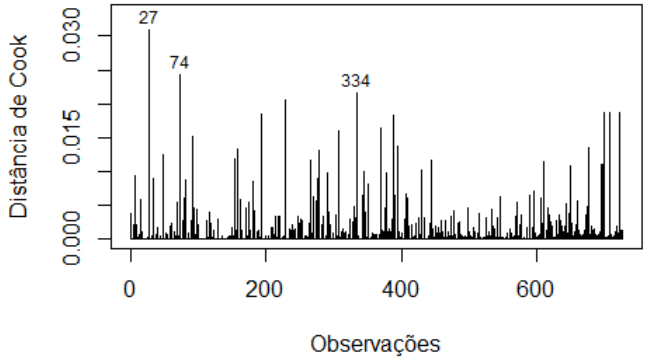
(a) Resíduos de Pearson

(b) Resíduos *Deviance*

- **Distância de Cook**

  A distância de Cook é utilizada para identificar possíveis observações influentes e o impacto delas nas estimativas dos parâmetros. O Gráfico 19 apresenta as distâncias de Cook para cada observação no modelo geral.

Gráfico 19: Distância de Cook modelo IRA. Licenciatura em Computação-UnB, 2012-2019



Nota-se que existem algumas observações consideradas influentes e por isso devem ser investigadas. A observação 27 por exemplo, trata-se um aluno ainda ativo com 15 semestres cursados, que adentrou no currículo antigo e possui 17 menções SR, que, apesar da variável não ter entrado no modelo, tem-se essa informação refletida no IRA, que é baixo. Já as observações 74 e 334 tratam-se de alunos que tiram IRA superior a 4, cursaram vários semestres do curso, taxa de reprovação zero, mas ainda assim evadiram do curso. Apesar de serem situações atípicas, são possíveis e

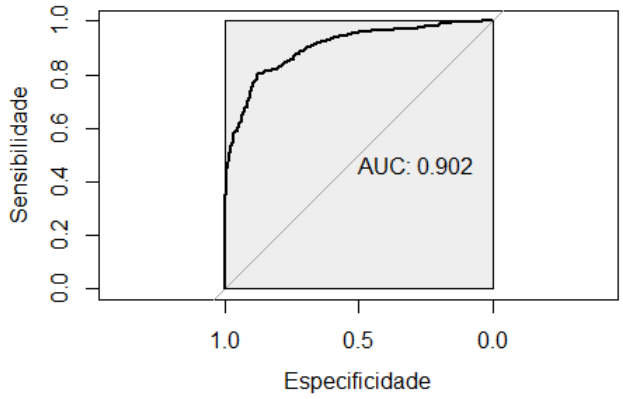
ocorreram algumas vezes, por isso foi decidido por mantê-las na modelagem.

- **Qualidade do ajuste - Curva ROC**

  Para verificar a qualidade do ajuste, a curva ROC (Receiver Operating Characteristic) é uma ferramenta para avaliação de um modelo de classificação, como o desenvolvido neste trabalho. A ROC mostra o quão bem um modelo criado pode diferenciar duas classificações, no caso deste trabalho, a evasão ou não evasão dos alunos do curso de Licenciatura em Computação.

  Como ressaltado anteriormente, a área debaixo da curva ROC (AUC) fornece a habilidade que o modelo tem que discriminar, sendo que quanto maior esse valor, melhor a capacidade de classificação do modelo. Para o modelo com IRA utilizando a base de construção, a área debaixo da curva foi de 0,9135, o que é considerado excelente. A seguir segue a curva ROC para o modelo geral.

Gráfico 20: Curva ROC modelo IRA. Licenciatura em Computação-UnB, 2012-2019



Para testar a capacidade preditiva do modelo, o modelo foi aplicado na amostra de validação e com os resultados de evasão observados e preditos, foi construído a matriz de confusão. A partir dela, e considerando um ponto de corte de 0,63, onde um valor predito maior que 0,63 é considerado evasão e abaixo ou igual não evasão, foi calculada a acurácia das predições do modelo nos dados da amostra de validação. A acurácia do modelo foi de 80,16%, indicando que o modelo acerta cerca de 80% todas as previsões.

4.5.2 Modelo Taxa de Reprovação

- **Testes de adequabilidade e resíduos**

  Para verificar a adequabilidade do modelo taxa de reprovação, aplicado o teste de Hosmer-Lemeshow, a estatística χ^2 obtida para esse modelo foi de 6,6554 com 8 graus de liberdade e um p-valor de 0,5742. Assim, não há evidências para rejeitar a hipótese que o modelo esteja bem ajustado.

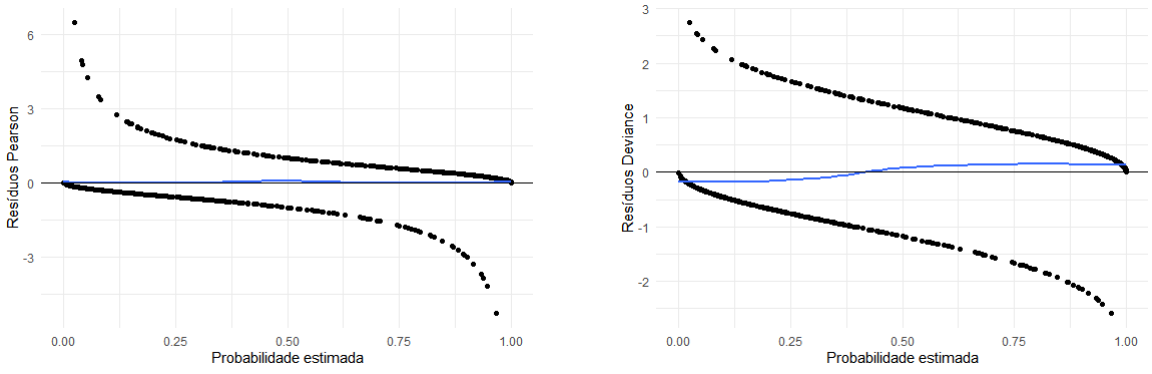
  Ao verificar a adequabilidade do ajuste através dos testes de resíduos de Pearson e *Deviance*, os resultados obtidos para este modelo se encontram na tabela abaixo. Aqui também, para nenhum deles há evidência para rejeitar a hipótese nula de que o modelo esteja bem ajustado, a um nível de significância de 5%.

Tabela 25: Testes de adequabilidade modelo taxa de reprovação.

Critério	Estatística	p-valor
Pearson	659,8545	0,9340451
Deviance	561,3476	0,9999948

Já ao olhar os resíduos do modelo taxa de reprovação, as mesmas considerações para os resíduos de Pearson e *Deviance* do modelo IRA podem ser utilizados aqui, visto que seus resultados são semelhantes.

Gráfico 21: Resíduos modelo taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019



(a) Resíduos de Pearson

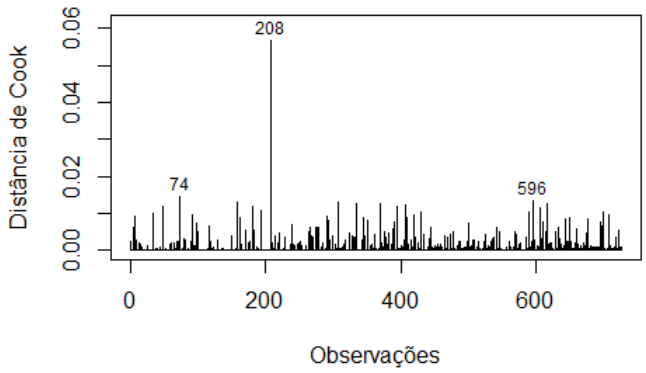
(b) Resíduos *Deviance*

- **Distância de Cook**

  A distância de Cook, como dito anteriormente, é utilizada para identificar possíveis pontos influentes. O Gráfico 22 apresenta a distância de Cook para cada observação

do modelo taxa de reprovação geral.

Gráfico 22: Distância de Cook modelo taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019

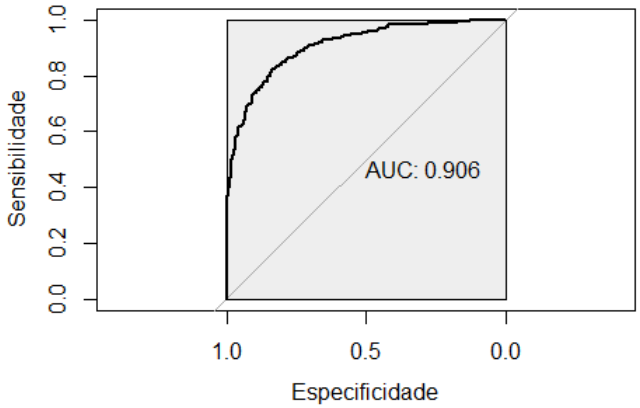


Para esse modelo, alguns pontos também precisam ser investigados. A observação 74 corresponde a um aluno que possui taxa de reprovação zero, nenhum SR e permaneceu no curso por 10 semestres, mas ainda assim evadiu. Já a observação 208 refere-se a um estudante que cursou verão, teve 10 SRs e cursou 11 semestres e evadiu do curso. Por fim, a observação 596 corresponde a um aluno com taxa de reprovação elevada igual a 1, que obteve 5 SRs, ainda ativo no curso com 4 semestres cursados. Essas situações são atípicas, mas ocorrem dentro do ambiente de curso, por isso ainda assim serão mantidas na análise.

- **Qualidade do ajuste - Curva ROC**

  A qualidade do ajuste para o modelo taxa de reprovação também foi verificada através da curva ROC. A área debaixo da curva ROC (AUC) fornece a habilidade de classificação do modelo. Para esse modelo, para a amostra de construção a AUC foi de 0,914, considerado muito bom. A figura abaixo apresenta a curva ROC e o AUC para o modelo geral.

Gráfico 23: Curva ROC modelo taxa de reprovação. Licenciatura em Computação-UnB, 2012-2019



Uma vez que a discriminação do modelo é muito boa, esse foi aplicado na amostra de validação e com os resultados dos valores preditos e observados, foi calculado a acurácia sobre os dados de validação. A acurácia do modelo foi de 79,88%, o que indica que o modelo acerta cerca de 80% de suas predições.

5 Conclusão

A motivação desse trabalho surgiu de um problema real: a evasão acadêmica no curso de licenciatura de Licenciatura em Computação da Universidade de Brasília e tem como objetivo de identificar quais são as características dos estudantes que estão associadas a evasão acadêmica e traçar o perfil dos alunos evadidos.

O curso de Licenciatura em Computação da UnB é composto majoritariamente por alunos do gênero masculino, cerca de 88% e, de todos os alunos, metade deles tem até 21 anos de idade ao ingressar no curso. Mais de 70% dos estudantes residiam em locais considerados de alta ou média alta renda quando ingressaram na UnB. A forma de ingresso mais utilizada por esses alunos é o vestibular da UnB.

Na vida dentro da universidade, os alunos costumam cursar poucos semestres do curso e poucos deles alcançam a marca de 9 semestres cursados, o mínimo para formar. O curso possui uma evasão de 52% dos estudantes avaliados, chegando a 80% ao avaliar a evasão por período de ingressos mais antigos.

Na construção do modelo de regressão logística optou-se por construir dois modelos, um considerando IRA entre as variáveis explicativas e outro considerando a taxa de reprovação, dada a alta correlação entre estas variáveis explicativas. Para o modelo IRA, demais fatores como forma de ingresso, currículo, semestres cursados e cursou verão foram significativas no modelo para explicar a evasão do curso. Aqui uma característica que se destaca é de que alunos do currículo velho possuem cerca de 16 vezes mais chances de evadir que os alunos do currículo novo. Vê-se aqui que a mudança de currículo ocorrida em 2015/2 está surgindo um efeito positivo. Para o modelo com taxa de reprovação, os fatores significativos foram os mesmos além da variável quantidade de menções SR.

Ambos os modelos encontrados, tanto o com IRA tanto o com taxa de reprovação apresentaram um bom ajuste com uma acurácia em torno dos 80%. Como ambos possuem resultados parecidos, a sugestão seria trabalhar com o modelo IRA, uma vez que ele é mais simples e as informações de IRA já são facilmente calculadas pela Universidade de Brasília.

Para trabalhos futuros, sugere-se como proposta:

- Ampliar o número de possíveis variáveis explicativas, coletando por exemplo dados que tragam uma perspectiva socioeconômica do aluno;
- Avaliar mais afundo o desempenho em disciplinas chaves do curso;

- Realizar um estudo semelhante por gênero para avaliar se realmente não há diferenças na evasão, dado que há uma disparidade expressiva no número de estudantes por gênero;
- Realizar um estudo semelhante para outros cursos de graduação da Universidade de Brasília.

Esse estudo buscou identificar fatores estão associados a evasão acadêmica do curso de Licenciatura em Computação e com isso fomentar mais estudos sobre o tema abordado e assim servir como base para elaborações de políticas voltadas ao combate da evasão no curso Licenciatura em Computação da Universidade de Brasília.

Referências

BARROSO, M. F.; FALCÃO, E. B. Evasão universitária: o caso do instituto de física da ufrj. *IX Encontro Nacional de Pesquisa em Ensino de Física*, v. 9, p. 1–14, 2004.

BRAGA, M. M.; CARMO, L. P. Maria do; BOGUTCHI, T. F. A evasão no ensino superior brasileiro: o caso da ufmg. *Avaliação: Revista da Avaliação da Educação Superior*, v. 8, n. 3, 2003.

BRASIL, M. Comissão especial de estudos sobre a evasão nas universidades públicas brasileiras. *http://www. dominiopublico. gov. br/download/texto/me001613. pdf Acesso em 19/09/2021*, v. 15, n. 01, p. 2007, 1997.

DIAS, E. C. M.; THEÓPHILO, C. R.; LOPES, M. A. Evasão no ensino superior: estudo dos fatores causadores da evasão no curso de ciências contábeis da universidade estadual de montes claros–unimontes–mg. In: *Congresso USP de Iniciação Científica em Contabilidade*. [S.l.: s.n.], 2010. v. 7, p. 1–16.

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

INEP. *Informe estatístico do MEC revela melhoria do rendimento escolar*. 1998. Disponível em: ⟨http://portal.inep.gov.br/artigo/-/asset_publisher/B4AQV9zFY7Bv/content/informe-estatistico-do-mec-revela-melhoria-do-rendimento-escolar/21206⟩.

MCMILLAN, J. Course change and attrition from higher education. *LSAY research reports*, p. 43, 2005.

SEMESP, I. *Mapa do Ensino Superior no Brasil. 10ª Edição*. 2020. Disponível em: ⟨https://www.semesp.org.br/wp-content/uploads/2020/04/Mapa-do-Ensino-Superior-2020-Instituto-Semesp.pdf⟩.

TINTO, V. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 45, n. 1, p. 89–125, 1975.

Apêndice

A Tabela do sistema de cotas por período de ingresso

Período	Não	Sim
2012/2	93%	7%
2013/1	89%	11%
2013/2	86%	14%
2014/1	72%	28%
2014/2	80%	20%
2015/1	74%	26%
2015/2	84%	16%
2016/1	67%	33%
2016/2	74%	26%
2017/1	54%	46%
2017/2	78%	22%
2018/1	69%	31%
2018/2	84%	16%
2019/1	46%	54%
2019/2	73%	23%