Universidade de Brasília
Departamento de Estatística

Análise de fatores que impactam no êxito de alunos em estado de vulnerabilidade social na UnB.

Juliana Paula Degani

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

Brasília
2022

Juliana Paula Degani

Análise de fatores que impactam no êxito de alunos em estado de vulnerabilidade social na UnB.

Orientador(a): Prof. Leandro Tavares Correia

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

Brasília
2022

Agradecimentos

Gostaria de agradecer em primeiro lugar aos meus pais Fabio e Jackeline Degani por sempre estarem presentes em minha trajetória acadêmica dando todo o apoio e incentivo possível.

Também agradeço à minha irmã Anelise Degani, pelo apoio emocional, pelo companheirismo e as risadas que deixaram todo esse caminho mais leve.

É claro que tenho um imenso agradecimento ao meu orientador Leandro Tavares, que foi tão prestativo e paciente durante todo o processo, por seus ensinamentos e orientações, foi um privilégio ser orientada pelo mesmo.

Agradeço também a professora Ana Maria Nogales, que admiro e respeito muito, uma grande inspiração como profissional.

À professora Juliana Betini Fachini, que ministrou a minha primeira aula na graduação, me marcou muito pela pessoa que é, pelo amor que tem pela profissão e pelos alunos.

Às minhas amigas desde a primeira semana de aula, pelo nosso "vetor aleatório de menina", Luísa, Carolina, Helena, Maria Eduarda e Yasmin, que em meio a muitas horas de estudo e trabalhos em grupo, tornaram essa experiência divertida e especial.

Aos meus amigos de curso, Amanda, Bruno, Victor Rezende, Beatriz, Rodrigo Berçott, Rayssa e Louise, por todo o companheirismo e horas auxílio com as matérias.

À Marilia Ferreira, servidora da Diretoria de Desenvolvimento Social da UnB que foi tão paciente e prestativa, me auxiliou em todo o processo para conseguirmos os dados de maneira que o presente trabalho se tornou possível.

Resumo

Este trabalho desenvolve uma análise estatística em um dos principais programas de auxílio estudantil existentes na Universidade de Brasília, o Auxíilio Socieconômico. Através dos dados de alunos que receberam o benefício no período compreendido entre 2016 e 2021, foi utilizada a técnica de regressão logística, por meio do software estatístico R.

Foi calculada a probabilidade de formatura por meio de um modelo de regressão logística binária envolvendo a formatura e a desistência desses alunos. Também foi ajustado um modelo de regressão logística multinomial incluindo os alunos ativos e observando o impacto no modelo.

Palavras-chave: Auxílio socioeconômico, UnB, vulnerabilidade social, assistência estudantil, formatura, evasão, regressão logística.

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

A desigualdade social é atualmente um grande problema sistêmico no Brasil e no mundo, sendo atualmente pauta de muitas ações para promover a redução destas desigualdades, pois ainda existe uma elevada concentração de renda nas mãos de uma parte seleta da população.

Embora o Brasil esteja entre os dez países com maior PIB - Produto Interno Bruto, ele é o oitavo país com o maior índice de desigualdade social e econômica do mundo. Segundo relatório de ONU (2010) as principais causas da desigualdade social são:

- Falta de acesso à educação de qualidade;
- Política fiscal injusta;
- Baixos salários;
- Dificuldade de acesso a serviços básicos de saúde, transporte público e saneamento básico.

Um estudo apresentado por Trovão (2020), baseado nos dados da Relação Anual de Informações Sociais (RAIS) de 2018, compara o rendimento domiciliar dos 10% mais ricos com os 10% mais pobres, e foi constatada uma diferença abismal. Onde os 10% mais ricos ganham em média R$7.302,46, enquanto os 10% mais pobres ganham em média R$117,96. Essa diferença também é observada dentro das Universidades brasileiras.

A Universidade de Brasília possui grande diversidade no perfil de seus alunos, e uma parcela que desperta atenção são os alunos em estado de vulnerabilidade socioeconômica e/ou risco social. São classificados assim pela UnB os alunos em estado de vulnerabilidade social, aqueles que possuem renda familiar de até um salário mínimo e meio.

Tendo em vista esse cenário, a Presidência da República publicou o decreto Republica (2010), que dispõe sobre o Programa Nacional de Assistência Estudantil (PNAE), que tem como objetivo principal ampliar as condições de permanência dos jovens na educação superior pública federal.

Nesse decreto o governo normatiza ações de assistência estudantil, que são: moradia estudantil, alimentação, transporte, atenção à saúde, inclusão digital, cultura, esporte, creche, apoio pedagógico e acesso, participação e aprendizagem de estudantes com deficiência, transtornos globais do desenvolvimento e altas habilidades e superdotação. Estas ações visam tornar o ensino superior em universidades públicas uma realidade possível dentre os jovens em estado de vulnerabilidade social.

Seu objetivo principal é minimizar os efeitos da desigualdade social, reduzindo as taxas de evasão do ensino e diminuição do tempo até a formatura, contribuindo desta maneira para a inclusão social.

Com base nessa necessidade, a Universidade de Brasília instituiu como responsável por alguns desses auxílios a DDS (Diretoria de Desenvolvimento Social) vinculada à DAC (Decanato de Assuntos Comunitários) que operacionaliza e monitora esses programas.

Dentre os auxílios proporcionados pela DDS, destaca-se o Auxílio Socieconômico, que segundo o site oficial da universidade (UNB, 2021), oferece bolsas mensais de R$ 465,00 (quatrocentos e sessenta e cinco reais) para os alunos selecionados no programa. Essa bolsa atualmente não exige contrapartida do estudante, não cobra desempenho nas matérias cursadas, nem justificativa de como esse dinheiro é gasto.

A seleção é feita com base em um critério principal, que é estar em estado de vulnerabilidade, além de outros fatores de desempate que são: turno do curso, identidade de gênero, nacionalidade, UF/região de proveniência, egresso de escola pública, participante do sistema de cotas raciais, situação de moradia do estudante e situação de moradia do grupo familiar.

Outros estudos com relação ao tema relatam sobre a importância dos auxílios oferecidos pelas universidades públicas aos alunos em estado de vulnerabilidade social. Um exemplo é o artigo Sant'anna e Almeida (2021), que também fala sobre o decreto Republica (2010), e sobre o desempenho da DDS na UnB, ressaltando que "A formação dos estudantes em vulnerabilidade socioeconômica impõe desafios complexos para a assistência estudantil", relacionando com a importância das políticas de assistência social na universidade. O autor coletou os dados de pessoas vinculadas à própria DDS e os utilizou na Escala de Likert (1932) e realizou a Análise de Cluster (AC), técnicas diferentes das escolhidas para o presente projeto.

Tendo em vista a relevância do tema, e o seu agravamento das complicações sociais decorrentes da Pandemia COVID-19, o mesmo foi escolhido visando auxiliar a universidade no entendimento dos resultados alcançados pelo auxílio vigente, e tambem fornecer insumos que permitam que a UnB aprimore cada vez mais este importante programa.

O objetivo principal deste trabalho é avaliar as características sociais dos alunos que usufruem do benefício **Auxílio Socieconômico**, e quais os impactos na probabilidade de formatura.

2 Revisão de Literatura

As técnicas estatísticas utilizadas no presente trabalhos estão descritas nessa sessão.

2.1 Distribuições de Probabilidade

Com base nos autores Bussab e Morettin (2017) e Magalhães (2015) serão utilizadas as definições abaixo.

2.1.1 Distribuição de Bernoulli

Para o presente trabalho foi utilizado como base os conhecimentos da distribuição de probabilidade de Bernoulli. Essa distribuição é utilizada nos casos em que são trabalhados experimentos a fim de obter resultados a serem classificados como "sucesso"ou "fracasso".

Para isso uma variável aleatória X é definida, assumindo o valor 1 em caso de "sucesso"e 0 em caso de "fracasso". Sendo assim assume-se que p é a probabilidade de se obter o "sucesso", com $0 < p < 1$.

Então essa variável aleatória X que segue a distribuição de Bernoulli possui função de probabilidade $(x, p(x))$ onde a probabilidade de sucesso é dada por:

$$P(X = 1) = p,$$

, e a probabilidade de fracasso é:

$$P(X = 0) = 1 - p.$$

E sua notação é dada por $X \sim \text{Ber}(p)$.

2.1.2 Distribuição Binomial

Também foi utilizado o conceito acerca da distribuição Binomial, que é explicada como sendo a repetição de n ensaios independentes de Bernoulli. Nesse caso supomos n repetições independentes do experimento, onde em cada uma das repetições existe a probabilidade p de "sucesso"ou "fracasso", com $0 < p < 1$.

Assim define-se como X uma variável aleatória que indica o número total de

sucesso dentre as n repetições.

Para calcular a probabilidade de se obter k $(k = 0, 1, ..., n)$ sucessos dentre as n repetições independentes utiliza-se a função de probabilidade:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$

Sua notação é dada por $X \sim B(n, p)$.

2.1.3 Distribuição Normal

Um conceito fundamental na Estatística é a distribuição normal de probabilidade. Também conhecida como Distribuição Gaussiana, sua função de densidade de probabilidade se aproxima das curvas de frequência de medidas físicas, como pode-se observar na figura abaixo:

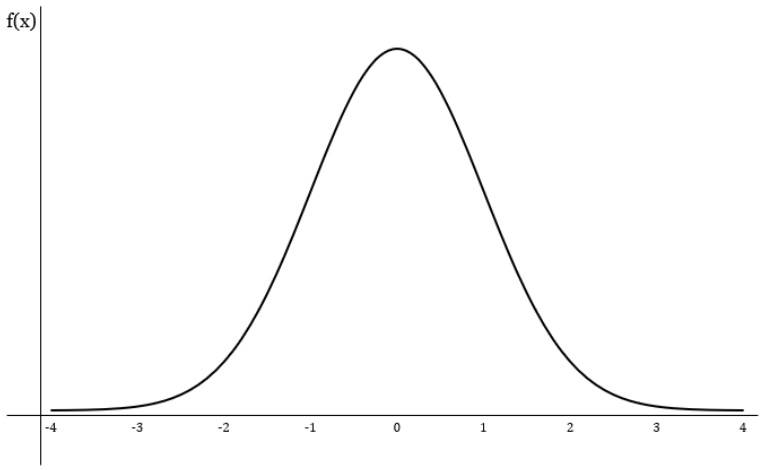

Figura 1: Distribuição Normal

Fonte: Site PróEducacional, disponível em: https://proeducacional.com/ead/curso-cga-modulo-i/capitulos/capitulo-4/aulas/distribuicao-de-probabilidades-distribuicao-normal/

Essa distribuição se utiliza dos parâmetros média μ e variância σ^2. Para calcular a sua função de densidade de probabilidade, a fórmula abaixo é utilizada:

$$f_x(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{1}{2}(\frac{x-\mu}{\sigma})^2},$$

onde $-\infty < x < \infty$.

2.1.4 Distribuição Multinomial

Na distribuição multinomial, foi trabalhado com o caso de n repetições independentes de um experimento que tem k resultados possíveis, com respectivas probabilidades de sucesso dadas por: $p_1, p_2, ..., p_k$ e $\sum_{i=1}^{k} P_k = 1$.

Sendo assim, vamos estudar o vetor aleatório $(X_1, X_2, ..., X_k)$, onde X_i é o número de ocorrências do i-ésimo resultado.

Segundo Oliveira e Silva (2018) os requisitos para uma distribuição ser considerada Multinomial são:

- Possuir n ensaios;
- Cada ensaio possui um número discreto de resultados possíveis;
- os ensaios são independentes entre si, com probabilidade constante de ocorrer um determinado resultado.

A distribuição de probabilidades é dada por:

$$P(X_1 = x_1, X_2 = x_2, ..., X_k = x_k) = \frac{n!}{x_1!x_2!, ..., x_k!} p_1^{x_1} p_2^{x_2} ... p_k^{x_k}.$$

Onde os x_is são números inteiros não negativos que satisfazem $\sum_{i=1}^{k} x_i = n$ e $\sum_{i=1}^{k} p_i = 1$.

2.2 Testes de Hipótese

O teste de hipótese é uma técnica estatística que tem como objetivo fornecer uma metodologia para verificar se os dados das amostras possuem indicativos que comprovem, ou não, uma hipótese previamente formulada.

$$\begin{cases} H_0 : \text{hipótese a ser testada (chamada de hipótese nula)} \\ H_1 : \text{hipótese alternativa que será aceita caso a hipótese nula seja rejeitada.} \end{cases}$$

Essa decisão é tomada por meio do p-valor, com base em um nível de significância, que normalmente é de 5%, segundo Dávila (2017).

Os testes estatísticos que foram utilizados como base estão descritos abaixo.

2.2.1 Teste qui-quadrado

O teste utiliza como base a estatística χ^2, apresentando mudanças nos graus de liberdade da sua distribuição de acordo com o teste que será utilizado. No geral,

$$\chi_v^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

em que v expressa o número de graus de liberdade, o_i é a frequência observada e e_i é chamado de valor esperado e representa a frequência que seria observada se H_0 fosse verdadeira.

A distribuição da estatística do teste segue a distribuição qui-quadrado com v graus de liberdade. Com $v = (r-1) * (c-1)$ graus de liberdade, sendo r e c são, respectivamente, o número de classes de resultados e o número de informações da amostra.

2.2.2 Teste de Wald

O teste de Wald é utilizado para verificar se o valor dos parâmetros de entrada verdadeiros tem a mesma probabilidade que os parâmetros calculados pela Estimativa de Máxima Verossimilhança.

Basicamente quanto maior o valor da estatística Wald, menor é a probabilidade de que os parâmetros de entrada sejam verdadeiros.

Ele é feito com base nas hipóteses:

$$\begin{cases} H_0 : \beta_j = \beta_j^{(0)} \\ H_1 : \beta_j \neq \beta_j^{(0)} \end{cases}$$

Em que $\beta_j^{(0)}$ é algum valor postulado para β_j. Esse teste é feito com base na estatística:

$$Z_t = \frac{\hat{\beta}_j - \beta_j^{(0)}}{\sqrt{Var(\hat{\beta}_j)}},$$

que sobre a hipótese nula, segue a distribuição Normal padrão.

3 Metodologia

Esta sessão discorre sobre as principais metodologias e ferramentas utilizadas no desenvolvimento do trabalho.

3.1 Banco de Dados

3.1.1 Origem

Tendo em vista esse contexto temos como base um banco de dados fornecido pela Diretoria de Desenvolvimento Social (DDS), que contém informações dos alunos que foram beneficiados pelo programa de auxílio pedagógico de 2016 a 2021.

A DDS possui registro anterior a 2016, entretanto foi optado por não trabalhar com esses dados por não serem tão confiáveis quanto os mais recentes, tendo em vista que o sistema no qual a UnB registra essas informações foi alterado algumas vezes e com isso ocorreu perda na qualidade de algumas informações, o que justifica a escolha do corte temporal utilizado no presente trabalho.

A solicitação destas informações estão registradas no sistema SEI do governo federal, sob a identificação 23106.045 687/2022-43.

3.1.2 Variáveis

O banco em questão possui as variáveis: Sexo, Data de nascimento, Raça/cor, Curso, Campus, Semestre/ano de ingresso, Tipo de ingresso (vestibular/PAS,SISU), Tipo de cota de ingresso (cota/universal), Ensino médio (escola pública/privada), Semestre/ano de ingresso nos programas, Trancamentos, Se está ativo na Universidade, Semestre/ano de saída. Com base nas informações disponibilizadas foi efetuado o estudo com relação ao tempo de formatura desses alunos e sua probabilidade de formatura.

3.2 Estatística Descritiva

Foi realizada a análise descritiva de todas as variáveis do banco de dados para se obter um melhor entendimento das variáveis e do perfil desses alunos que são atendidos pelo programa de auxílio socieconômico da UnB.

3.3 Dashboard

Com a finalidade de auxiliar na compreensão dos dados por parte da Diretoria de Desenvolvimento Social (DDS) foi construído um Dashboard contendo as informações do banco de dados de forma descritiva. Essa forma visual de apresentar os dados facilita a compreensão do perfil dos alunos que usufruem do benefício de auxílio pedagógico fornecido pela Universidade para esses alunos em estado de vulnerabilidade socioeconômica.

Cabe ainda uma aprimoramento sobre os dados a serem apresentados, de modo que o painel possa refletir somente os dados mais relevantes para a análise gerencial da área.

Para esta construção foi utilizada a ferramenta **DataStudio** por ser gratuita e de fácil compressão para os usuários.

Figura 2: DataStudio

Fonte: Blog DPC, Disponível em:
https://blog.dp6.com.br/8-hacks-do-data-studio-que-voc%C3%AA-precisa-conhecer-f602448e048a

3.4 Modelo Logístico

Como a variável resposta que será utilizada no modelo é uma variável categórica que segue a distribuição Binomial (Sucesso ou Fracasso), foi utilizado o modelo de regressão logística Binária.

Na Figura 3 é possível observar um exemplo ilustrativo desse caso.

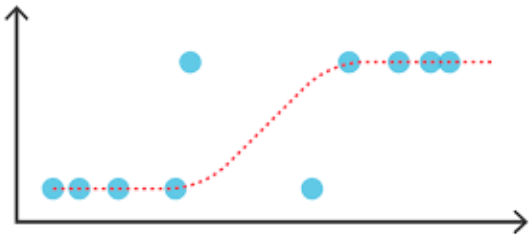

Figura 3: Exemplo Modelo Logístico

Fonte: Site tibco, Disponível em:
https://www.tibco.com/pt-br/reference-center/what-is-logistic-regression

Na Figura 3 é possível observar no eixo Y, a variável resposta categórica que assume somente duas categorias, e a linha vermelha seria um possível modelo que se ajusta aos dados.

3.4.1 Análise de Regressão Logística Binária

Segundo Hosmer e Lemeshow (2000) a técnica consiste em criar um modelo de regressão para uma variável resposta que é dicotomizada em "sucesso"ou "fracasso". Esse modelo traz como resultado a probabilidade de "sucesso".

A modelagem é feita por meio da equação:

$$P(Y_i = 1|X_{1i}, \ldots, X_{pi}) = \pi(X_i) = \frac{e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}}{1 + e^{\beta_0+\beta_1 X_{1i}+\ldots+\beta_p X_{pi}}},$$

em que a probabilidade de sucesso da variável resposta ($Y = 1$) está em função das variáveis explicativas X_i, $i = 1, 2, \ldots, p$.

Tal equação pode ser escrita de maneira linear pela transformação *logito*:

$$\pi^*(X_i) = ln\left(\frac{\pi(X_i)}{1 - \pi(X_i)}\right) = \beta_0 + \beta_1 X_{1i} + \ldots + \beta_p X_{pi}.$$

O parâmetro β_j corresponde à taxa de crescimento ou redução da curva para cada unidade de X_j sobre o logaritmo neperiano da chance de sucesso ($Y = 1$), mantendo as demais variáveis constantes. Dessa forma, e^{β_j} tem como efeito a multiplicação na *odds* de $Y = 1$ para o aumento de uma unidade de X_j, mantendo as variáveis constantes. A *odds*, mencionada no parágrafo anterior, é uma razão, dada pela probabilidade de um evento acontecer dividido pela probabilidade dele não ocorrer. Ao deixar as demais variáveis constantes, por meio da odds é possível saber, para cada impacto na variável X, qual o aumento da chance de ocorrer o evento sobre a chance do mesmo não ocorrer.

3.4.2 Análise de Regressão Logística Multinomial

O modelo logístico binário é um caso particular do modelo Multinomial, que é usado em situações onde a variável dependente tem múltiplas categorias. O modelo possui uma expressão alternativa em termos das respostas com múltiplas categorias:

$$\pi_{ij} = \frac{e^{\alpha_j+\beta_j x}}{\sum_h e^{\alpha_h+\beta_h x}}, \qquad j = 1, ..., J.$$

- $\sum_j \pi_j = 1$
- j representa a j-ésima categoria das J categorias da variável resposta.

O parâmetro β_j corresponde ao efeito do aumento de uma unidade de X_j sobre o logaritmo neperiano da chance de sucesso ($Y = 1$), mantendo as demais variáveis constantes, de uma categoria de referência com alguma outra das demais. Dessa forma, e^{β_j} tem como efeito a multiplicação na *odds* de $Y = 1$ para o aumento de uma unidade de X_j, mantendo as variáveis constantes, de uma categoria de resposta com uma outra categoria de referência.

Isso por ser realizado pelos testes: Teste de Wald, Teste de Razão de Verossimilhança e/ou Teste Score.

3.4.3 Teste da significância dos Coeficientes

Após ajustar o modelo é essencial avaliar a significância de cada um dos coeficientes, o que envolve a formulação de um teste com as seguintes hipóteses:

$$\begin{cases} H_0 : \beta_j = 0, \text{ na ausência de influência dessa variável} \\ H_1 : \beta_j \neq 0, \text{ influência dessa variável} \end{cases}$$

Ou seja, verifica-se se as variáveis independentes no modelo estão "significativamente"relacionadas à variável independente.

Para isso utiliza-se a distribuição assintótica dos estimadores de máxima verossimilhança e obtém-se distribuição normal para estatística do teste, que é dada por:

$$Z = \frac{\hat{\beta}_j}{\widehat{SE}(\hat{\beta}_j)}.$$

Onde:

- $\hat{\beta}_j$ é o parâmetro estimado de β_j;
- $\widehat{SE}(\beta_j)$ é o estimador do erro padrão para o respectivo estimador do parâmetro.

3.4.4 Teste de razão de Verossimilhança

Esse teste compara a verossimilhança sob a suposição de hipótese nula verdadeira, (θ_o), com a maior verossimilhança $(\hat{\theta})$. Segundo as hipóteses:

$$\begin{cases} H_0 : \text{Não existe diferença significativa entre os modelos} \\ H_1 : \text{Existe diferença significativa entre os modelos} \end{cases}$$

A estatística do teste é dada por:

$$Q_L = -2log\left[\frac{L(\theta_0)}{L(\hat{\theta})}\right].$$

Que segue a distribuição qui-quadrado com dois graus de liberdade. Onde,

- Q_L é a estatística do teste;
- θ_0 é o estimador de máxima verossimilhança
- $\hat{\theta}$ o estimador de máxima verossimilhança em H_0
- L representa a função de verossimilhança.

3.4.5 Intervalo de Confiança para os Coeficientes

O cálculo e a interpretação dos intervalos de confiança para os parâmetros de interesse compõem um importante passo na hora de testar a significância do modelo. Esses estimadores para o intervalo de confiança são baseados em seus respectivos testes Wald.

O intervalo para o intercepto, que é o β_0 é dado por:

$$\hat{\beta}_0 \pm z_{1-\alpha/2}\widehat{SE}(\beta_0).$$

E o intervalo dos demais coeficientes, β_i, $1 = 1, 2, ...$ são dados por:

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

Onde:

- $z_{1-\alpha/2}$ é o ponto superior de $100(1 - \frac{\alpha}{2})$ da distribuição normal parão;
- $\widehat{SE}(\beta_i)$ é o estimador do erro padrão para o respectivo estimador do parâmetro.

3.5 Avaliar a Qualidade do Ajuste

Existem alguns métodos para avaliar a qualidade do ajuste de modelos de regressão logística, com a suposição inicial de que estamos preliminarmente satisfeitos com os modelos candidatos.

3.5.1 Erro quadrático médio

Segundo (RODRÍGUEZ, 2020) o Erro quadrático médio é uma forma de medir o desempenho da função ou função de perda é denominada como MSE, (Mean Squared Error), que avalia a qualidade da função de previsão, que é dado pela fórmula:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

Onde n é o tamanho da amostra, y_i são os valores observados na amostra e $\hat{y}_i$ os respectivos valores estimados.

3.5.2 Critério de Informação de Akaike (AIC)

O critério de informação de Akaike é calculado por:

$$AIC = -2\log(L(\hat{\theta})) + 2K, \tag{3.5.1}$$

onde:

- K é o número de parâmetros do modelo;
- n é o tamanho da amostra em questão.

3.5.3 Critério de Informação Bayesiano (BIC)

O critério de informação Bayesiano é calculado por:

$$BIC = -2\log(L(\hat{\theta})) + K\log(n), \tag{3.5.2}$$

onde:

- n é o tamanho da amostra em questão.

Os modelos que apresentam menores valores de BIC e AIC são preferíveis.

3.5.4 Deviance

A estatística *Deviance* é utilizada para avaliar o ajustamento do modelo aos dados, seguindo a expressão matemática:

$$deviance = -2\log(L(\hat{\theta} + log(L(\hat{\theta}_{ModeloSaturado})),$$

onde:

- L é a função de verossimilhança do modelo;

Em geral, os modelos com um menor *deviance* se encaixam melhor do que os modelos com maior *deviance*.

3.5.5 Teste de Hosmer-Lemeshow

Esse teste foi proposto por Hosmer e Lemeshow com base no agrupamento dos valores das probabilidades estimadas, para avaliar se o modelo ajusta bem os dados.

Foram propostas duas estratégias diferentes de agrupamento, a primeira ocorre ao escolher a tabela baseada nos percentis das probabilidades estimadas, a segunda estratégia é utilizada escolhendo a tabela com base nos valores fixos da probabilidade estimada.

Independentemente da estratégia de agrupamento utilizada, o teste é feito com base na estatística de Hosmer-Lemeshow ($\hat{C}$), que é obtida calculando a estatística de Pearson a partir da tabela de g linhas e 2 colunas observadas nos dados e as frequências estimadas.

Essa estatística é baseada no χ^2, sendo calculada pela fórmula abaixo.

$$\hat{C} = \sum_{k=1}^{g} \frac{(O_k - \acute{n}_k\overline{\pi_k})^2}{\acute{n}_k\overline{\pi_k}(1-\overline{\pi_k})}.$$

Em que:

- $\acute{n}_k$ é o número total de indivíduos no k-ésimo grupo;
- $O_k = \sum_{c_k}^{j=1} y_j$ é o número de respostas entre os padrões de covariáveis c;
- c é o número de covariáveis padrões no k-ésimo decil;
- $\overline{\pi_k} = \sum_{c_k}^{j=1} \frac{m_k\hat{\pi_j}}{\acute{n}_k}$ é a média das probabilidades estimadas.

3.6 Interpretação dos Parâmetros

Após um modelo de regressão logística ser construído, ajustado e ter tido sua significância testada é importante se ter uma interpretação clara do que significam os parâmetros desse modelo. A partir da interpretação dos parâmetros, é possível extrair inferências práticas dos coeficientes estimados.

O intercepto do modelo indica onde o modelo cruza o eixo Y. Ele é importante quando o modelo está em um contexto onde seria razoável as variáveis explicativas serem iguais a zero, que indica o ponto onde o modelo cruza o eixo Y.

Os coeficientes estimados para as variáveis independentes representam a inclinação, ou seja, a taxa de variação, de uma função da variável dependente por unidade de variação da variável independente.

Basicamente o processo de interpretação dos parâmetros envolve determinar a relação funcional entre a variável dependente e a variável independente e definir a unidade de mudança da variável independente.

No modelo de regressão logística, o β_j indica a mudança na variável preditora correspondente a uma mudança de uma unidade na variável independente.

3.6.1 Interação e Confundimento

O termo confundidor é usado por epidemiologistas para descrever uma covariável que está associada tanto à variável de resultado de interesse quanto a uma variável independente primária ou fator de risco. Quando ambas as associações estão presentes, diz-se que a relação entre o fator de risco e a variável de resultado está confundida.

Um método para verificar o status do confundidor de uma covariável é comparar o coeficiente estimado para a variável fator de risco de modelos contendo e não contendo a covariável. Qualquer mudança "estatisticamente comprovada"no coeficiente estimado para o fator de risco sugere que a covariável é um fator de confusão e deve ser incluída no modelo, independentemente da significância estatística de seu coeficiente estimado.

3.7 Medidas Preditivas

A qualidade do modelo de pode ser avaliada a partir de algumas medidas que falam sobre o seu poder preditivo, o qual é expresso pelas classificações corretas dado o ajuste do modelo.

Neste trabalho, o poder preditivo foi analisado com base na matriz de confusão,

curva ROC, acurácia, sensibilidade e especificidade.

A **matriz de confusão**, ilustrada na Tabela 1, consiste em uma tabela entre os valores preditos pelo modelo e os valores reais, mostrando a relação entre os verdadeiros positivos, falsos positivos, falsos verdadeiros e falsos negativos. A partir dessa matriz é possível calcular as demais medidas preditivas.

Tabela 1: Matriz de confusão

		Amostra	
		Fracasso	Sucesso
Previsto	Fracasso	**TN**	**FN**
	Sucesso	**FP**	**TP**

Em que,

- **Verdadeiro Positivo (TP)**: Os valores são verdadeiros e foram classificados como verdadeiros;
- **Verdadeiro Negativo (TN)**: Os valores são falsos e foram classificados como falsos;
- **Falso Positivo (FP)**: Os valores são falsos e foram classificados como verdadeiros;
- **Falso Negativo (FN)**: Os valores são verdadeiros e foram classificados como falsos.

A partir das informações da matriz de confusão são calculadas as medidas preditivas do modelo.

A **Acurácia** de um modelo fala sobre a probabilidade dele acertar as previsões, com base dos dados de validação. Seu cálculo é feito por meio da fórmula:

$$Acurácia = \frac{TP + TN}{TP + TN + FP + FN}.$$

A **Sensibilidade** é um indicador baseado em proporção do resultado do modelo, cujo objetivo é apresentar o quão bom o modelo na previsão de sucessos, dado pela fórmula:

$$Sensibilidade = \frac{TP}{TP + FN}.$$

A **Especificidade** também é um indicador baseado em proporção do resultado do modelo, mas o objetivo é apresentar quão bom o modelo na previsão dos fracassos, através da fórmula:

$$Especificidade = \frac{FP}{FP + TN}.$$

Após o cálculo desses indicadores a **curva ROC** pode ser construída, a qual é um gráfico da sensibilidade em função de 1 menos a especificidade para diferentes valores da probabilidade de sucesso, resumindo o poder preditivo do teste.

3.8 Seleção de Variáveis

A abordagem tradicional para a construção de modelos estatísticos envolve a escolha de um modelo que seja mais parcimonioso. Parcimônia no dicionário é "ação ou hábito de fazer economia, de poupar", na prática, quando se trata de modelos, o modelo mais parcimonioso é aquele que com a menos quantidade de variáveis irá representar melhor aquele conjunto de dados. Esse procedimento, que será descrito a seguir, é baseado ao proposto por Hosmer e Lemeshow (2000).

A justificativa para isso é de que minimizar o número de variáveis no modelo faz com que o modelo resultante tenha maior probabilidade de ser numericamente estável sendo mais facilmente generalizado. Quanto mais variáveis incluídas em um modelo, maiores se tornam os erros padrão estimados e mais dependente o modelo se torna dos dados observados.

Sendo assim, inicialmente é feita uma análise individual das variáveis, por meio de teste de razão de verossimilhança com $k - 1$ graus de liberdade, com o objetivo de verificar se elas são estatisticamente significantes no modelo.

Feito isso, foram selecionadas as variáveis para a análise multivariada. Para isso, com base no estudo de Bendel e Afifi (1977) e no artigo de Mickey e Greenland (1989), vamos considerar que qualquer variável que em sua análise preliminar obteve um p-valor menor que 0,25, como candidata ao modelo final. Nos trabalhos desses autores, foi apontado que o uso de um nível mais tradicional (como 0,05) muitas vezes não identifica variáveis conhecidas como importantes. O uso do nível superior tem a desvantagem de incluir variáveis de importância questionável na fase de construção do modelo. Por esta razão, é importante rever todas as variáveis adicionadas a um modelo criticamente antes que uma decisão seja tomada em relação ao modelo final.

Diante dos resultados é possível ajustar o modelo de regressão logística multinomial, verificando a importância de cada variável que foi incluída no modelo, por meio do teste de Wald, citado na revisão teórica. As variáveis que não contribuem para o modelo baseado nestes critérios devem ser eliminadas e um novo modelo deve ser ajustado. O novo modelo deve ser comparado ao antigo, por meio do teste da razão de verossimilhança.

Além disso, os coeficientes estimados para as demais variáveis devem ser comparados com os do modelo completo.

Por fim, se refina o modelo de efeitos principais, verificando se as variáveis contínuas estão dimensionadas corretamente, e verifica-se as interações entre as variáveis do modelo.

Antes de usarmos qualquer modelo para inferências, devemos avaliar sua adequação e verificar seu ajuste.

Existem métodos de seleção para determinar quais as variáveis entram em um modelo. O **Método Forward** ocorre quando iniciamos com um modelo contendo apenas o intercepto, em seguida são adicionadas as variáveis, a cada variável adicionada é feito o teste estatístico para verificar se o coeficiente dessa variável adicionada é significante ou não. Caso positivo, essa variável entra no modelo, caso negativo a variável não entra no modelo. Dessa forma esse método percorre todas as variáveis e apenas adiciona ao modelo. O **Método Backward** faz o procedimento o oposto ao método Forward. Aqui o softeware inicia com um modelo inicial contendo o intercepto e todas as variáveis. Assim as variáveis vão sendo testadas uma a uma para verificar quais devem sair do modelo. Existe também o **Método Stepwise**, que é a combinação dos dois métodos acima, onde os coeficientes são testados na hora de entrar e sair do modelo.

4 Manipulação dos dados

A partir dos dados fornecidos pela DDS, a primeira planilha fornecida contém informação dos 3118 que são os alunos de 2016 até 2021 que se cadastraram no SIGAA (sistema institucional vigente). A segunda planinha fornecida contém os alunos que não se cadastraram nesse sistema acrescentando os antigos, totalizando dados de 5585 alunos. Todas essas informações foram fornecidas sem os dados pessoais dos alunos envolvidos, respeitando as diretrizes da Lei Geral de Proteção de Dados (LGPD).

O banco final a ser utilizado possui alguns dados faltantes, por falta desses registros no sistema, a quantidade de valores faltantes por variável está na Tabela 2.

Tabela 2: Variáveis do Banco

Variável	**Descrição**	**Qtd. NA's**
Identificador	Id de 1 a 5585 dos alunos	0
Sem.ano	Semestre e ano de ingresso no Programa	0
2016	Indica se o aluno estava ativo no ano	0
2017	Indica se o aluno estava ativo no ano	0
2018	Indica se o aluno estava ativo no ano	0
2019	Indica se o aluno estava ativo no ano	0
2020	Indica se o aluno estava ativo no ano	0
2021	Indica se o aluno estava ativo no ano	0
Sexo	Sexo declarado pelos alunos	0
Data de Nascimento	Data de Nascimento	1146
Raça	Raça declarada pelos alunos	1984
Curso	Curso de graduação dos alunos	43
Campus	Campus da UnB que o aluno estuda	47
Semestre de Ingresso	Semestre de Ingresso na UnB	720
Tipo de Ingresso	Forma de Ingresso na UnB	2219
Cota de Ingresso	Qual a cota utilizada ao ingressar na UnB	2649
Ensino médio	Cursou ensino médio público ou privado	774
Trancamentos	Qtd. de disciplinas trancadas na graduação	2225
Ativo	Indica se o aluno ainda está ativo na UnB	1167
Semestre de Saída	Sem. de saída do auxílio socieconômico	701
Motivo de Saída	Motivo de saída do auxílio socieconômico	1756

O primeiro passo para a organização desses dados foi encontrar erros, visto que de 2016 para 2021 a UnB utilizou três sistemas diferentes para o cadastro desse alunos que utilizam o benefício socieconômico.

Apesar da variável Raça não possuir nenhum valor faltante, exitem 1984 alunos que não quiseram declarar sua raça ou cor, e ara este caso essas informações foram dadas como NAs.

Alunos que realizaram o vestibular novamente ou realizaram transferência interna, por exemplo, trocam de matrícula e o sistema mais recente utilizado pela universidade identifica essas matrículas diferentes como sendo alunos diferentes. Para regularizar essa situação onde cada aluno pode apresentar mais de uma linha no banco de dados com suas informações, que muitas vezes eram informações distintas entre si, foram estabelecidos critérios para a padronização e organização dessas informações.

Esses critério foram feitos com o auxílio de servidores da Diretoria de Desenvolvimento Social da UnB que trabalham há anos com esses alunos e com base no site oficial da UnB.

Os critérios estão descritos abaixo:

- Caso o aluno tenha mais de uma data de início na utilização do benefício em virtude de novo vestibular realizado ou mudança de curso, foi considerada como a data de início no benefício a primeira já realizada pelo aluno, enquanto a data final que utilizou o auxílio será considerada a última.
- Caso o aluno tenha realizado a inscrição para receber o auxílio e preenchido o campo "Raça"de forma diferente, o último cadastro foi levado em consideração.
- Em alunos que mudaram de curso após ingressarem a UnB, foi registrada a primeira forma como o aluno ingressou na universidade.
- No caso de alunos que mudaram de campus, foi registrado o último campus do registro.
- Em alunos que trocaram de curso e possuem trancamento em disciplinas dos cursos anteriores, a quantidade total de trancamentos foi a soma de todos os cursos que a pessoa passou.
- No caso dos alunos que mudaram de curso, no registro da sua primeira matrícula o motivo de saída é dado como "Mudança de Curso"ou "Novo vestibular", nesses casos foi registrado como o real motivo de sua saída o último registo no sistema, referente ao último curso.

A DDS não exige que os alunos comprovem a renda semestralmente. Até 2015 a validade das avaliações socioeconômicas era de 4 semestres. Em 2016, a validade foi para 10 semestres, ou seja, o estudante só precisava refazer a avaliação depois de 5 anos, dessa

forma, grande parte dos estudantes nem precisaria renovar antes de formar e se mantêm no programa.

A primeira manipulação nesse banco foi criar a variável **Idade**, com base nas datas de nascimento dos alunos, mas comparando com a data que início de sua participação no programa. Isso foi feito pois o banco possui alunos de 2016 a 2021, com essa transformação é possível comparar os alunos com a idade que esses possuíam no momento que ingressaram no auxílio. Como não faz parte dos dados a informação sobre a data exata em que os alunos ingressaram no auxílio, foi colocado, a fim de comparação, uma estimativa onde os alunos que ingressaram no primeiro semestre letivo a data de entrada foi 01/03 e os do segundo semestre letivo foi 01/08 dos respectivos anos. Essas datas foram escolhidas por serem próximas aos semestres usuais em que a UnB inicia as aulas.

Em seguida a formatação das variáveis foi organizada, pois em decorrência das constantes mudanças no sistema utilizado pela UnB, haviam discrepâncias.

Para a variável Raça foram utilizadas: preta, parda, indígena, branca e amarela; que seguem o sistema classificatório de "cor ou raça"do IBGE. Assim como é explicado e discutido no artigo de Osório (2003).

Segundo o próprio site na Universidade de Brasília (BRASÍLIA, 2022a) existem 16 formas de ingresso. Como muitas dessas formas possuem uma frequência absoluta muito baixa, essas classes foram agrupadas. Foram mantidas as classes dos alunos que ingressaram por meio do Vestibular tradicional, ENEM/Sisu e PAS (Programa de Avaliação Seriada), por serem os meios que juntos representam mais de 90% das informações disponíveis. As classes Mudança de Curso, transferência obrigatória e facultativa, foram agrupadas como sendo transferência interna. As demais classes, que são: Acordo Cultural PEC-G, Convênio Andifes - Mobilidade Acadêmica Nacional, Convênio Interinstitucional - Internacional, Matrícula Cortesia, Portadores de Diploma de Curso Superior (DCS), Refugiado e Vestibular para o mesmo curso foram classificadas como "Outros".

Dentre os cursos da UnB divididos nos 4 campus, existem dois cursos denominados Comunicação Social - Publicidade e Propaganda e Comunicação Social - Audiovisual, esses dois foram agrupados como "Comunicação Social", visto que os registros foram feitos de formas diferentes, o que poderia comprometer as análises, dificultando diferenciar o real curso dos alunos.

Como são muitos cursos presentes na UnB e a frequência absoluta de alunos em cada uma das mais de 70 classes acabou apresentando muitos valores pequenos, é de interesse agrupar os cursos por área do conhecimento. Esse agrupamento foi feito com base no site prática.org (2021) que especifica quais os cursos ficam em cada uma das 8 áreas existentes.

A variável Motivo de saída, que fala sobre o motivo pelo qual o aluno deixou

o programa, traz informações de extrema importância para o presente trabalho. Dela podemos tirar quais os alunos formaram, que é um dos alvos do presente trabalho. Além disso, foram agrupadas as categorias que possuem uma frequência absoluta muito baixa ao se comparar com o todo. Primeiro, foram classificados os alunos que ainda estão ativos, ou seja, não possuem motivo de saída por não terem saído da universidade e continuarem no programa, aqui também são considerados os alunos que trocaram que curso. Depois, aqueles que foram desligados no programa, seja por desistência, por terem sido jubilados, por não se encaixarem mais nos requisitos do auxílio ou por solicitação espontânea. Por último, o objetivo de fato, que são aqueles alunos que saíram do benefício, pois se formaram.

Por conseguinte foram criadas novas variáveis que contém informações importantes, que são:

- O tempo de ingresso na UnB até o ingresso no benefício.
- Tempo de entrada no benefício até a saída, seja por desistência ou formatura.

Ambas variáveis são contadas em semestres, sempre contando o início do semestre como sendo dia 01 de março, para o primeiro semestre letivo e 01 de agosto para o segundo. E para a formatura ou desistência, como temos apenas a informação sobre em qual semestre a pessoa se formou ou desistiu, vamos considerar dia 01 de julho como final do primeiro semestre, e 01 de dezembro como o final do segundo semestre.

Para alguns alunos existe o registro no banco apenas após trocarem de curso ou realizarem um novo vestibular, nesse caso aparece que eles entraram no auxílio antes de ingressarem na UnB. A única forma de isso ocorrer é quando esses alunos já utilizaram o auxílio com uma matrícula anterior. Como essa informação foi utilizada para construir a variável "O tempo de ingresso na UnB até o ingresso no benefício", existem alunos onde o tempo apresentou um resultado negativo. Para não viesar o trabalho, a informação referente a essa variável, foi substituída por NAs (valores faltantes), com 187 alunos.

Também foi criada a variável "Razão de Tempo", que é calculada com base na divisão da quantidade de semestres que o aluno ficou no auxílio sobre a quantidade de semestres que o aluno ficou na UnB. Assim, o aluno que não ficou na UnB sem o auxílio socieconômico assume valor 1, os demais apresentam valores no intervalo contínuo de 0 a 1.

5 Análise Exploratória

Abaixo foi realizada a análise exploratória dos dados para o melhor entendimento do banco e do perfil dos alunos.

5.1 Dashboard - Data Studio

Como proposto, foi elaborado um dashboard através da plataforma do DataStudio. O mesmo está disponível pelo link.

5.2 Perfil dos Alunos

O banco de dados disponível possui informação de mais de 5 mil alunos que utilizam ou já utilizaram o auxílio socieconômico oferecido pela UnB. Dentre esses, alguns alunos já se formaram e outros não, seja por estarem ativos da universidade, por teres desistido, trancado, ou apenas saído do programa.

Abaixo pode-se observar algumas das características importantes desses alunos.

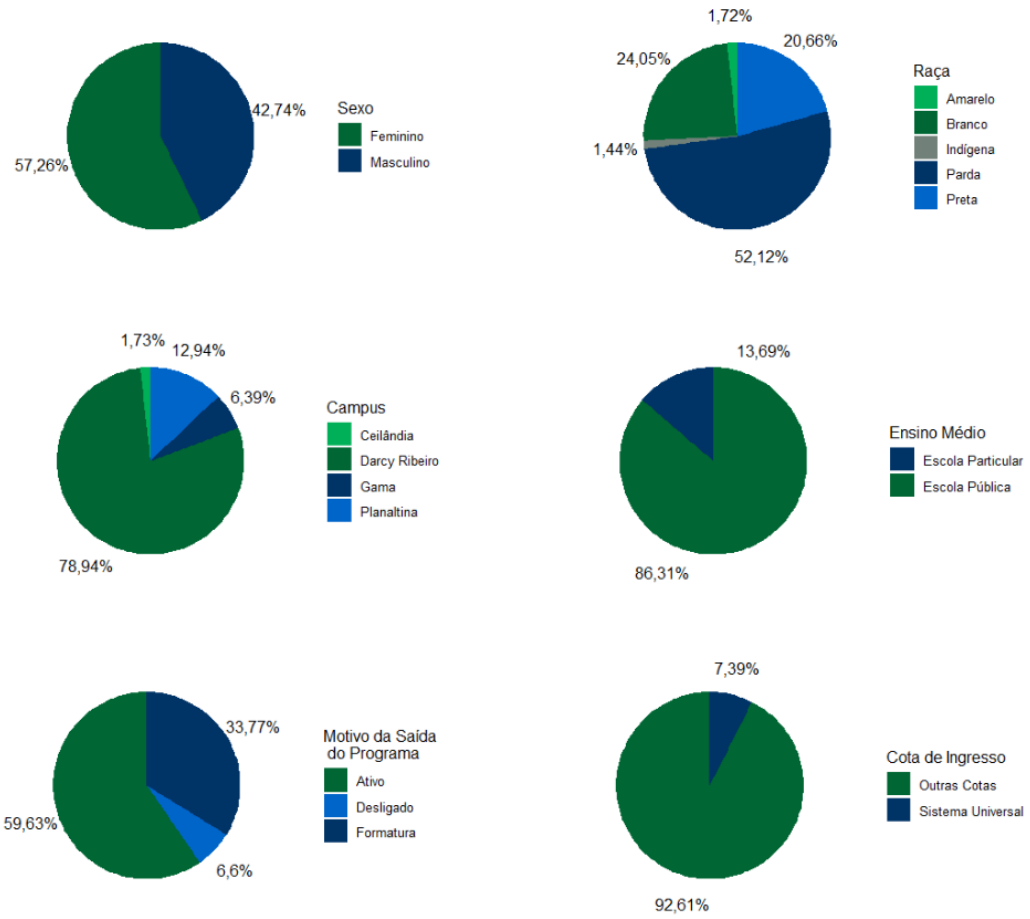

Figura 4: Perfil dos Alunos

Pela Figura 4 é perceptível que trata-se de um cenário de predominância feminina, com aproximadamente 57% dos alunos em questão sendo mulheres. Dentre os alunos em questão existe uma predominância daqueles que se declaram Pardos (33,61%), e duas minorias de Amarelos (1,11%) e Indígenas (0,93%). Brancos e Pretos possuem uma representatividade semelhante no banco de dados, de respectivamente 15,51% e 13,32%.

A Universidade de Brasília é multicampi, e atualmente possui 4 campus diferentes, sendo o Darcy Ribeiro o maior deles.

Com relação ao tipo de escola que esses alunos realizaram o ensino médio, sua maioria foi em escola pública, mas 11,17% foi em colégio particular.

A DDS oferece o auxílio socieconômico com o objetivo de auxiliar os alunos durante a graduação, estes permanecem por um tempo participando do programa. Aqui o banco dados dos alunos que utilizam/utilizaram o auxílio socieconômico possui 59,6% dos alunos que estão ativos na universidade e no programa, outros 33,8% não estão mais ativos no programa pois já se formaram e os outros 6,6% não estão mais ativos, seja porque desistiram, trancaram o curso ou não atingiam mais os pre-requisitos para continuarem

recebendo o benefício.

Dentre esses alunos apenas 7,4% ingressaram na Universidade pelo sistema universal, os demais utilizaram as demais cotas.

5.3 Formação dos alunos

Atualmente na UnB existem 73 cursos de graduação, sendo que 61 deles são no Campus Darcy Ribeiro. (BRASÍLIA, 2022b). Os cursos que lideram na quantidade de alunos que receberam/recebem o benefício estão na tabela abaixo:

Tabela 3: Frequência de alunos nos 5 cursos com a maior quantidade de alunos da amostra

Curso	**Freq. Absoluta**	**Freq. Relativa**
Letras	451	8,08%
Ciências Naturais	249	4,46%
Serviço Social	206	3,69%
Gestão de Agronegócios	200	3,58%
Agronomia	188	3,37%

Curioso destacar que um curso conhecido por ser elitizado, a medicina, está em 4° lugar entre os cursos que possuem menos alunos no banco de dados em questão, com 23 alunos.

Devido à grande diversidade de cursos na UnB é interessante observar como esses dados se distribuem segundo as 8 áreas do conhecimento entre os sexos.

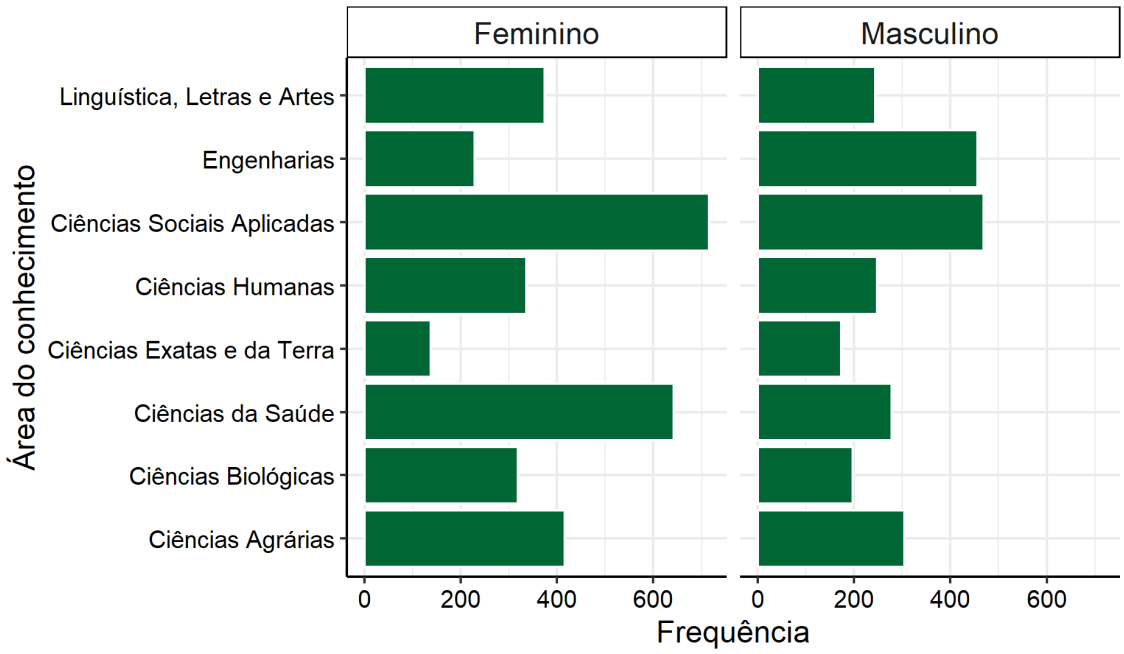

Figura 5: Gráfico de Barras da Área de Conhecimento versus Sexo dos alunos

De modo geral, como já foi visto anteriormente, as mulheres estão em maioria dentre os alunos do banco, então já era esperado que estivessem em maioria em grande parte das áreas do conhecimento. Apesar disso, os cursos de Engenharias possuem maior quantidade de homens matriculados.

Todo semestre ingressam novos alunos da universidade, a princípio existem as vagas reservadas para o ENEM, PAS e pelo Vestibular. Existem outras formas de ingresso, mas essas não possuem quantidade de vagas pré-definias. Vemos a distribuição dessas vagas por oferta e pela quantidade de alunos do banco, para observarmos se existem diferenças entre as vagas oferecidas e os alunos do auxílio.

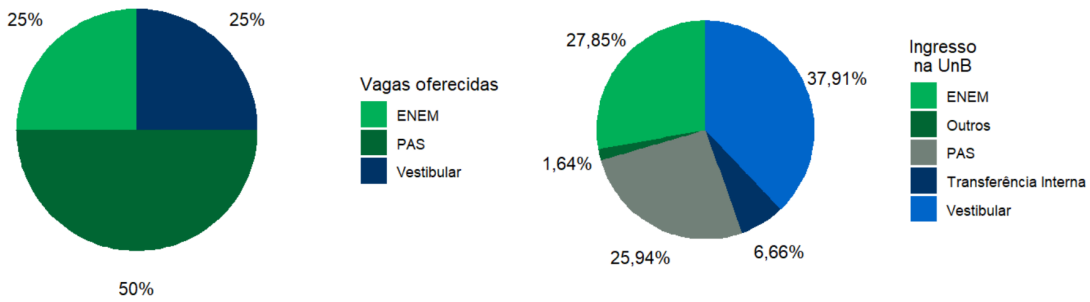

Figura 6: Gráficos de Setores com o comparativo entre as vagas oferecidas pela UnB e tipo de ingresso dos alunos do auxílio socieconômico.

A grande parte dos alunos do auxílio ingressaram pelo Vestibular tradicional, mas

apesar da UnB reservar apenas 25% das vagas para o Vestibular, nem sempre foi assim. Antes as vagas de Vestibular correspondiam à 50% do total, esse histórico é refletido nos alunos atuais.

Abaixo é possível observar a idade dos alunos e a quantidade de disciplinas que os alunos trancaram ao longo da graduação.

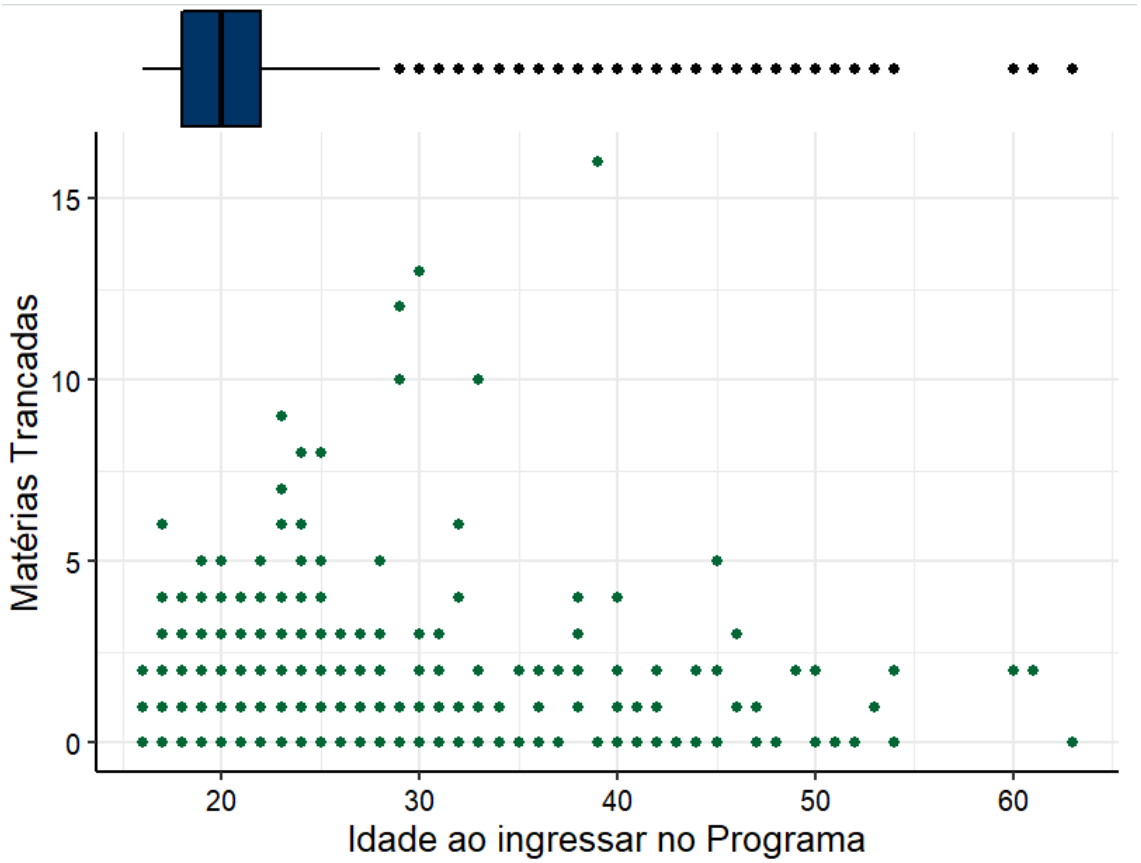

Figura 7: Gráfico de Dispersão entre a quantidade de disciplinas trancadas e a idade dos alunos ao ingressar no programa

A mediana para a idade mostrou que 50% dos alunos ingressaram com 20 anos ou menos no auxílio. Como o valor mínimo observado é de 16 anos, metade dos alunos estão nesse intervalo de 16 a 12 e os outros 50% estão no intervalo de 20 a 63 anos.

Sabe-se que dentre os mais de 5 mil alunos do banco, apenas 604 deles trancaram alguma matéria.

Além disso, foi realizado o teste de Correlação Linear de Pearson, baseado nas hipóteses:

$$\begin{cases} H_0 : \text{Correlação linear nula} \\ H_1 : \text{Correlação linear não nula} \end{cases}$$

Tabela 4: Teste de Correlação Linear entre a idade dos alunos que trancaram alguma disciplina e a quantidade de disciplinas trancadas

Resultados do teste	
Estatística do Teste	4,29
Correlação	0,17
P-valor	$< 0{,}001$
Graus de Liberdade	602

A medida do coeficiente de correlação linear indicou que as duas variáveis apresentam correlação linear fraca positiva. O teste em concordância com o coeficiente mostrou que para o nível de significância de 5% existem evidências estatísticas suficientes para rejeitar a hipótese nula de que não existe correlação.

5.4 Ingresso e Permanência

Um ponto de suma importância para ser analisado com relação aos alunos beneficiados pelo auxílio socieconômico é como esses alunos ingressaram a universidade e como está sendo o período de permanência dos mesmos, sempre visando a formatura.

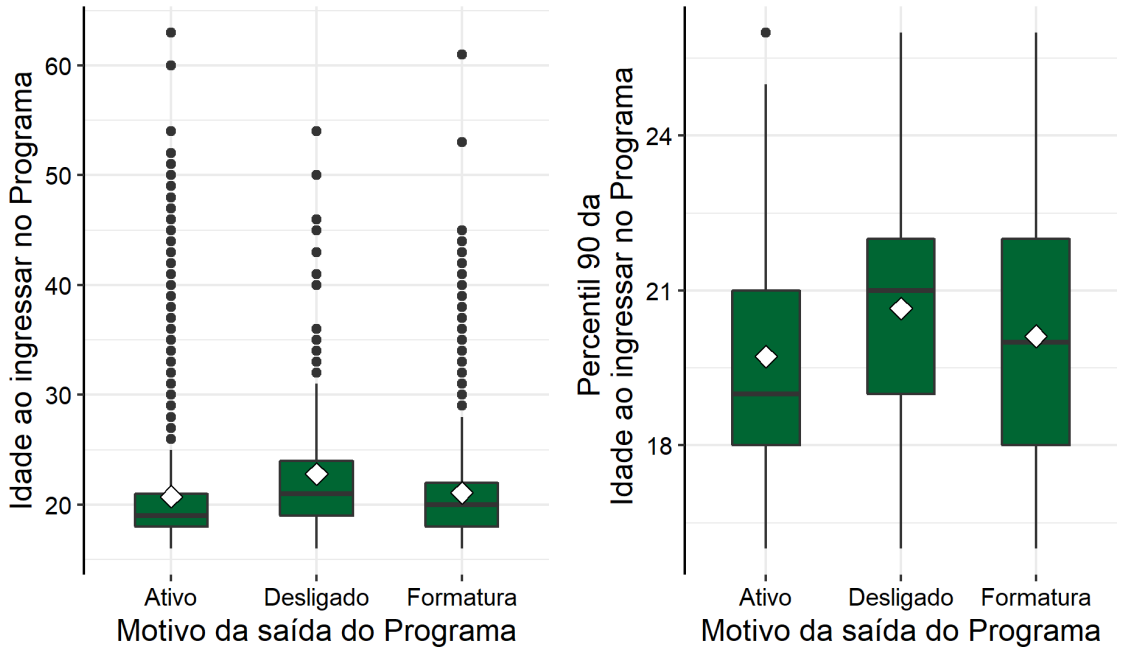

Figura 8: Boxplot para as idade dos alunos segundo o motivo de saída do programa

Observando a idade que os alunos tinham ao ingressarem no programa segundo

ao motivo pelo qual deixaram o mesmo. O gráfico da direita é filtrado pelo percentil 90 da idade dos alunos, filtrando apenas os alunos de 26 anos ou menos, com o intuito de facilitar a visualização.

A média e a mediana da idade de ingresso na UnB dos alunos que formaram é respectivamente de 20 e 21 anos, enquanto que a menor idade de ingresso na UnB de aluno que já se formou é de 16 anos e a maior idade é de 61 anos.

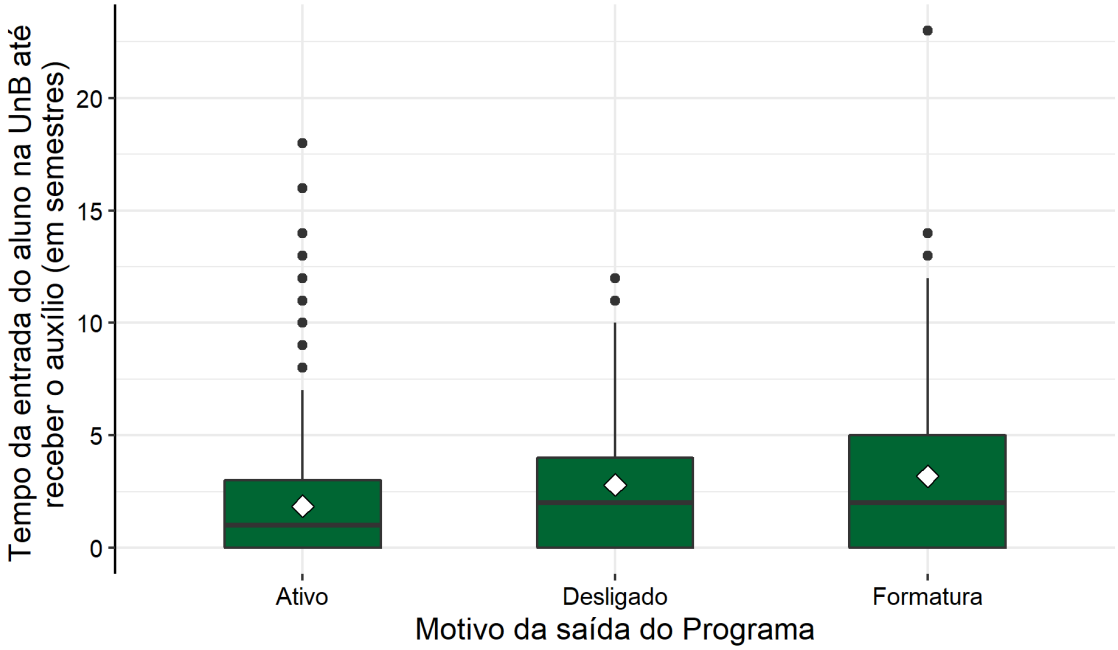

Figura 9: Tempo que os alunos permaneceram na UnB até começarem a receber o auxílio segundo o motivo pelo qual saíram do mesmo

Acima vemos em semestres quanto tempo os alunos permaneceram na UnB sem receber o auxílio. É importante ressaltar que podem existir alguns casos onde o aluno trancou o curso nesse período, trocou de curso, fez dupla graduação ou realizou um novo vestibular, sendo assim nem todos os alunos seguiram um fluxo normal de entrar na universidade, cursar as matérias e se formar. Isso explica a grande quantidade de outliers.

A média da quantidade de semestres dentre os alunos ativos é de 1,85, dos desligados é de 2,64, e dos que se formaram é de 3,8 semestres. A mediana dentre os formados é de 3 semestres e entre os deligados é de 2 semestres, ao contrário do que era de se esperar.

Também é interessante observarmos quantos semestres os alunos ficaram no auxílio segundo o motivo de saída do programa, se ele foi desligado ou se formou, em conjunto com uma divisão entre os alunos que trancaram alguma disciplina na graduação, ou não.

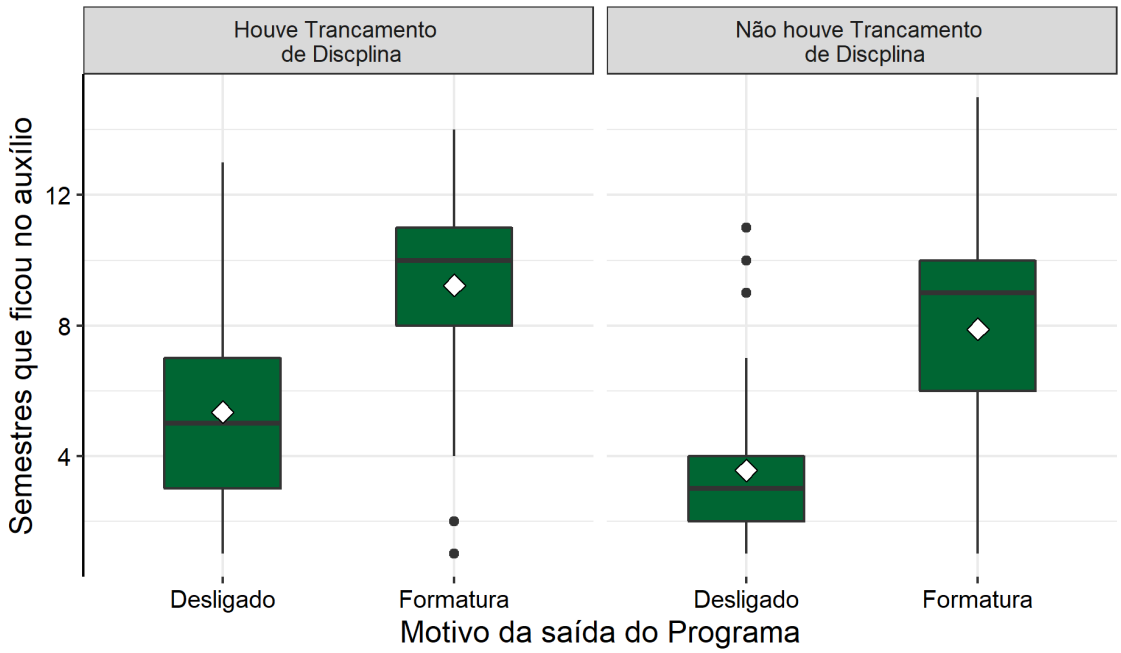

Figura 10: Tempo que os alunos permaneceram utilizando o auxílio segundo o motivo se saída e se estes trancaram alguma disciplina na graduação

Observa-se um comportamento semelhante entre os alunos que trancaram alguma matéria e os demais no sentido de que os alunos que se formaram permaneceram mais tempo no auxílio do que os se desligaram do auxílio. A grande parte das observações referentes aos alunos que passaram mais semestres no auxílio são de alunos que se formaram, onde o trancamento de disciplinas não pareceu influenciar significativamente.

Os alunos que trancaram alguma matéria e se desligaram no programa ficaram em média 5,34 semestres no mesmo, enquanto os que não trancaram e também se desligaram ficaram no auxílio por 3,56 semestres. Importante levar em consideração que as bolsas são de R$ 465,00 mensais, totalizando R$ 2.790,00 por semestre.

Aproveitando a variável Trancamentos por semestre na UnB, pode-se observar a sua distribuição conjuntamente com a idade dos alunos ao ingressarem no programa segundo o motivo de saída do mesmo.

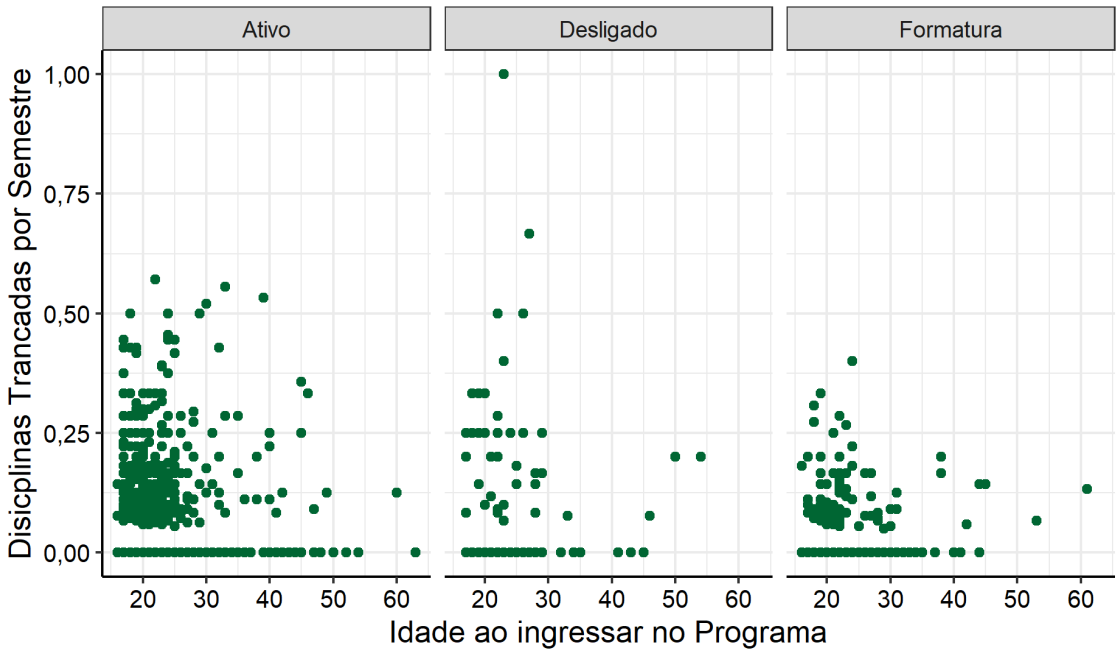

Figura 11: Gráficos de Dispersão entre a idade dos alunos ao ingressar no auxílio socieconômico e a quantidade de disciplinas trancas por semestre durante a graduação, segundo a variável Motivo de Saída

O trancamento é um recuro que a universidade disponibiliza para ajudar os alunos. A ideia é saber qual a melhor forma de organizar as matérias e obter êxito ao final de cada semestre.

Um exemplo que chama a atenção é o caso de um aluno que trancou uma matéria por semestre que cursou, dentre os demais existe uma concentração maior de 0 a 0,25 disciplinas trancadas por semestre.

Os alunos que ingressaram mais velhos no programa parecem trancar menos matérias, do que os mais jovens. Já os que se formaram parecem ter trancados menos matérias do que os que se desligaram.

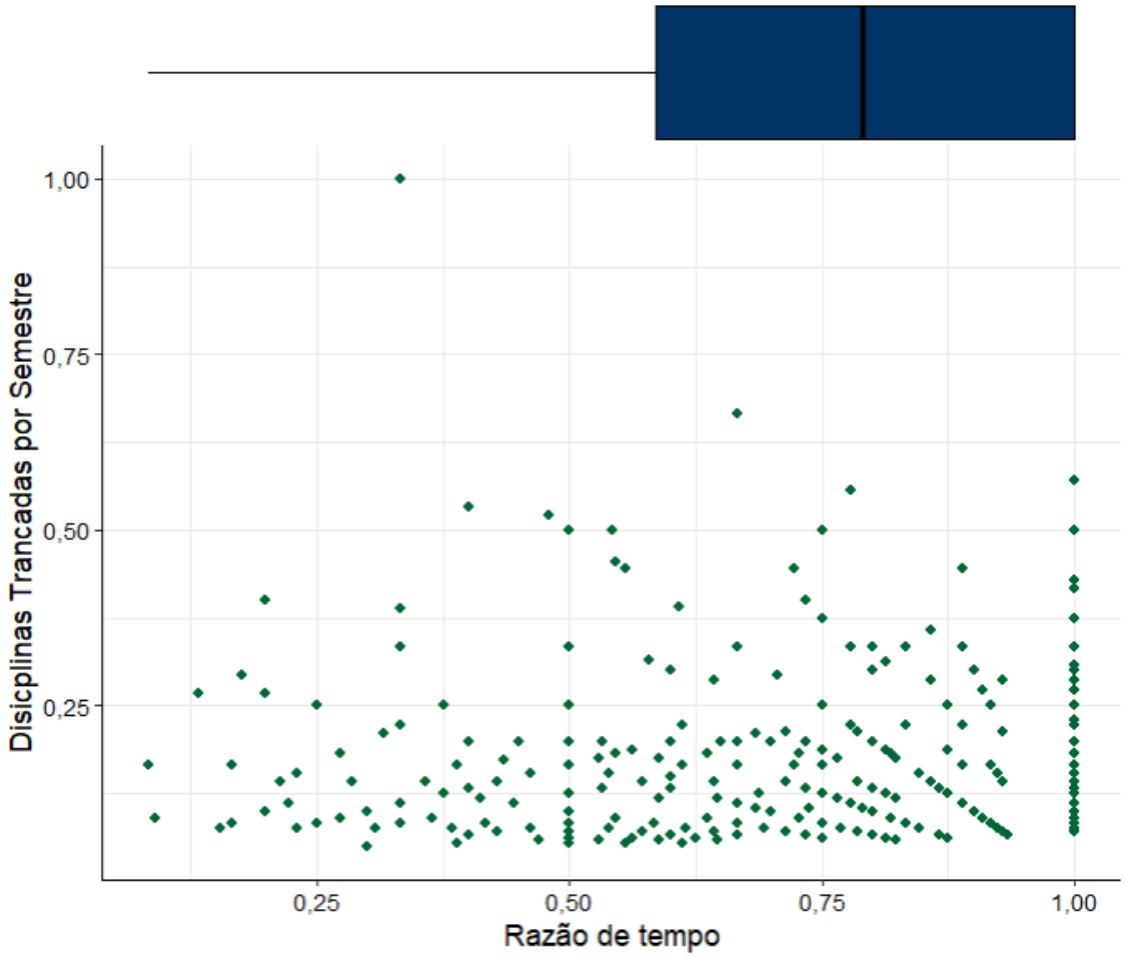

Figura 12: Gráfico de Dispersão entre as variáveis Razão de Tempo e Trancamentos por semestre dentre os alunos que trancaram pelo menos uma matéria

Como foi descrito na metodologia foi criada a variável razão de tempo, que relaciona a quantidade de semestres no auxílio dividido pela quantidade de semestres na UnB, para que seja possível trabalhar com a porcentagem de semestres que eles permaneceram com o auxílio.

A mediana da razão de tempo igual a 0,78 indicando que metade dos alunos ficaram mais de 78% do tempo que estiveram na UnB recebendo o auxílio socieconômico.

Importante ressaltar o ponto que não foi disponibilizado o motivo pelo qual os alunos não ingressaram antes no auxílio, mas hipóteses foram levantadas sobre o aluno não ter ingressado antes, por não estar em vulnerabilidade social previamente. Visto que a UnB mantém esforços com o objetivo de atender a maior quantidade possível de alunos que necessitam do auxílio, em muitos semestres foi possível atender a todos os que solicitaram.

Deseja-se observar se existe correlação linear entre a razão de tempo e a quantidade de trancamentos por semestre dentre os alunos que trancaram pelo menos uma matéria com base no teste de correlação linear de Pearson, que é feito segundo as hipóteses

abaixo:

$$\begin{cases} H_0 : \text{Correlação linear nula} \\ H_1 : \text{Correlação linear não nula} \end{cases}$$

Tabela 5: Teste de Correlação Linear entre as variáveis Razão de Tempo e Trancamentos por semestre

Resultados do teste	
Estatística do Teste	0,6
Correlação	0,026
P-valor	0,5
Graus de Liberdade	525

O p-valor do teste não trouxe evidências estatísticas suficientes para rejeitar a hipótese nula, assim, assume-se que não existe correlação linear entre as duas variáveis.

Outra visualização interessante para o estudo é a relação entre a variável razão de tempo segundo os motivos de saída do aluno do programa.

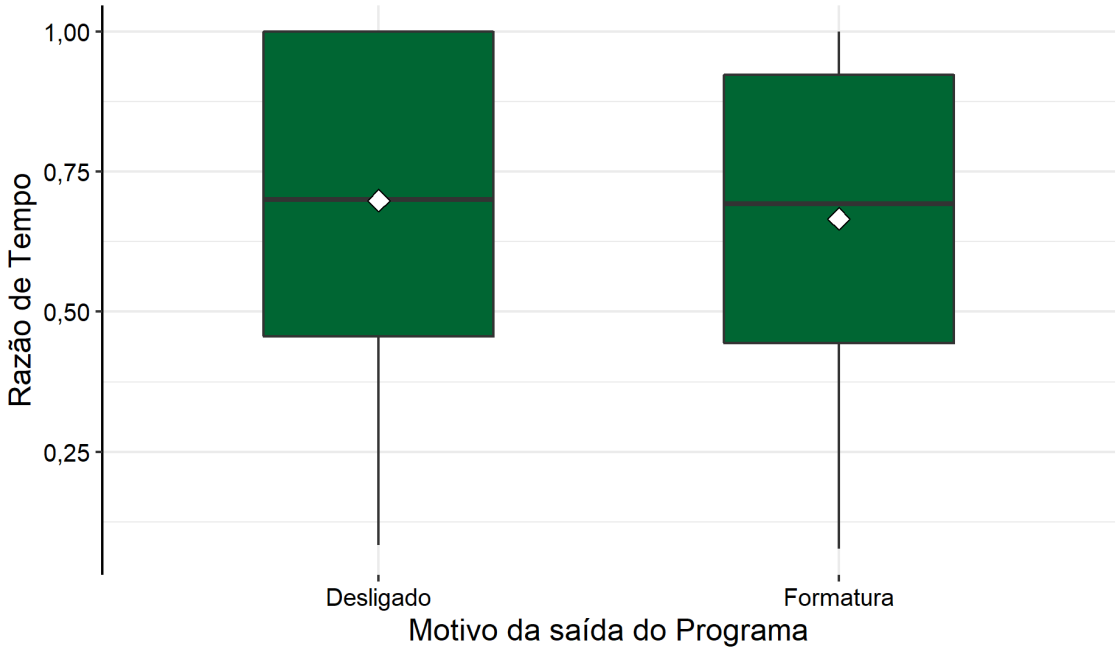

Figura 13: Boxplot entre a variável Razão de Tempo e Motivo de Saída dentre os alunos que saíram da UnB

Existe uma leve discrepância entre os alunos que se desligaram e os que se formaram, mas a mediana das duas distribuições é muito semelhante.

O terceiro quartil dos alunos que se desligaram está em 1, ou seja, 25% dos alunos que se desligaram ficaram recebendo o auxílio durante todo o período que estiveram na UnB. Já entre os formados, foi observado um resultado diferente, o terceiro quartil é de 0.92.

Esse resultado pode abrir portas para uma interpretação incorreta sobre o assunto. Parece não fazer sentido os alunos que se desligaram terem uma razão de tempo maior, mas tem-se que se eles ficaram esse tempo no auxílio, é uma comprovação de seu estado de vulnerabilidade social durante seu período na UnB. Já com os demais alunos, não tem-se a informação do motivo pelo qual não permaneceram todo o tempo no auxílio, mas pode abrir portas para a hipótese de que esses alunos não necessitavam desse auxílio antes, ou possuíam uma rede de apoio maior.

Com relação aos trancamentos por semestre e o tipo de escola que os alunos curaram seu ensino médio temos duas visualizações do mesmo gráfico abaixo:

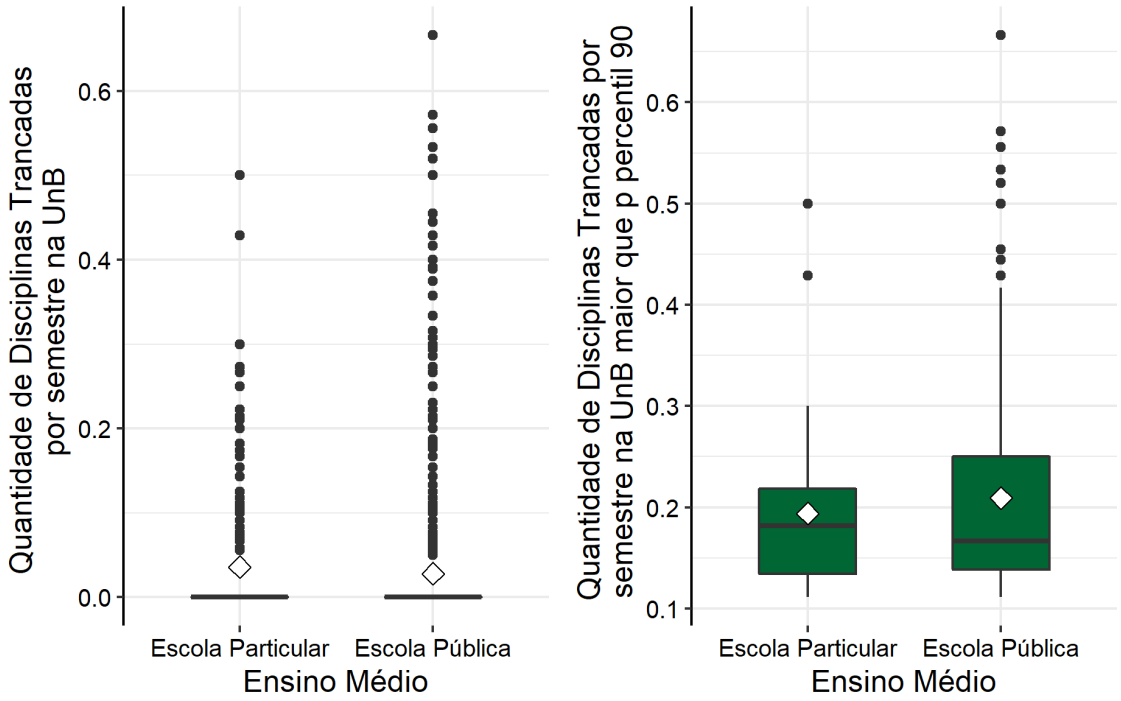

Figura 14: Boxplots entre a quantidade de disciplinas trancadas por semestre e o tipo de escola que o aluno cursou o ensino médio

Foi calculado o percentil 90 da distribuição da variável trancamentos por semestre e foi filtrado pelos maiores valores da distribuição, já que como é possível observar na figura da esquerda mais da metade dos alunos não trancaram nenhuma matéria, o que acaba dificultado a visualização dessa distribuição.

Para a escola particular a distribuição aparenta ser mais simétrica do que os alunos da escola pública e apresenta uma menor quantidade de valores extremos superiores.

Dentre essa classificação do ensino médio dos alunos, deseja-se observar qual a proporção deles que já se formou ou não formou por terem desistido do curso, pedido desligamento do auxílio ou ainda está ativo.

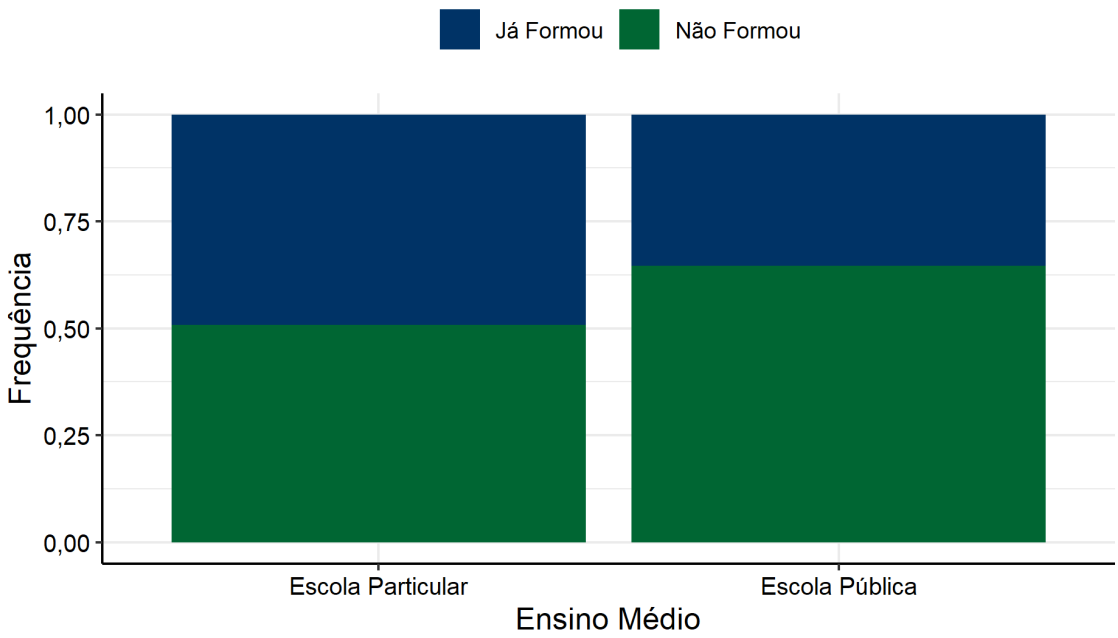

Figura 15: Gráfico de Barras entre a variável que indica a formatura ou não do aluno e o tipo que escola que o mesmo cursou o ensino médio

É visível uma diferença onde os alunos de escola particular, apensar de estarem em minoria no banco quase metade deles já se formaram (49%) enquanto que entre os de escola pública apenas 35% já se formaram.

Para identificarmos se essa diferença é significante ou não, foi realizado o teste qui-quadrado de independência com base nas hipóteses abaixo.

$$\begin{cases} H_0 : \text{As variáveis são Independentes} \\ H_1 : \text{As variáveis não são Independentes} \end{cases}$$

O resultado do teste é dado por:

Tabela 6: Teste Qui-Quadrado de Independência entre as variáveis Formatura e Ensino médio

Resultados do teste	
Estatística do Teste	26
P-valor	<0,001
Graus de Liberdade	1

O teste mostrou que existem evidências estatísticas suficientes para rejeitar a hipótese nula, assim aceitamos a hipótese alternativa de que as variáveis que indicam a formatura do aluno e o tipo de escola que cursou o ensino médio não são independentes.

De maneira semelhante deseja-se observar como as variáveis indicadoras de formatura e cotas se apresentam.

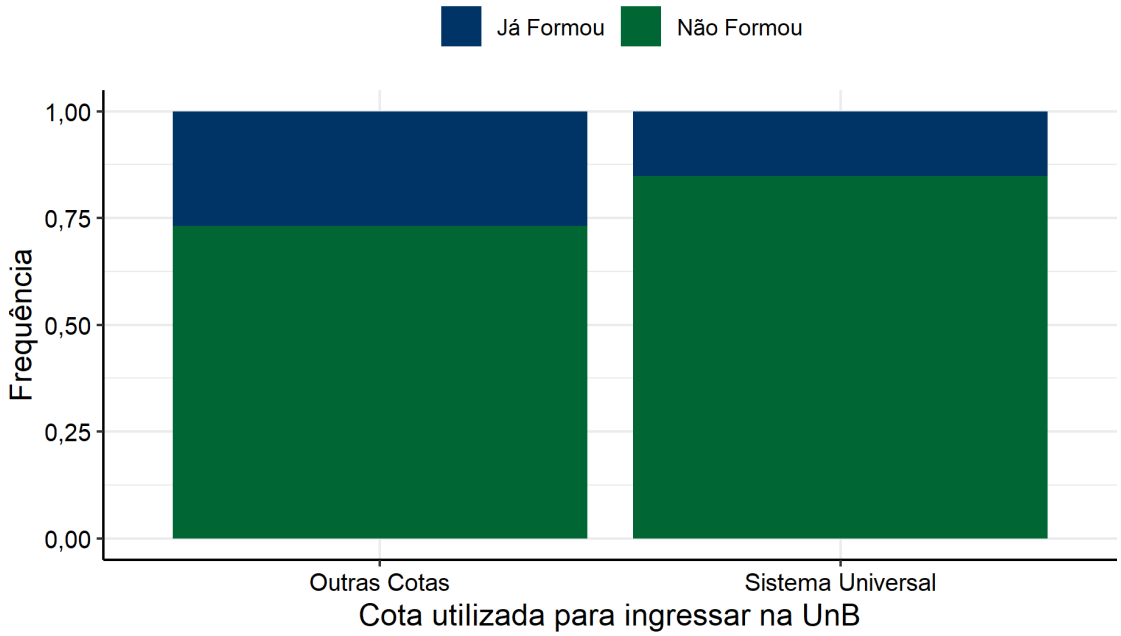

Figura 16: Gráfico de Barras entre a variável que indica a formatura ou não do aluno e se utilizou cotas para ingressar na UnB ou se utilizou o sistema universal

É importante relembrar que apenas 7,39% dos alunos presentes no banco em questão utilizaram o sistema universal. E alunos do sistema universal apresentam uma menor proporção de formaturas do que quando comparado com os alunos das demais cotas. Para verificar se isso é algo pontual ou não será realizado o teste qui-quadrado de independência com base nas hipóteses abaixo:

$$\begin{cases} H_0 : \text{As variáveis são Independentes} \\ H_1 : \text{As variáveis não são Independentes} \end{cases}$$

Tabela 7: Teste Qui-Quadrado de Independência entre as variáveis Usou cota e Formatura

Resultados do teste	
Estatística do Teste	14
P-valor	<0,001
Graus de Liberdade	1

Como o p-valor é menor que o nível de significância, existem evidências estatísticas para rejeitar a hipótese de que essas variáveis sejam independentes. Assim, assume-se que as variáveis que indicam a formatura do aluno e a utilização das cotas não são independentes.

6 Modelo Logístico Binário

O objetivo dessa sessão é modelar a probabilidade de formatura dos alunos em estado de vulnerabilidade social que utilizam o auxílio socieconômico da Universidade de Brasília. Identificando quais os fatores que estão relacionados com essa probabilidade.

6.1 Seleção das Variáveis

A variável resposta do modelo indica a formatura ou não do aluno. Existem 20 opções de variáveis explicativas, nessa sessão será descrito o processo de escolha.

Como existem valores faltantes do banco, os dados foram filtrados excluindo os NA's da variável resposta. Também foram desconsiderados os alunos ativos, pois entende-se que estes não tiveram a oportunidade de se formar ainda, como o objetivo é modelar a probabilidade de formatura, não faria sentido utilizar a informação desses alunos nesse momento.Foram retiradas, também, os alunos que saíram do auxílio de forma espontânea antes de sua formatura, (apenas 12 estavam nessa categoria). A partir dessas mudanças, o modelo de regressão logística binária foi construído com base na informação de 1541 alunos.

Inicialmente foram construídos 20 modelos com apenas uma variável explicativa, a fim de entender o comportamento e a influência das mesmas individualmente.

Dentre esses modelos, aqueles que apresentaram p-valor indicando que a inserção da variável é significativa são modelos das variáveis: Semestre de entrada na Unb, Ano de entrada na Unb, Sexo, Forma de Ingresso, Campus (Darcy e Planaltina), Idade ao ingressar no auxílio, quantidade de trancamentos, trancamentos por semestre, Cota de Escola pública, Cota Universal, área do conhecimento (Ciências Exatas e da Terra e Engenharias), variável que indica se houve algum trancamento de disciplina na graduação, quantidade de semestres antes do auxílio, quantidade de semestres no auxílio, quantidade de semestres na UnB e a variável que indica se o aluno utilizou cotas ou o sistema universal para ingressar na Universidade de Brasília.

Consequentemente os modelos onde o p-valor do teste de Wald para a significância dos coeficientes aceitou a hipótese nula de que $\beta_i = 0$, são os modelos das variáveis: Campus (Gama), Ensino médio, Raça, Cota de baixa renda, cota de indígena, cota de negros, Área do Conhecimento (Ciências Biológicas,Ciências da Saúde, Ciências Humanas, Ciências Sociais Aplicadas e Engenharias Linguística, Letras e Artes) e a Variável Razão de Tempo.

Analisando as tabelas de contingência e entre a variável que indica formatura

ou desistência com cada variável quantitativa, identificamos algumas onde as categorias possuíam poucas observações. Além disso, para algumas variáveis categóricas que possuem mais de duas classes, nem todas apresentaram significância. Então estas foram agrupadas de maneira que fizesse sentido lógico na sua interpretação, e que fossem significantes no modelo. Em virtude disso, as variáveis campus, raça, forma de ingresso, área do conhecimento, tiveram suas categorias reorganizadas.

A variável do Campus ficou agrupada como "Darcy Ribeiro"e os "Demais Campus". A variável Raça ficou definida também em duas categorias: "Branco ou Amarelo"e "Parda, Preta ou Indígena". A variável que fala sobre a forma de ingresso, os alunos da categoria "Transferência Interna"foram agrupados junto à categoria "Outros".

Apenas as categorias de Engenharias e Ciências Exatas e da Terra apresentaram significância para o modelo, mas somente essas duas não representam uma quantidade grande de observações. Então a variável de área do conhecimento foi agrupada entre (1) Ciências Agrárias, Ciências Biológicas, ciências da Saúde , (2) Ciências Exatas e da Terra e Engenharias e (3) Ciências Humanas, Ciências Sociais Aplicadas , Linguística, Letras e Artes. Porém as mesmas duas áreas do conhecimento Engenharias e Ciências Exatas e da Terra continuaram sendo as únicas a apresentarem significância nos modelos. Então a categoria que será utilizada é a "Engenharia, Ciências Exatas e da Terra"e as demais em outra classe.

Após essa organização das categorias, a variável Campus deixou de ser significativa no modelo, e Raça se manteve como não significativa e a variável da forma de ingresso se manteve como significante.

Ao utilizar a seleção automática Stepwise o modelo não convergiu, então foi realizada a seleção manual das variáveis.

Iniciou-se o precedimento com o modelo contendo todas as variáveis que se mostraram significante de maneira isolada no modelo.

Sabe-se que quando existe Multicolinearidade entre as variáveis resposta do modelo, algumas podem não ser significantes. Sendo assim foi realizado o teste de correlação linear de Pearson ao nível de significância de 5% entre as variáveis quantitativas presentes no modelo para identificarmos as correlações.

Cada uma das variáveis que apresentou correlação com pelo menos uma outra variável. Então todas elas foram retiradas e serão inseridas uma a uma caso necessário. Assim as variáveis foram sendo retiradas uma a uma com base no p-valor para a significância do modelo. Além disso para comparar os modelos foi realizado Teste da razão de verossimilhança com base nas hipóteses:

$$\begin{cases} H_0 : \text{Não Existe diferença significativa entre os modelos} \\ H_1 : \text{Existe diferença significativa entre os modelos} \end{cases}$$

Vale ressaltar que, caso o teste da razão de verossimilhança não acuse diferença entre modelos, será utilizado o critério da Parcimônia, escolhendo o modelo mais simples.

Então, o modelo apresentou as variáveis: Semestre de entrada na UnB, área do conhecimento nas duas categorias, ingresso e a variável se houve trancamento. Da seguinte forma:

Feito esses procedimentos foram testadas as variáveis que ainda não haviam entrado no modelo: Razão de tempo, Ensino médio, Raça, as cotas de baixa renda, indígena e negro.

Ao incluir a variável razão de tempo, o p-valor correspondente ao teste de significância do coeficiente não indicou que essa nova variável é significante no modelo. O teste de razão de verossimilhança indicou que a retirada da variável não teve impacto significativo, portanto a variável não entra. O mesmo ocorreu com as variáveis Ensino Médio, Raça, Cota de baixa renda e Cota para Negros.

Para a variável indicativa de Cotas para indígenas, essa não apresentou p-valor significativo no modelo, apesar do teste de razão de máxima verossimilhança indicar que ela possui um impacto. Essa variável não foi incluída no modelo, uma vez que não existe nenhum aluno na base de treino e que tenha se formado e entrado na UnB por cota indígena, o que não trará uma boa estimativa caso entre no modelo.

Em seguida as variáveis contínuas que foram retiradas no começo foram testadas novamente, e observado se ao acrescenta-las, trouxeram efeitos significantes. Que são as variáveis: Trancamentos, trancamentos por semestre, semestres no auxílio, semestres na UnB semestres antes do auxílio, ano de entrada na UnB e idade do aluno ao ingressar no auxílio.

Dentre essas, 2 entraram no modelo: trancamentos por semestre, semestres na UnB, e semestres antes do auxílio. Após esse passo, identificamos que as variáveis Semestre de entrada na Unb, e Houve trancamento poderiam sair do modelo. Feito esses procedimentos a variável razão de tempo passou a ser significante.

Na Tabela 8 observa-se a relação de variáveis que foram significativas para o modelo e as que não foram.

Tabela 8: Relação de Variáveis no modelo logístico binário

Presentes no modelo	**Ausentes no modelo**
	Quantidade de Semestres na UnB
	Quantidade de Semestres no Auxílio
Trancamentos por semestre	Semestre de entrada na UnB
	Ano de entrada na UnB
Área do Conhecimento	Campus
	Sexo
Quantidade de Semestres antes do auxílio	Ensino Médio
	Raça
Razão de Tempo	Qtd de Trancamentos
	Houve Trancamento
Forma de ingresso	Usou Cota
	Cota para Escola Pública
Idade ao ingressar no auxílio	Cota para Negros
	Cota para Indígenas
	Cota Universal
	Cota Baixa Renda

Curioso observar que nenhuma das variáveis de cota se mostrou significativa, assim como o momento de ingresso do aluno da Universidade.

O modelo final é dado por:

Tabela 9: Modelo de Regressão Logística Binária

Coeficiente	**Estimativa β_i**	**Erro Padrão**	**Z**	**P-valor**
Intercepto β_0	-4,08	1,38	-2,97	0,003
Trancamentos por semestre	-9,78	2,00	-4,89	<0,001
Área do conhecimento - (1)	-1,55	0,32	-4,85	<0,001
Semestres antes do auxílio	0,90	0,14	6,30	<0,001
Ingresso - Outros	0,52	0,65	0,81	0,42
Ingresso - PAS	0,36	0,39	0,93	0,35
Ingresso - Vestibular	1,27	0,35	3,62	<0,001
Idade ao ingressar no auxílio	-0,13	0,03	-4,32	<0,001
Razão de tempo	- 8,91	1,38	6,48	<0,001

Todas as variáveis do modelo apresentaram pelo menos um coeficiente significativo. Os resultados foram muito interessantes do ponto de vista do auxílio socieconômico

e da probabilidade de formatura. Fato esse que já era de se esperar, uma vez que a quantidade de trancamentos por semestre impacta negativamente na probabilidade de formatura do aluno.

Outro resultado interessante é que o modelo trouxe a informação de que os alunos de cursos de Ciências Exatas e da Terra e os alunos de Engenharias possuem menor probabilidade de formatura, quando comparado com os cursos das demais áreas do conhecimento.

6.2 Qualidade do ajuste e diagnóstico

Foi realizado o teste de Hosmer-Lemeshow ao nível de significância de 5%, para ajudar a identificar se o modelo é adequado, com base nas hipóteses:

$$\begin{cases} H_0 : \text{O modelo está bem ajustado aos dados} \\ H_1 : \text{O modelo não está bem ajustado aos dados} \end{cases}$$

E o resultado se encontra na tabela a seguir.

Tabela 10: Teste de Hosmer-Lemenshow

Resultados do teste	
Estatística do Teste	12
P-valor	0,2
Graus de Liberdade	8

O p-valor indica que não existem evidências estatísticas suficientes para rejeitar a hipótese nula de que o modelo está bem adequado aos dados.

Tabela 11: Medias de Diagnóstico do Modelo de Regressão Logística Binária

Medida	**Valor**
AIC	356
BIC	397
Deviance	338

Para avaliar a qualidade do ajuste foi calculado o critério de informação de Akaike, o critério de informação bayesiano e o Deviance. Essas medidas também foram utilizadas para encontrar o modelo selecionado, visto que valores menores são preferíveis.

Sabe-se que modelos onde existe multicolinearidade, os erros padrão dos coeficientes são mais elevados.

Tabela 12: Fator de influência da Variância Generalizado (GVIF) por variável do modelo

	GVIF	Graus de Liberdade
Trancamentos por semestre	1,05	1,00
Área do Conhecimento	1,07	1,00
Semestres antes do auxílio	7,50	1,00
Ingresso	1,16	3,00
Idade no ingresso	1,27	1,00
Razão de Tempo	7,33	1,00

Como o VIF não é aplicável a modelos que contém preditores categóricos com mais de 2 categorias, que é o caso em questão devido a presença da variável de ingresso. Tem-se então o Fator de Inflação da Variância generalizado (gVIF), onde apesar de um valor mais elevado entre duas das variáveis, nenhum passa de 10 e devido a importância dessas variáveis, elas são mantidas no modelo.

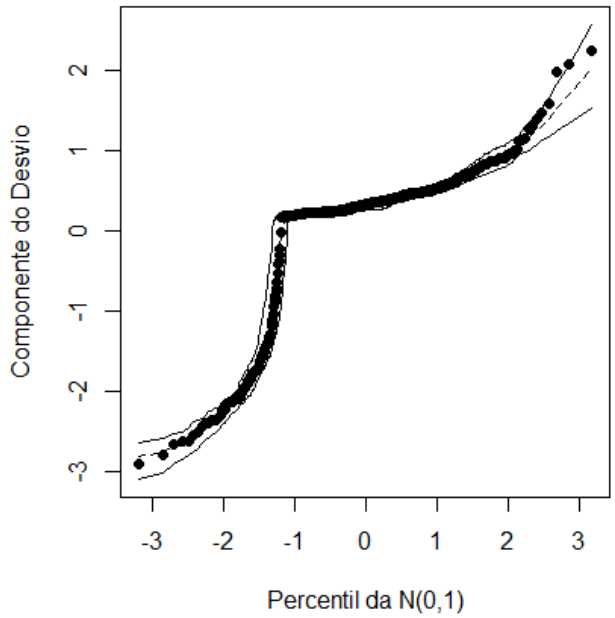

Figura 17: Gráfico de envelope dos resíduos Deviance do modelo de regressão logística binária

Pela figura 17 os resíduos aparentam um comportamento aleatório em torno do zero e se concentram dentro da banda de confiança indicando um bom ajuste dos dados.

6.3 Poder preditivo do modelo

Com relação ao poder preditivo, abaixo tem-se a curva ROC.

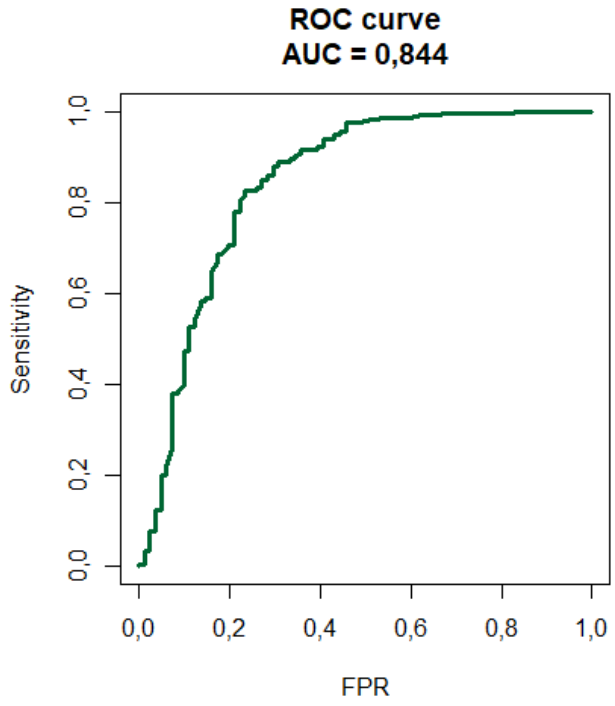

Figura 18: Curva ROC

A área abaixo da curva ROC indica que o modelo está acertando mais de 84% das previsões, com base nos dados de treino e validação do modelo.

A base de validação possui 441 observações, mas retirando os NA's tem-se 281, com base nesse banco foi realizada a matriz de confusão com base no ponto de corte escolhido de 0,85, que obteve os melhores resultados.

Tabela 13: Matriz de confusão do poder preditivo do modelo logístico bonimial

		Amostra	
		Desistência	Formatura
Previsto	Desistência	21	31
	Formatura	14	215

Por meio desta, tem-se a tabela com as medidas:

Tabela 14: Medidas sobre o poder preditivo

Medida	**Valor**
Acurácia	83,99%
Sensibilidade	87,4%
Especificidade	60%

O modelo classificou corretamente quase 84% dos dados da amostra (acurácia). A sensibilidade de 87% é um indicador sobre o quanto o modelo está classificando como Formatura dos alunos que de fato se formaram. Já a especificidade de 60% é um indicador do quanto o modelo está classificando como Desistência os alunos que desistiram realmente do curso.

6.4 Interpretação dos Parâmetros

6.4.1 Variáveis

O coeficiente da variável **Trancamentos por semestre** indica que quanto maior a quantidade de trancamentos o aluno tem por semestre ao longo da sua graduação, menor será a sua probabilidade de formatura.

Já com relação á **Área de Conhecimento** do curso, o modelo trouxe uma conclusão que já era de se esperar. Os alunos dos cursos de Engenharia e Ciências Exatas e da Terra possuem menor probabilidade de formatura que os demais.

A **Quantidade de semestres antes do auxílio**, que fala sobre quantos semestres o aluno ficou na UnB sem receber o auxílio, trouxe um resultado surpreendente. Um aumento na probabilidade de formatura dos alunos que ficaram na UnB mais tempo sem o auxílio, antes de receberem o mesmo. Isso pode estar ocorrendo devido a diversos fatores que não são o objeto de estudo, que por isso não podem ser inferidos. Mas podem levantar algumas hipóteses, de que talvez fosse o caso de alunos que antes não se enquadravam nos requisitos de vulnerabilidade social antes de receberem o auxílio ou de alunos que precisaram lidar com a situação de vulnerabilidade de outras formas enquanto não tinham o auxílio e isso os ajudou.

Com relação à forma do **Ingresso** dos alunos na universidade, somente o vestibular apresentou um p-valor indicando significância com relação ao ENEM. Dessa forma, os alunos que ingressaram pelo Vestibular possuem um impacto positivo na probabilidade de formatura em relação aos que ingressaram pelo ENEM.

A **Idade do aluno ao ingressar no auxílio socieconômico**, ao contrário do que era de se esperar, possui um coeficiente negativo, indicando que quanto maior a idade que o aluno possui no momento que começa a receber o auxílio, isso traz um impacto negativo em sua probabilidade de formatura.

A **Razão de Tempo**, que é calculada através da divisão entre a quantidade de semestres que o aluno ficou no auxílio e a quantidade de semestres que o aluno ficou na UnB, basicamente fala sobre a porcentagem de semestres que o aluno ficou no auxílio. O resultado do coeficiente dessa variável foi muito curioso, ao contrário do que era de se esperar, o aluno que possui uma porcentagem maior da sua graduação recebendo o auxílio socieconômico, possui uma menor probabilidade de formatura. Novamente nos deparamos com algo que pode levantar algumas hipóteses sobre o motivo pelo qual isto ocorre, talvez o objetivo do auxílio seja algo realmente temporário para que o aluno consiga sair dessa situação social e progredir financeiramente ainda durante a graduação. Vale ressaltar que pelo estudo em questão isso não pode ser inferido, apenas uma hipótese levantada.

A entrada das variáveis **Semestres antes do auxílio** e **Razão de tempo** abrem portas para interpretações equivocadas, que devem ser esclarecidas. Não é possível inferir que o tempo maior no auxílio é o responsável por impactar negativamente na probabilidade de formatura como se fosse uma influência negativa do auxílio. O que de fato dificulta a formatura é o estado de vulnerabilidade socieconômica. Seria de muita relevância para o estudo caso fosse disponibilizado pela UnB, informações sobre o motivo pelo qual os alunos não ingressaram antes no auxílio. Levanta-se a hipótese de que no começo do curso, esses alunos que demoraram mais a ingressar no auxílio, possuíam mais recursos familiares, e esse apoio inicial fez diferença para seu êxito.

6.4.2 Razão de Chances (odds ratio)

Além disso é importante calcular a razão de chances dos coeficientes, a qual é dada pelo exponencial dos coeficientes.

Tabela 15: Razão de Chances Modelo Logístico Binomial

Variável	**Razão de Chances**
Trancamentos por semestre	0,00006
Área do conhecimento	0,21
Semestres antes auxílio	2,47
Ingresso - Outros	1,68
Ingresso - PAS	1,44
Ingresso - Vestibular	3,58
Idade no ingresso	0,88
Razão de tempo	0,007

Dado os demais fatores contantes temos algumas conclusões a partir da interpretação dessa medida.

Disso tiramos que ao aumentar cada unidade da variável trancamentos por semestre a probabilidade de formatura é multiplicada por 0,00006.

Caso um aluno mude de algum curso para outro de Ciências Exatas e da Terra ou Engenharia, a probabilidade de formatura desse aluno cai na proporção de 0,21.

A cada semestre que o aluno ficou na UnB antes de começar a receber o auxílio, a probabilidade de formatura aumenta 2,47 vezes.

No caso de um aluno que ingressou na universidade pelo vestibular, esse possui uma probabilidade de formatura 3,58 vezes maior do que algum aluno que entrou pelo ENEM.

A cada aumento na unidade de razão de tempo, que é dada pela divisão entre a quantidade de semestres que o aluno ficou no auxílio e a quantidade de semestres que o aluno ficou na UnB isso possui um impacto negativo de 0,007 na probabilidade de formatura do aluno.

A partir do modelo podemos traçar perfis de alunos e compara-los com relação a probabilidade de formatura.

Por exemplo um aluno que ingressou na universidade em um curso da área da Saúde pelo Vestibular, tem 15 vezes mais chance de se formar do que um aluno que ingressou para um curso de Engenharia pelo ENEM.

7 Modelo Logístico Multinomial

O objetivo dessa sessão é criar um modelo de regressão logístico multinomial entre as categorias da variável que indica o motivo da saída do aluno do auxílio, sendo divida em 3 categorias, que são a Formatura, Desistência do Curso ou alunos em estado de vulnerabilidade social que utilizam o auxílio socieconômico que ainda estão ativos na UnB.

7.1 Seleção das Variáveis

Novamente foram testados os modelos com apenas uma variável explicativa, mas desta vez apenas as variáveis: Semestre de entrada na UnB, Campus, Raça, idade no ingresso e cota indígena não foram significativas. Todas as demais apresentaram p-valor maior que 0,25 (sugestão de Hosmer-Lemenshow) para o teste de significância dos coeficientes.

A seleção foi feita manualmente através do Stepwise, testando primeiro a inserção daquelas variáveis que foram significantes no modelo do caso binomial, sempre observando o p-valor para a significância dos coeficientes e o teste de máxima verossimilhança; onde a forma de ingresso e a idade não foram significativas. Em seguida as demais também foram testadas.

Após esse procedimento o modelo ficou com as variáveis: trancamentos por semestre, área de conhecimento, semestres antes do auxílio, razão de tempo, semestres na UnB, semestres no auxilio, ensino médio, usou cota e cota universal. Com 9 variáveis explicativas.

Ao rodar a função "multinom"para construir o modelo, o software desconsidera as observações que possuem valores faltantes as variáveis ensino médio, usou cota e cota universal passaram a ter zeros na tabela de contingência com a variável resposta. Devido a esse motivo exclusivo essas variáveis saem do modelo.

Sendo assim, com base no AIC e no p-valor da significância dos coeficientes, com o objetivo de encontrar um modelo mais parcimonioso, foi testada a saída de cada variável e o modelo final é dado por:

Tabela 16: Modelo Logístico Multinomial

	Variável	β_i	**P-valor**
Ativo	Intercepto β_0	-1,91	0,24
	Trancamentos por semestre	8,12	<0,001
	Área do conhecimento	0,51	0,01
	Semestres antes do auxílio	0,10	0,47
	Razão de Tempo	5,16	0,01
	Semestres na UnB	-0,23	<0,001
Desligado	Intercepto β_0	23,01	0,07
	Trancamentos por semestre	14,38	<0,001
	Área do conhecimento	3,41	0,01
	Semestres antes do auxílio	-5,61	0,07
	Razão de Tempo	-25,59	0,05
	Semestres na UnB	-0,29	<0,001

Importante ressaltar que a categoria utilizada como referência na variável resposta é a "Formatura".

Ao comparar os alunos ativos com os demais que foram estudados no modelo logístico binomial, as variáveis trancamentos por semestre, área do conhecimento, razão de tempo e semestres na UnB foram significativas.

Para o modelo em questão foi escolhido o α (significância) de 10%, o que justifica a entrada da variável Semestres antes do auxílio. Indo de acordo com o que foi observado na sessão anterior, ocorre uma diminuição da probabilidade do aluno se desligar quando aumenta a quantidade de semestres que o aluno ficou na UnB antes de receber o auxílio.

Seguindo o que foi observado no modelo binomial, a razão de tempo entre Formatura e Desligamento mostra que o aluno que possui uma porcentagem maior da sua graduação recebendo o auxílio socieconômico, possui uma menor probabilidade de formatura. E o diferencial desse modelo é que adicionando os alunos ativos, esses possuem mais chances de se formarem quanto maior a razão de tempo.

7.2 Qualidade do ajuste, diagnóstico e Poder Preditivo

O objetivo desta sessão é avaliar a qualidade do ajuste do modelo, realizar seu diagnóstico e falar sobre seu poder preditivo, para isso, algumas das medidas sobre o modelo escolhido estão na Tabela 17 abaixo.

Tabela 17: Medias de Diagnóstico do Modelo de Regressão Logística Multinomial

Medida	**Valor**
AIC	1203
BIC	1262
Deviance	1179

Essas medidas foram levadas em consideração no momento de escolher o modelo mais apropriado, sabendo que valores menores são preferíveis.

Além disso, foi realizado o teste de Hosmer-Lemeshow ao nível de significância de 5%, para ajudar a identificar se o modelo é adequado, com base nas hipóteses:

$$\begin{cases} H_0 : \text{O modelo está bem ajustado aos dados} \\ H_1 : \text{O modelo não está bem ajustado aos dados} \end{cases}$$

E o resultado se encontra na tabela a seguir.

Tabela 18: Teste de Hosmer-Lemenshow

Resultados do teste	
Estatística do Teste	21,27
P-valor	0,17
Graus de Liberdade	16

Não encontrando evidências para rejeitar a hipótese nula de que o modelo está bem ajustado aos dados.

Com relação a capacidade preditiva do modelo observa-se a matriz de confusão e as medidas calculadas a partir da mesma:

Tabela 19: Matriz de confusão do poder preditivo do modelo logístico multinomial

		Amostra		
		Formatura	Ativo	Desistência
Previsto	Formatura	183	147	15
	Ativo	380	1269	86
	Desistência	0	1	10

Tabela 20: Medidas sobre o poder preditivo do modelo logístico multinomial

	Acurácia	**Sensibilidade**	**Especificidade**
Formatura	70%	32%	89%
Ativo	70%	90%	31%
Desistência	70%	9%	100%

Como existem muitos alunos ativos, é de se esperar que o modelo aloque muitas das informações nessa categoria, o que de fato foi observado.

Observou-se, que para os alunos ativos estes se aproximam mais da formatura do que da desistência.

O modelo apresentou um resultado satisfatório com acurácia de 70%.

A Sensibilidade nos indica, com base nos dados de validação, a porcentagem de vezes que o modelo classificou como sucesso cada categoria da variável resposta. Dentre os alunos ativos esse indicador é igual a 90% para os alunos ativos, os quais são o principal objetivo deste modelo.

A Especificidade, fala sobre a proporção em que o modelo calculou como "fracasso"nos momentos em que de fato é uma categoria de "fracasso". Foi obtido um resultado acima de 80% somente para os que se formaram ou desistiram, dentre os alunos ativos a especificidade do modelo foi de apenas 31%.

8 Conclusão

Tendo em vista o objetivo principal de avaliar as características sociais dos alunos que recebem ou já receberam o benefício Auxílio Socieconômico, que impactam na sua probabilidade de formatura, esta avaliação foi sendo construída no decorrer do trabalho através das seguintes etapas:

Etapa 1: Iniciou-se a identificação dos dados recebidos, de modo a uniformizar e organizar as informações necessárias;

Etapa 2: Efetuada a analise exploratória para observar a qualidade dos dados recebidos e identificar e tratar as divergências;

Etapa 3: Efetuado ajustes nos modelo logístico binário e multinomial. O primeiro modelo levou em consideração os alunos que já se formaram e os alunos que saíram o auxílio por desistirem do curso. O segundo modelo acrescentou os alunos que estão ativos na universidade, com o objetivo de classifica-los da melhor forma e identificar quais fatores os aproximam da formatura ou da desistência;

Etapa 4: Efetuada a análise a partir das informações estatísticas dos dois modelos ajustados. Com o modelo de regressão logística binária pôde-se observar que o modelo se ajustou bem aos dados, com uma acurácia de 84%. Onde foi observado que as variáveis trancamentos por semestre, área do conhecimento, quantidade de semestres antes do auxílio, razão de tempo, forma de ingresso e idade ao ingressar na UnB foram significativos e impactam a probabilidade de formatura dos alunos que recebem o auxílio socieconômico da Universidade de Brasília.

A partir do modelo multinomial, que permitiu avaliar também os alunos em atividade da universidade, possui acurácia de 70%. Novamente as variáveis trancamentos por semestre, área do conhecimento, quantidade de semestres antes do auxílio e razão de tempo foram significativas.

A razão de tempo se destacou com um resultado diferente do que foi encontrado no modelo binário, o que já era de se esperar, visto que os alunos ativos, em geral, passaram menos semestres na faculdade do que os formados. Foi adicionada no modelo a variável semestres na UnB, que foi significativa ao diferenciar os grupos de alunos entre ativos, formados e desistentes.

Apos a construção destas etapas, conclui-se neste presente trabalho que:

- Infere-se a partir do desenvolvimento do projeto que existem pontos de melhoria na gestão do auxilio que com certeza trariam mais efetividade do auxilio para a sociedade, e traria um melhor aproveitamento dos recursos públicos investidos pelo estado em alunos em estado de vulnerabilidade social. Destacam-se dois pontos

principais de melhoria: O saneamento e a gestão dos dados disponíveis, e um sistema de monitoramento do aproveitamento do aluno, com uma prestação recorrente da assiduidade dos alunos nos cursos e seu desempenho.

- A qualidade dos dados é muito importante para a eficiência e eficácia das analises, e foi observado uma taxa muito alta de dados errados, divergentes e sem relações com outros sistemas, o que exigiu um esforço muito grande na uniformização e manipulação dos dados. Neste item destacam-se as recentes mudanças em sistemas na UnB, que não possuem algumas integrações de dados e validações de campos, e também carecem de informações chave em comum de alunos, o que permitiria que dados de um sistema fossem facilmente recuperados de outro sistema;

- A base de dados analisada não possui confiabilidade em algumas informações que seriam muito relevantes para este projeto, como por exemplo a Renda dos alunos;

- Observou-se a falta de algumas informações complementares que seriam relevantes para a evolução do tema, como por exemplo, se houvesse informações sobre os alunos que não foram selecionados para receber o auxílio;

- Por fim, conclui-se pela análise estatística que a falta de qualidade dos dados e a inexistência de um processo eficaz de monitoramento dos dados está causando uma baixa efetividade no atingimento da principal função do auxílio, que é ajudar os alunos em estado vulnerável a se manter na faculdade, cumprir a etapa de formatura, e finalmente devolver o investimento feito pelos cidadãos para eles próprios, através de profissionais formados que, sem o auxilio, dificilmente conseguiriam cumprir estas etapas.

Referências

BENDEL, R. B.; AFIFI, A. A. Comparison of stopping rules in forward "stepwise" regression. *Journal of the American Statistical association*, Taylor & Francis, v. 72, n. 357, p. 46–53, 1977.

BRASÍLIA, U. de. *Formas de Ingresso*. 2022. Ultimo acesso em 11 de julho de 2022 às 18:22. Disponível em: ⟨https://saa.unb.br/graduacao/formas-de-ingresso⟩.

BRASÍLIA, U. de. *Formas de Ingresso*. 2022. Ultimo acesso em 13 de julho de 2022 às 16:28. Disponível em: ⟨https://www.unb.br/graduacao/cursos⟩.

BUSSAB, W. de O.; MORETTIN, P. A. *LaTeX :Estatística B´sica. 9. ed. [S.l.]: Editora Saraiva, 2017.*

DÁVILA, V. H. L. Teste de hipóteses. *Instituto de Matemática, Estatística e Computação, UNICAMP. Disponível em:¡ https://www. ime. unicamp. br/˜hlachos/Inferencia_ Hipo1.pdf*, p. 3, 2017.

HOSMER, D. W.; LEMESHOW, S. *LaTeX :Applied Logistc Regression*. 2. ed. Columbus, Ohio: [s.n.], 2000.

LIKERT, R. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

MAGALHÃES, M. N. *LaTeX :Probabilidade e Vari´veis Aleatórias. 3. ed. [S.l.]: Editora da Universidade de S˜o Paulo, 2015.*

MICKEY, R. M.; GREENLAND, S. The impact of confounder selection criteria on effect estimation. American journal of epidemiology*, Oxford University Press, v. 129, n. 1, p. 125–137, 1989.*

OLIVEIRA, J. P. D. D.; SILVA, T. T. D. Sobre as distribuições binomial e multinomial. Revista de Matemática*, v. 5, n. 1, p. 1–28, 2018.*

OSÓRIO, R. G. O sistema classificatório de cor ou raça do ibge. Instituto de Pesquisa Econômica Aplicada (Ipea), 2003.

PRÁTICA.ORG, N. Conheça as áreas de conhecimento. *2021. Ultimo acesso em 14 de julho de 2022 às 13:24. Disponível em: ⟨https:// www.napratica.org.br/ conheca-areas-do-conhecimento/⟩.*

REPuBLICA, P. da. DECRETO Nº 7.234. *2010. Ultimo acesso em 08 de março de 2021 às 13:15. Disponível em: ⟨http:// www.planalto.gov.br/ ccivil_03/ _ato2007-2010/ 2010/ decreto/ d7234.htm⟩.*

RODRÍGUEZ, E. Y. A. Técnicas de aprendizado de máquina para predição do custo da logística de transporte: uma aplicação em empresa do segmento de autopeças. Universidade Estadual Paulista (UNESP), 2020.

SANT'ANNA, M. C. de; ALMEIDA, A. N. de. Processos de trabalho da assistência estudantil no ensino superior: uma percepção dos assistentes sociais da universidade de brasília (unb). Administração Pública e Gestão Social*, 2021.*

TROVÃO, C. A pandemia da covid-19 e a desigualdade de renda no brasil: um olhar macrorregional para a proteção social e os auxílios emergenciais. Natal: Universidade Federal do Rio Grande do Norte*, 2020.*

UNB, D. Programa Auxíílio Socioeconômico*. 2021. Ultimo acesso em 08 de março de 2021 às 17:22. Disponível em: ⟨http:// dds.dac.unb.br/ index.php/ auxilio-socioeconomico⟩.*