UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

AMANDA DA SILVA VASCONCELOS

ANÁLISE SOBRE A EVASÃO DE ALUNOS DO CURSO DE ESTATÍSTICA NA FCT/UNESP

PRESIDENTE PRUDENTE
2023

AMANDA DA SILVA VASCONCELOS

ANÁLISE SOBRE A EVASÃO DE ALUNOS DO CURSO DE ESTATÍSTICA NA FCT/UNESP

Relatório Final de Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/UNESP para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientador: Profª. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE

2023

V331a

Vasconcelos, Amanda da Silva
Análise sobre a evasão de alunos do curso de Estatística na FCT/Unesp / Amanda da Silva Vasconcelos. -- Presidente Prudente, 2023
50 p. : il., tabs.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientadora: Miriam Rodrigues Silvestre

1. Evasão escolar. 2. Estatística. 3. Análise de Regressão logística. 4. Árvores de Decisão. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

AMANDA DA SILVA VASCONCELOS

ANÁLISE SOBRE A EVASÃO DE ALUNOS DO CURSO DE ESTATÍSTICA NA FCT/UNESP

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____
Profª. Dra. Miriam Rodrigues Silvestre
Departamento de Estatística

_____
Prof. Dr. Klaus Schlünzen Junior
Departamento de Estatística

_____
Profª. Dra. Silvely Nogueira de Almeida Salomão Néia
Departamento de Estatística

Presidente Prudente, 27 de janeiro de 2023.

RESUMO

Um dos grandes problemas que aflige a educação há uma longa data até a atualidade é a evasão escolar, principalmente no ensino superior. Essa é uma realidade também no curso de graduação em Estatística na FCT/UNESP de Presidente Prudente - SP. Visando identificar os perfis de alunos mais propensos a evadirem do curso, foram utilizadas duas técnicas estatísticas neste trabalho. A regressão logística e a árvore de decisão são técnicas de classificação, ou seja, a partir delas conseguimos prever uma variável categórica, no caso do presente estudo, se o aluno irá ou não evadir do curso. Essas técnicas foram aplicadas em uma base de dados com variáveis sociodemográficas, como idade, sexo, tipo de escola cursada no ensino médio, etc., de alunos ingressantes entre os anos de 2008 e 2017 no curso de Estatística, com o objetivo de terem seus resultados comparados. As análises apresentadas foram realizadas no software R. Ao comparar as duas técnicas pela métrica F1-score, a árvore de decisão se ajustou melhor no treino, porém obteve-se maior poder de previsão na regressão logística.

Palavras-chave: Evasão Escolar. Estatística. Análise de Regressão Logística. Árvores de Decisão.

ABSTRACT

One of the major problems that has afflicted education for a long time until today is school dropout, especially in higher level of education. This is also a reality in the undergraduate course in Statistics at FCT/UNESP in Presidente Prudente - SP. In order to identify the profiles of students most likely to dropout of the course, two statistical techniques were used in this work. Logistic regression and decision tree, which are classification techniques, that is, from them we can predict a category variable, in the case of this study, whether or not the student will dropout of the course. These techniques were applied to a database with sociodemographic variables, such as: age, gender, type of high school attended, etc., of students entering the Statistics course between 2008 and 2017, with the aim of having their results compared. The analyzes presented were performed in the programming language R. When comparing the two techniques by the F1-score metric, the decision tree fitted better in training, but greater predictive power was obtained in logistic regression.

Keywords: School Dropout. Statistics. Logistic Regression Analysis. Decision Trees.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A evasão escolar, ato de deixar de frequentar as aulas em que se está matriculado ou abandono total dos estudos, é um fenômeno que afeta todos os níveis educacionais. Se tratando do ensino superior, as perdas em decorrência disso, assim como nos demais graus educacionais, são enormes. É uma perda social, pois é uma vaga que poderia estar sendo ocupada por um aluno que futuramente se formaria e contribuiria para a sociedade. As perdas pessoais também podem ser grandes, pois o estudante perde tempo fazendo um curso que, provavelmente, não usará no futuro até que decida evadir, além do gasto da família, principalmente se a universidade for em outra cidade. Além disso, aquela vaga evadida traz também prejuízos a universidade, seja ela pública ou privada.

Cardoso (2008) diferencia dois tipos de evasões, uma nomeada por evasão aparente, que se refere ao estudante que faz transferência de curso dentro da universidade ou para outra e a evasão real, quando o estudante abandona e não conclui o curso.

Estudos feitos em várias universidades relatam que a maior porcentagem de evasão está em cursos da área de ciências exatas, como na UDESC – Universidade do Estado de Santa Catarina, onde:

> “Os resultados apontam que os menores índices de evasão, no período analisado, ocorreram nos cursos da área de Ciências da Saúde (19,6%) e os maiores nos cursos das áreas de Ciências Exatas e da Terra (58,6%), Engenharia (41,0%), e Linguística, Letras e Artes (45,9%).” (DAVOK, BERNARD, 2016, p.503)

Segundo dados sobre evasão universitária, da Secretaria de Educação Superior do Ministério da Educação (SESU/MEC), de 87 cursos avaliados em universidades federais no ano de 2018, Estatística ocupou a 9ª posição, com 27,9% de alunos que deixaram o curso sem se formar (PINTO, 2019).

Este presente trabalho visa apresentar a análise da evasão dos estudantes da estatística da FCT/Unesp através de variáveis sociodemográficas, a fim de traçar o perfil de estudantes mais propensos a evadirem do curso, fazendo o uso da regressão logística e árvore de decisão, que são técnicas estatísticas e posteriormente comparar seus resultados.

2 FUNDAMENTOS TEÓRICOS

2.1 Regressão Logística

A regressão logística é uma técnica estatística que, segundo Hosmer e Lemeshow (2013), relaciona uma variável dependente a uma ou mais variáveis independentes, onde, geralmente, a variável dependente é binária. Tachibana (2020) reforça que a análise com essa técnica procura encontrar o modelo mais parcimonioso.

Esta técnica estatística determina a probabilidade de um evento acontecer, para isso, a variável resposta deve ser categórica, e é mais comumente utilizada para variáveis dicotômicas, exemplo: pagou ou não pagou, fraudou ou não fraudou, sobreviveu ou morreu, etc. Também é muito empregada por ser de fácil interpretação.

2.1.1 Modelo de Regressão Logística

Segundo Hosmer e Lemeshow (2013), considerando um vetor de p variáveis independentes $x^t = (x_1, x_2, \dots, x_p)$ e sendo a probabilidade condicional dada por $P(Y = 1|x) = \pi(x)$, o logito do modelo de regressão logística múltipla é denotado pela seguinte equação:

$$g(x) = ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p, \tag{2.1}$$

onde, o modelo de regressão logística é:

$$\pi(x) = \frac{e^{g(x)}}{1 + e^{g(x)}}, \tag{2.2}$$

logo, o modelo de regressão logística múltipla é dado por:

$$\pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}}, \tag{2.3}$$

onde $\beta_1, \beta_2, \dots, \beta_p$ são parâmetros que são estimados pela função log-verossimilhança, dada por:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} [y_i ln\pi(x_i) + (1 - y_i)\ln\,(1 - \pi(x_i))]. \tag{2.4}$$

2.1.2 Testes de significância do modelo

Para avaliar o modelo, o primeiro passo é testar se pelo menos uma variável independente é significativa para v ariável resposta, para isso compara-se o log da verossimilhança do modelo completo com o log da verossimilhança para o modelo com apenas o intercepto. Esse é o teste da Razão da Verossimilhança, com as hipóteses:

$$\begin{cases} H_0: \beta_1 = \beta_2 = \cdots = \beta_p = 0 \\ H_1: pelo\ menos\ um\ \beta_j \neq 0 \end{cases}. \tag{2.5}$$

A estatística do teste de Razão de Verossimilhança é dada por:

$$G = -2ln\left[\frac{verossililhança\ do\ modelo\ sem\ a\ variável}{verossimilhança\ do\ modelo\ com\ a\ variável}\right], \tag{2.6}$$

ou então, por:

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i \ln(\hat{\pi}_i) + (1 - y_i)\ln\,[1 - \hat{\pi}_i]\right] - \{n_1 ln(n_1) + n_0 \ln(n_0) - nln(n)\}\right\}, \tag{2.7}$$

em que a estatística G segue distribuição χ^2 com p (número de variáveis) graus de liberdade. Então se χ^2 calculado for maior que o valor de χ^2 tabelado, rejeita-se H_0 e conclui-se que há pelo menos uma variável significativa para o modelo.

Outra etapa muito importante na validação do modelo de regressão logística, é testar se cada coeficiente é diferente de zero, para isso usa-se o teste de Wald, que

verifica se há relação significativa entre cada uma das variáveis independentes com a variável resposta, suas hipóteses são:

$$\begin{cases} H_0: \beta_j = 0 \\ H_1: \beta_j \neq 0 \end{cases}, \qquad j = 0,1, \dots, p. \tag{2.8}$$

A estatística do teste de Wald é dada por:

$$W_j = \frac{\hat{\beta}_j}{\sqrt{\widehat{Var}(\hat{\beta}_j)}}, \tag{2.9}$$

onde W_j segue distribuição normal padrão, assim, ${W_j}^2$ segue uma distribuição χ^2 com 1 grau de liberdade. Então se χ^2 calculado for maior que o valor de χ^2 tabelado, rejeita-se H_0 e conclui-se que a variável é significativa para o modelo.

2.1.3 Intervalos de Confiança para os parâmetros

Os Intervalos de Confiança (I.C) para inclinação e para o intercepto são baseados no teste de Wald. O I.C. para inclinação é dado por:

$$I.C.(\beta_j)_{100(1-\alpha)\%}: \left[\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} D.P(\hat{\beta}_j)\right] \; com \; j = 1,2, \dots, p. \tag{2.10}$$

E para o intercepto, como:

$$I.C.(\beta_0)_{100(1-\alpha)\%}: \left[\hat{\beta}_0 \pm z_{1-\frac{\alpha}{2}} D.P.(\hat{\beta}_0)\right]. \tag{2.11}$$

O intervalo de confiança para o logito é dado como:

$$I.C.(g(x))_{100(1-\alpha)\%}: \left[\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} D.P.(\hat{g}(x))\right]. \tag{2.12}$$

E o intervalo de confiança para o valor ajustado é:

$$I.C.(\pi)_{100(1-\alpha)\%}: \left[\frac{e^{\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} D.P.(\hat{g}(x))}}{1 + e^{\hat{g}(x) \pm z_{1-\frac{\alpha}{2}} D.P.(\hat{g}(x))}} \right]. \quad (2.13)$$

2.1.4 Odds Ratio

Diferentemente da regressão linear, que estima-se um valor de -∞ a +∞, na regressão logística estima-se uma probabilidade, que varia de 0 a 1. Outra diferença entre os dois modelos de regressão é a interpretação dos parâmetros, na regressão linear essa interpretação é direta enquanto na logística é necessário a Odds Ratio ou Razão de Chances. Kleinbaum e Mitchel (2006) enfatiza que, a razão de chances não é a razão de risco, e que ela é a única estimativa de associação direta da regressão logística.

A Odds Ratio (OR) é a exponencial do parâmetro que se quer interpretar, isso porque:

$$\mathrm{OR} = \frac{\frac{\pi(1)}{[1-\pi(1)]}}{\frac{\pi(0)}{[1-\pi(0)]}} = \frac{\frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)}{\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)}}{\frac{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)}{\left(\frac{1}{1+e^{\beta_0}}\right)}} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{(\beta_0+\beta_1)-\beta_0}. \quad (2.14)$$

Tem-se, portanto, que:

$$\mathrm{OR} = e^{\beta_1}. \quad (2.15)$$

Suponhamos que Y indica se o paciente teve ou não infarto, e X se o paciente fuma ou não fuma. Presumindo ainda que ao calcular a odds ratio, obteve-se OR=3, interpreta-se então que, a chance de infarto em pacientes que fumam é três vezes maior do que a chance de infarto em pacientes que não fumam.

Seu intervalo de confiança é dado por:

$$I.C.(OR)_{100(1-\alpha)\%}: \left[exp\left[\hat{\beta}_j \pm z_{1-\frac{\alpha}{2}} D.P(\hat{\beta}_j) \right] \right] \; com \; j = 1,2,\dots,p. \tag{2.16}$$

2.1.5 Medidas resumo da qualidade do ajuste

As medidas resumo da qualidade do ajuste não dizem sobre as variáveis do modelo individualmente, mas valores exorbitantes indicam algum problema no modelo. As medidas utilizadas serão: deviance e a estatística qui-quadrado de Pearson.

2.1.5.1 Deviance

Os valores ajustados na regressão logística são dados por:

$$\hat{y}_j = m_j \hat{\pi}_j = m_j \left\{ \frac{e^{\hat{g}(x_j)}}{1 + e^{\hat{g}(x_j)}} \right\}. \tag{2.17}$$

A deviance residual será dada por:

$$d(y_j, \hat{\pi}_j) = \pm \left\{ 2 \left[y_j ln\left(\frac{y_j}{m_j \hat{\pi}_j} \right) + (m_j - y_j) ln\left(\frac{(m_j - y_j)}{m_j(1 - \hat{\pi}_j)} \right) \right] \right\}^{1/2}, \tag{2.18}$$

onde o sinal + ou – é o mesmo de $(y_j - m_j \hat{\pi}_j)$. Para covariável padrão $y_j = 0$, a deviance residual será dada por:

$$d(y_j, \hat{\pi}_j) = -\sqrt{2m_j \left| ln(1 - \hat{\pi}_j) \right|}. \tag{2.19}$$

Quando $y_j = m_j$, a deviance residual é:

$$d(y_j, \hat{\pi}_j) = \sqrt{2m_j \left| ln\hat{\pi}_j \right|}. \tag{2.20}$$

A estatística resumo baseada na deviance residual é a deviance, que é:

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2, \tag{2.21}$$

com D seguindo distribuição ${\chi^2}_{J-(p+1)}$.

2.1.5.2 Estatística Qui-Quadrado de Pearson

A medida de Pearson para a diferença dos valores observados e predito é:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1 - \hat{\pi}_j)}}. \tag{2.22}$$

E a estatística qui-quadrado de Pearson é dada por:

$$X^2 = \sum_{j-1}^{J} [r(y_j, \hat{\pi}_j)]^2. \tag{2.23}$$

2.2 Árvore de Decisão

A árvore de decisão é uma técnica estatística que auxilia na tomada de decisão e é descrita por Lauretto (1996) como “Uma coleção de elementos chamados nós, dentre os quais um é distinguido como uma raiz, juntamente com uma relação de “paternidade” que impõe uma estrutura hierárquica sobre os nós.” Essa é uma técnica construída a partir de um conjunto de filtros, sua premissa é decompor um problema complexo em problemas menores, sendo um algoritmo de simples visualização, que lida bem com dados missings, diferente da regressão logística.

Para construir uma árvore de decisão, vamos supor que você irá fazer um piquenique com seus amigos, mas para isso vocês se fazem algumas perguntas, o tempo está ensolarado ou chuvoso? Se estiver ensolarado vocês vão fazer piquenique, se estiver chuvoso, não. Estando ensolarado, vocês pensam em um local, o shopping não pareceu muito agradável, portanto, vocês escolhem um parque. Pensando em um período do dia, de manhã e à tarde pareceu legal para todos, já a noite não, com isso, tem-se visualmente a árvore dada pela Figura 1:

Figura 1 - Árvore de decisão para a escolha de fazer um piquenique

Fonte: Elaborada pela autora

Na árvore de decisão as perguntas são os nós, as opções são os ramos e a resposta são as folhas. No exemplo ilustrado acima, o tempo, o local e o período são chamados de nós, onde o tempo por ser a primeira pergunta é o nó raiz, e o local e o período são os nós intermediários, as opções, chuvoso, ensolarado, shopping, parque, manhã, tarde e noite, são os ramos, e as respostas sim e não são as folhas.

2.2.1 Escolha dos melhores atributos

O processo de criação de uma árvore pode ser muito infactível computacionalmente se não existirem algumas regras, pois existem inúmeras possibilidades no processo de criação de uma árvore. Portanto, existem algumas medidas que auxiliam os algoritmos nesse processo e que maximizam o ganho de informação, sendo elas a entropia, o ganho de informação e o índice de Gini.

2.2.1.1 Entropia

A entropia é uma medida de impureza no conjunto, ou seja, o quão heterogêneo é a sua composição. É uma medida que varia de 0 a 1, onde 0 indica pureza no conjunto e 1 indica impureza.

Simões (2008) denota a entropia de um conjunto S de s amostras com m classes distintas C_i (i=1,2,...,m), por:

$$Entropia(S) = -\sum_{i=1}^{m} p_i \, log_2(p_i), \tag{2.24}$$

onde p_i é a probabilidade de um elemento qualquer pertencer a classe C_i, calculada por s_i/s.

2.2.1.2 Ganho de Informação

O ganho de informação é uma medida que calcula a diminuição da entropia esperada ao utilizar um atributo A na partição do conjunto de dados. Sua fórmula é dada por:

$$Ganho(S, A) = Entropia(S) - \sum_{j=1}^{m} \frac{|S_j|}{|S|} Entropia(S_j). \tag{2.25}$$

2.2.1.3 Índice de Gini

O índice de Gini é uma medida de impureza no nó, ou seja, o quão heterogêneo é a sua composição. É uma medida que varia de 0 a 1, quando o Gini é 0 indica um nó puro e quando é 1 indica um nó impuro:

$$Gini(S) = 1 - \sum_{i=1}^{m} {p_i}^2, \tag{2.26}$$

onde p_i é a frequência relativa de cada classe em cada nó e m é o número de classes.

Silva (2005) diz que para partições binárias, o índice de Gini é usado para isolar os registros que caracterizam a classe com maior frequência em um ramo e a entropia é utilizada para balancear o número de registros nos ramos.

2.3 Métricas de Qualidade de Ajuste dos Modelos

Depois da criação dos modelos de classificação é preciso calcular algumas métricas a fim de ver a performance do modelo, se ele está sendo capaz de discriminar as observações de forma correta. Existem diversas métricas de qualidade de ajuste, mas a que serão abordadas no estudo serão acurácia, precisão, recall e F1-score, e para calculá-las é utilizada a matriz de confusão.

A matriz de confusão é uma tabela que apresenta os valores preditor pelo modelo vs o valor real observado, com isso tem-se os verdadeiros positivos, que são quando o modelo classifica como categoria de interesse e realmente era; os falsos negativos, que são quando o modelo classifica como não categoria de interesse, mas era; os falsos positivos, que são quando o modelo classifica como categoria de interesse, mas não é; e os verdadeiros negativos, que são quando o modelo classifica como não categoria de interesse e realmente não é. A matriz de confusão é definida como segue no Quadro 1.

Quadro 1 - Matriz de confusão

	Predito		
Real		$Y = 1$	$Y = 0$
	$Y = 1$	Verdadeiro Positivo (VP)	Falso Negativo (FN)
	$Y = 0$	Falso Positivo (FP)	Verdadeiro Negativo (VN)

Fonte: Elaborado pela autora

2.3.1 Acurácia

A acurácia é uma métrica que indica a proporção de classificações corretas do modelo em relação a todas. A acurácia é dada por:

$$Acurácia = \frac{VP + VN}{VP + FN + FP + VN}. \tag{2.27}$$

2.3.2 Precisão

A precisão é uma métrica que indica a proporção de classificações corretas do modelo em relação aos preditos como categoria de interesse. A precisão é dada por:

$$Precisão = \frac{VP}{VP + FP}. \tag{2.28}$$

2.3.3 Recall

O recall é uma métrica que indica a proporção de classificações corretas feitas pelo modelo dentre todos os valores esperados da categoria de interesse. O recall é dado por:

$$Recall = \frac{VP}{VP + FN}. \tag{2.29}$$

2.3.4 F1-score

O F1-score é uma métrica que possibilita a observação da precisão e do recall em um número. O F1-score é dado por:

$$F1 = \frac{2 * Precisão * Recall}{Precisão + Recall}. \quad (2.30)$$

3 APLICAÇÃO

3.1 Base de Dados e Descrição

O objeto de estudo são estudantes do curso de Estatística da FCT/Unesp, câmpus de Presidente Prudente – SP. A base de dados foi fornecida pela seção técnica de graduação e contém informação de 216 alunos ingressos no curso de graduação em Estatística na FCT/UNESP de Presidente Prudente entre 2008 e 2017, contendo 139 formados e 77 evadidos.

São considerados como evadidos todos os ingressos: desistentes, ingressantes em outro curso, que cancelaram no vestibular, que abandonaram, que jubilaram ou que não renovaram a matrícula. Nessa base os alunos estão mascarados, ou seja, não há qualquer informação pessoal do discente (como nome, cpf, telefone, etc.).

A base disponibilizada não continha a variável distKM, que é referente a distância da cidade de origem do aluno a Presidente Prudente -SP em KM, ela foi criada a partir da variável cidadeorigem.

As variáveis presentes na base de dados que serão utilizadas são as apresentadas no Quadro 2.

Quadro 2 - Dicionário de dados

Variável	Descrição	Tipo
classe	Aluno evadido ou formado	Qualitativa nominal
sexo	Sexo do aluno: feminino ou masculino	Qualitativa nominal
cidadeorigem	Cidade de origem do aluno	Qualitativa nominal
distKM	Distância da cidade de origem do aluno a Presidente Prudente em KM	Quantitativa contínua
cor	Cor do aluno: amarelo, branco, pardo, preto ou não informado	Qualitativa nominal
ensinomedio	Aluno concluiu toda ou maior parte do ensino médio em escola: privada ou pública	Qualitativa nominal

Continua.

Cont. Quadro 2.

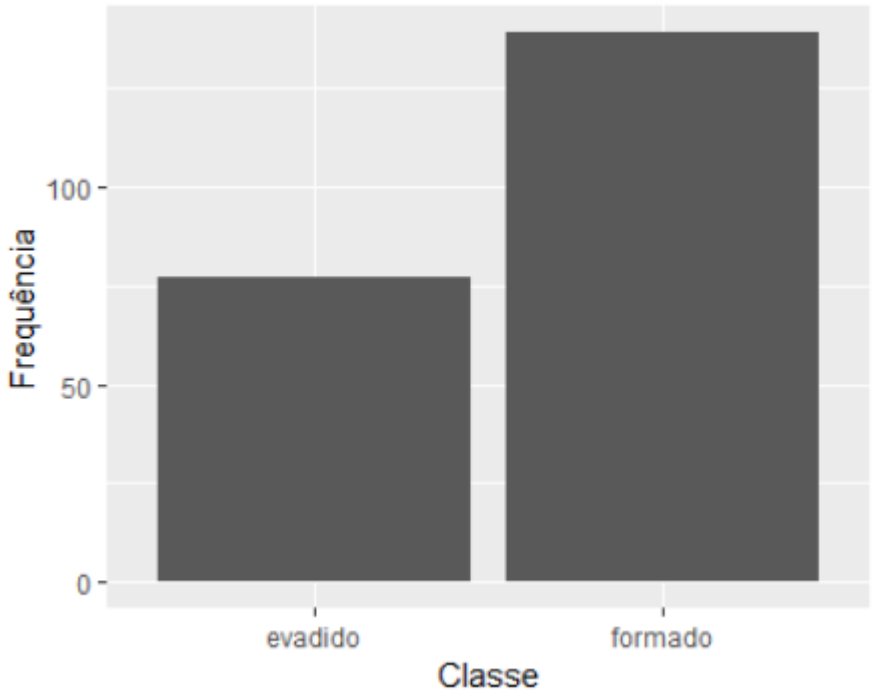
cotas	Aluno entrou pelo sistema de cotas: sim (ensino público ou PPI – Preto, pardo ou índio) ou não	Qualitativa nominal
classvest	Classificação do aluno no vestibular	Quantitativa discreta
notavest	Nota do aluno no vestibular	Quantitativa contínua
idade	Idade do aluno ao ingressar no curso	Quantitativa discreta

Fonte: Elaborado pela autora

A análise exploratória dos dados é uma etapa de extrema importância, pois a partir dela entendemos o comportamento das variáveis, sua relação com a variável respostas e assim, quais serão importantes para a etapa de modelagem. Existe a análise univariada, onde observamos uma variável por vez e a bivariada ou multivariada, onde observamos a relação entre duas ou mais variáveis. A análise exploratória do banco de dados estudado será apresentada a seguir da Figura 2 a Figura 18 e da Tabela 1 a Tabela 9.

A variável classe é a variável resposta, ou seja, o que queremos prever no processo de modelagem. A proporção de evadidos é um pouco inferior a de formados, como pode-se ver na Figura 2 e na Tabela 1, mas dado o contexto, observamos que realmente a evasão no curso de Estatística da FCT/UNESP é alta.

Figura 2 - Gráfico de barras para a variável classe



Fonte: Elaborada pela autora

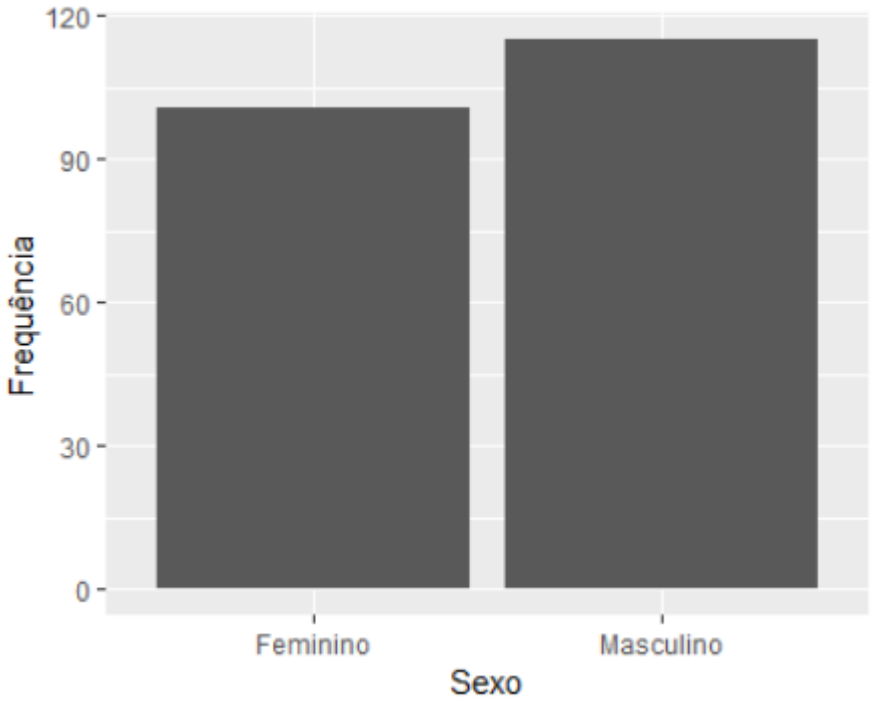
Tabela 1 - Proporção da variável classe

Evadido	Formado
35,65%	64,35%

Fonte: Elaborada pela autora

Observamos, a partir da Figura 3 e da Tabela 2, que embora os sexos estejam próximos, ainda existe uma predominância de ingressantes do sexo masculino no curso.

Figura 3 - Gráfico de barras para a variável sexo



Fonte: Elaborada pela autora

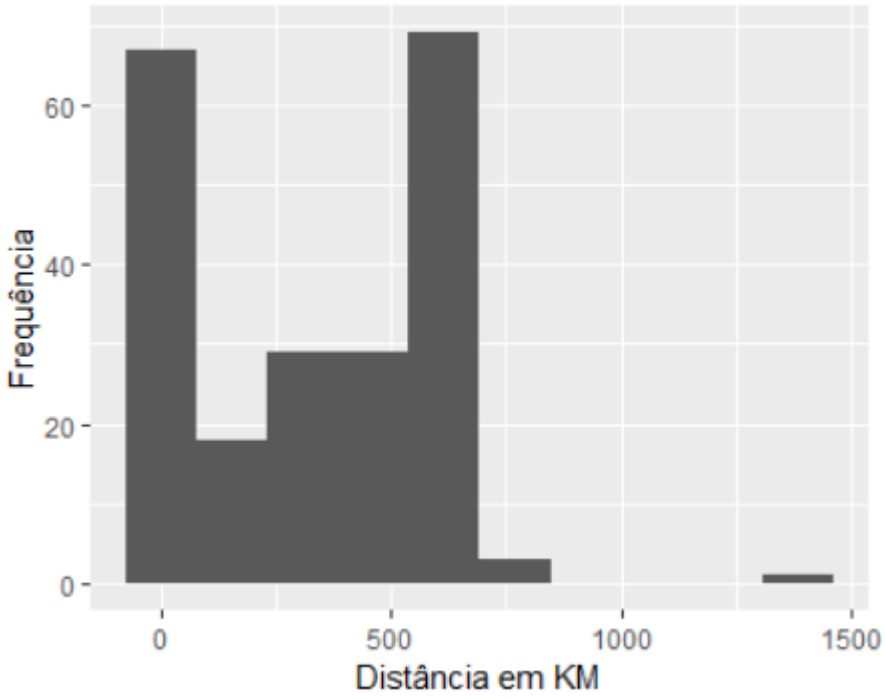
Tabela 2 - Proporção da variável sexo

Feminino	**Masculino**
46,76%	53,24%

Fonte: Elaborada pela autora

A variável distKM mostra a distância da cidade de origem do aluno a Presidente Prudente em quilômetros. A partir da Figura 4 e da Tabela 3, observa-se que muitos ingressantes são de Presidente Prudente ou cidades próximas, e 75% está em um raio de até 554km e apenas um está a mais de 1000km de Presidente Prudente, vindo de São Francisco – MG.

Figura 4 - Histograma da variável distKM



Fonte: Elaborada pela autora

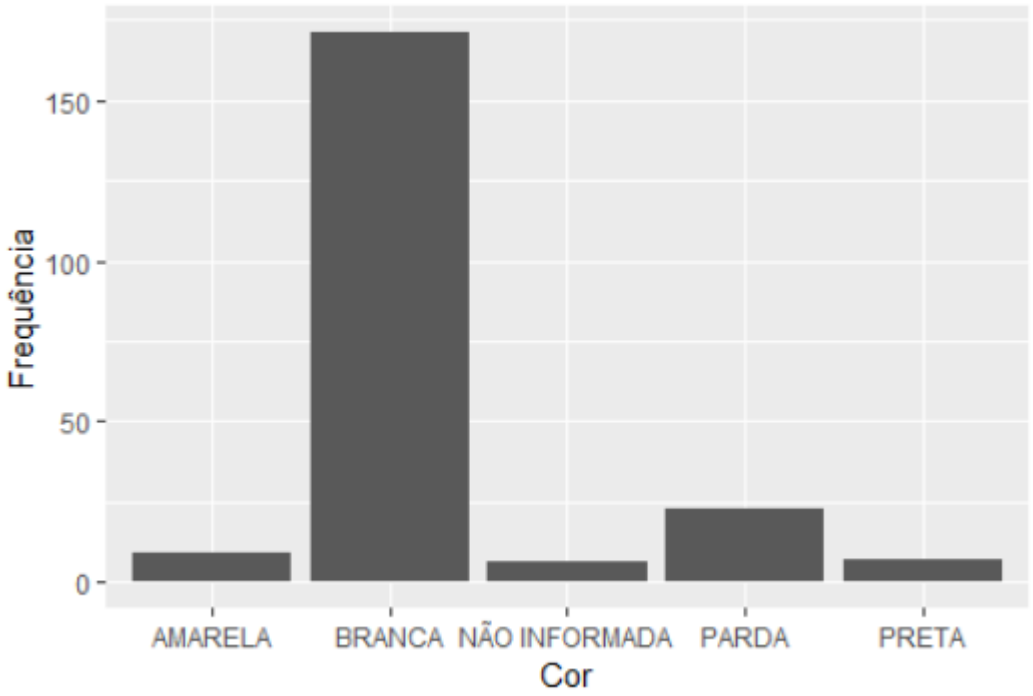
Tabela 3 - Medidas resumo da variável distKM

Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
1	21	359	554	1386	318,7	255,35

Fonte: Elaborada pela autora

Vemos na Figura 5 e Tabela 4 que a grande maioria dos ingressantes do curso, com quase 80%, são brancos, seguidos de pardos, amarelos, pretos e a menor parte não informou sua cor.

Figura 5 - Gráfico de barras para a variável cor



Fonte: Elaborada pela autora

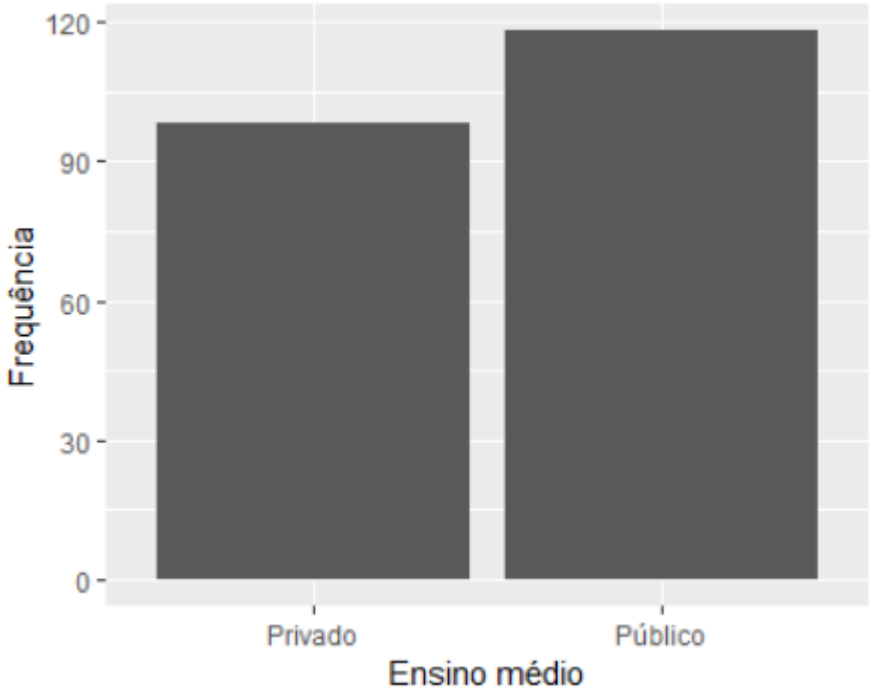
Tabela 4 - Proporção da variável cor

Amarela	Branca	Não informada	Parda	Preta
4,17%	79,17%	2,78%	10,64%	3,24%

Fonte: Elaborada pela autora

A variável ensino médio diz respeito a formação majoritária ou total em escola pública ou privada. Dadas a Figura 6 e a Tabela 5, observamos que, apesar de próximas, a maioria dos ingressantes fizeram a maior parte ou todo o ensino médio em escola pública.

Figura 6 - Gráfico de barras para a variável ensinomedio



Fonte: Elaborada pela autora

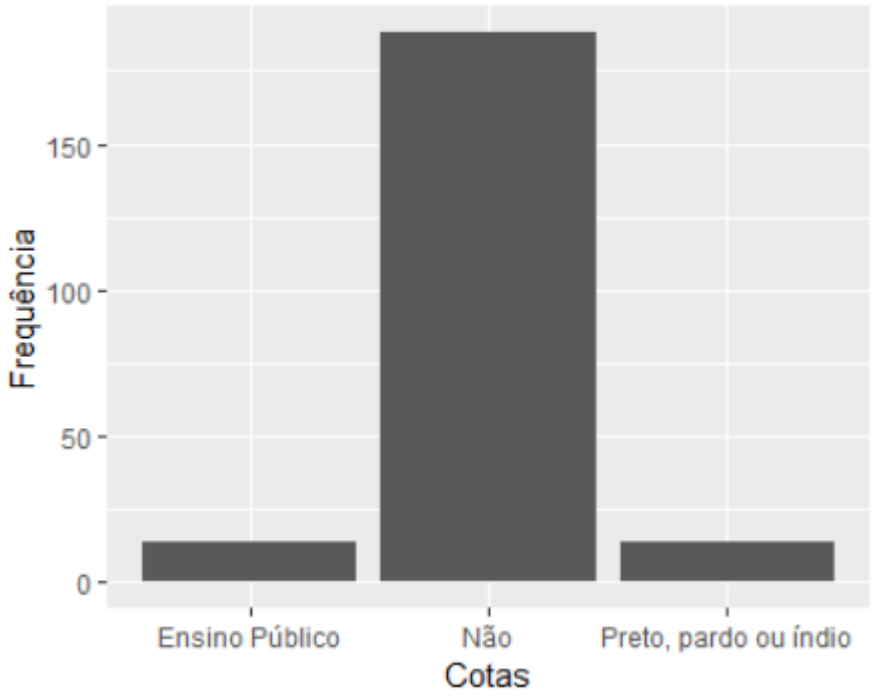
Tabela 5 - Proporção da variável ensinomedio

Privado	Público
45,37%	54,63%

Fonte: Elaborada pela autora

A partir de 2016 a UNESP passou a reservar 50% das vagas para o sistema de cotas, antes eram apenas 15% das vagas. Porém, como estamos analisando dados de 2008 a 2017, observa-se na Figura 7 e Tabela 6 que há um predomínio de não cotistas, enquanto ingressantes pelo sistema de cotas não somam nem 13%.

Figura 7 - Gráfico de barras para a variável cotas



Fonte: Elaborada pela autora

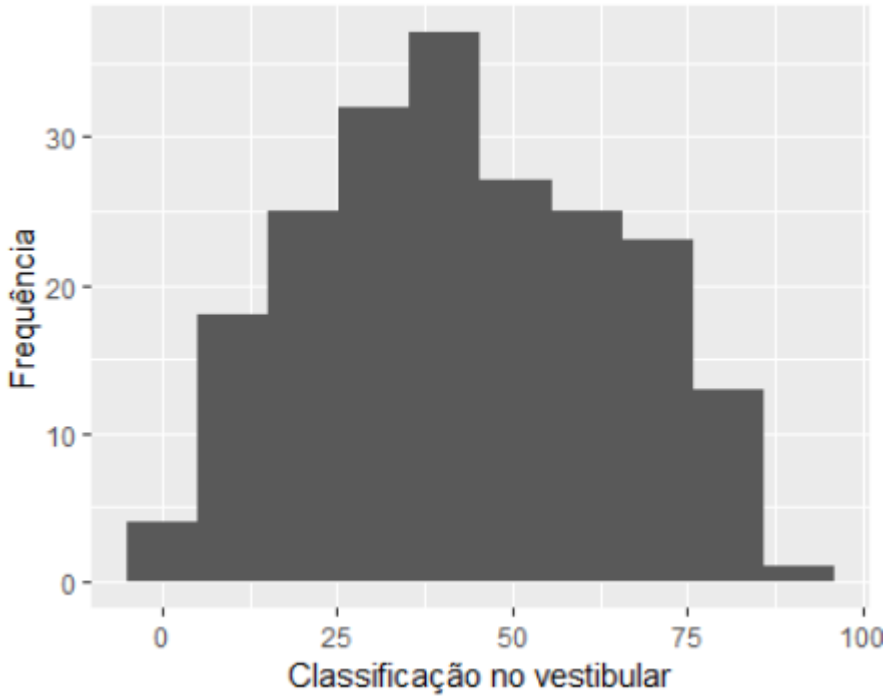
Tabela 6 - Proporção da variável cotas

Ensino Público	Não	Preto, pardo ou índio
6,48%	87,04%	6,48%

Fonte: Elaborada pela autora

Graficamente, pela Figura 8, a classificação no vestibular dos alunos aparenta seguir distribuição normal, porém ao fazer o teste de normalidade de Shapiro-Wilk obteve-se p-valor = 0,01145, portanto a variável não segue distribuição normal. Essa variável contém 11 valores missings, portanto todas as medidas apresentadas na Tabela 7 foram calculadas retirando os valores faltantes.

Figura 8 - Histograma para a variável classvest



Fonte: Elaborada pela autora

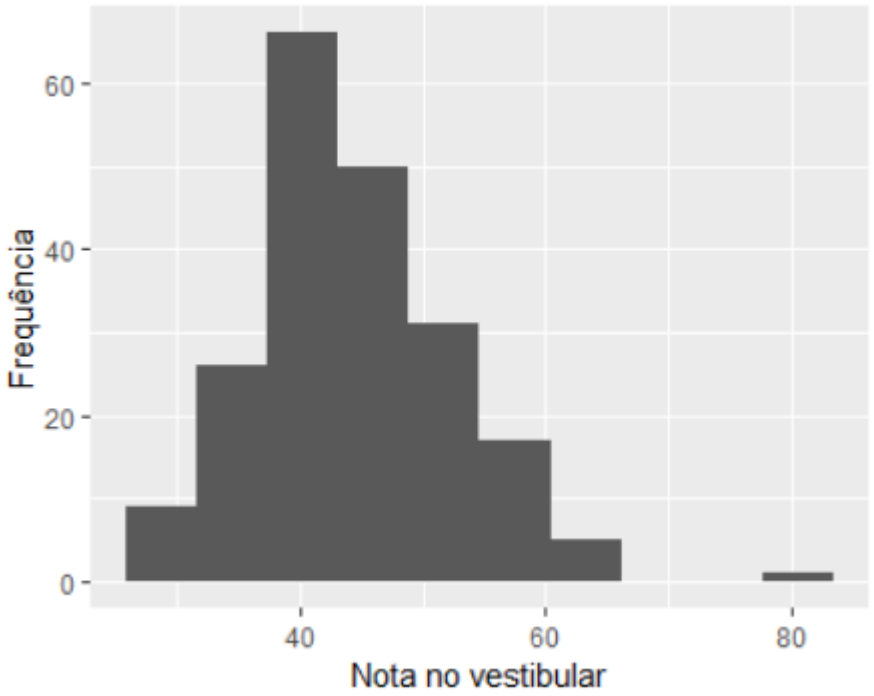
Tabela 7 - Medidas resumo da variável classvest

Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
1	28	43	58	92	43,35	21,08

Fonte: Elaborada pela autora

Assim como a classificação no vestibular, o histograma apresentado na Figura 9 da nota do aluno no vestibular também parece seguir distribuição normal, mas ao testar a normalidade pelo teste de Shapiro-Wilk obteve-se p-valor = 0,01555, portanto a variável não segue distribuição normal. Essa variável contém 11 valores missings, portanto todas as medidas da Tabela 8 foram calculadas retirando os valores faltantes.

Figura 9 - Histograma para a variável nota no vestibular



Fonte: Elaborada pela autora

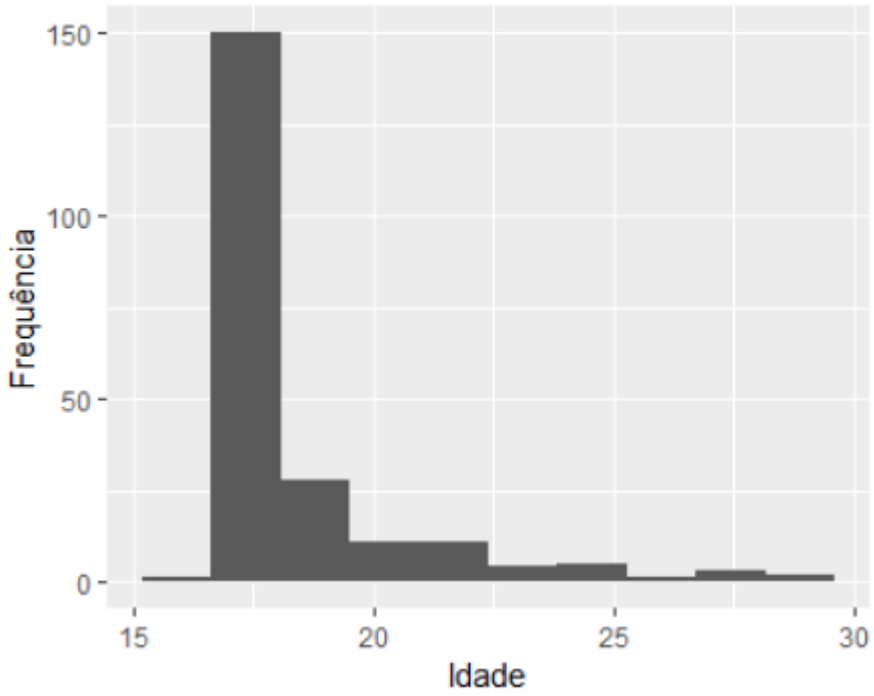
Tabela 8 - Medidas resumo da variável notavest

Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
26	39	43,83	49,18	77,76	44,45	8,15

Fonte: Elaborada pela autora

Nota-se, pela Figura 10 e Tabela 9, que 25% dos alunos ingressam no curso de Estatística na FCT/UNESP com até 17 anos e 75% com até 19 anos. No intervalo de tempo analisado, a idade máxima foi de 29 anos.

Figura 10 - Histograma para a variável idade



Fonte: Elaborada pela autora

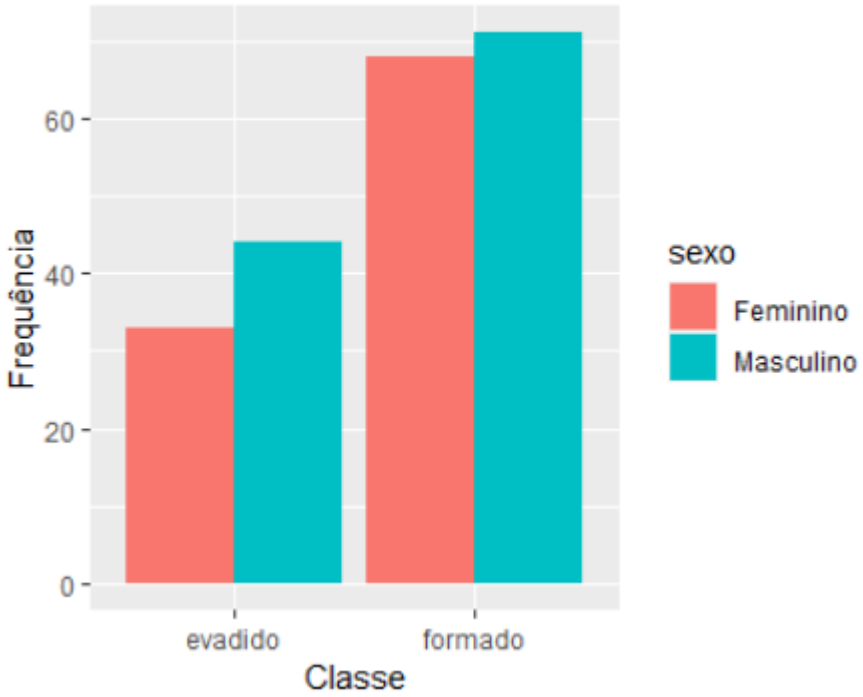
Tabela 9 - Medidas resumo da variável idade

Mínimo	1º Quartil	Mediana	3º Quartil	Máximo	Média	Desvio Padrão
16	17	18	19	29	18,58	2,27

Fonte: Elaborada pela autora

Observa-se na Figura 11 que, entre os evadidos, a proporção de pessoas do sexo masculino é superior, assim como entre os formados, porém com uma diferença menor.

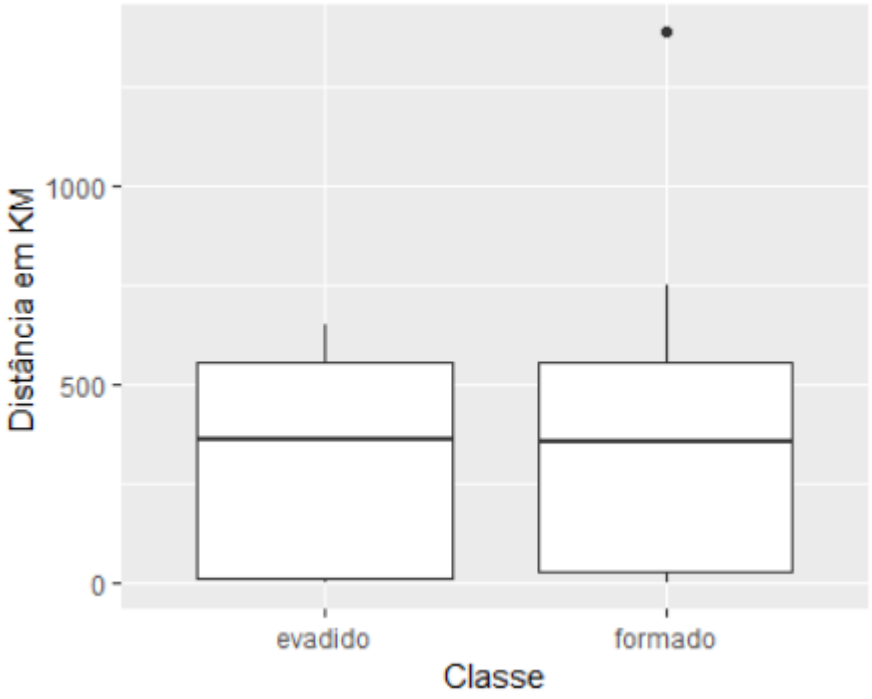
Figura 11 - Gráfico de barras para a variável sexo por classe



Fonte: Elaborada pela autora

É possível notar na Figura 12 que o boxplot para a classe dos formados, embora parecido com o dos evadidos, contém São Franscisco – MG, que está a 1.386 quilômetros de Presidente Prudente – SP, como outlier e seu limite superior é levemente maior, indicando que os alunos que vêm de mais de 700km de Presidente Prudente - SP aparentemente tendem a se formar.

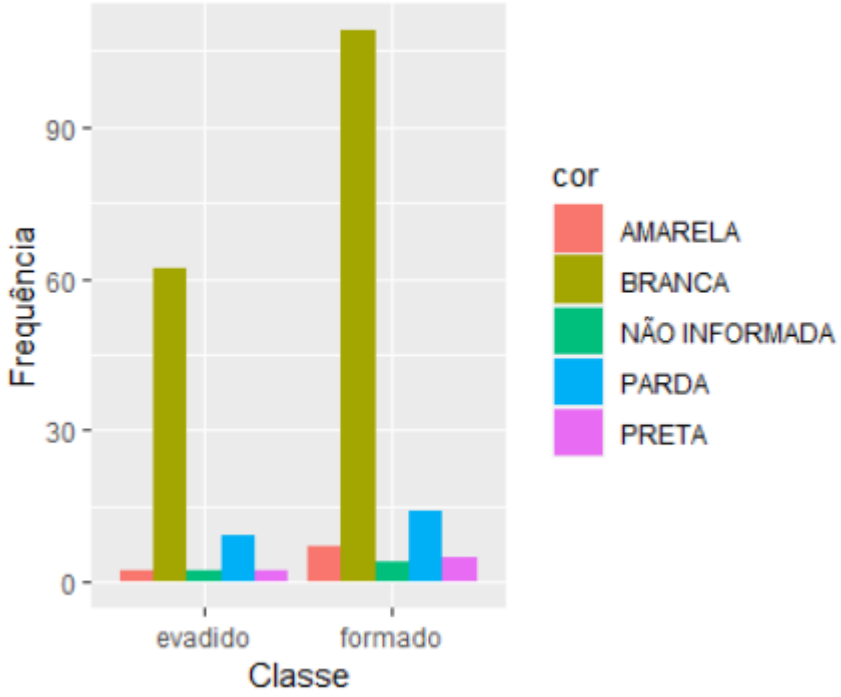
Figura 12 - Boxplot para a variável distKM



Fonte: Elaborada pela autora

Observa-se, na Figura 13, que todas as categorias de cor são maiores na classe de formados do que a de evadidos.

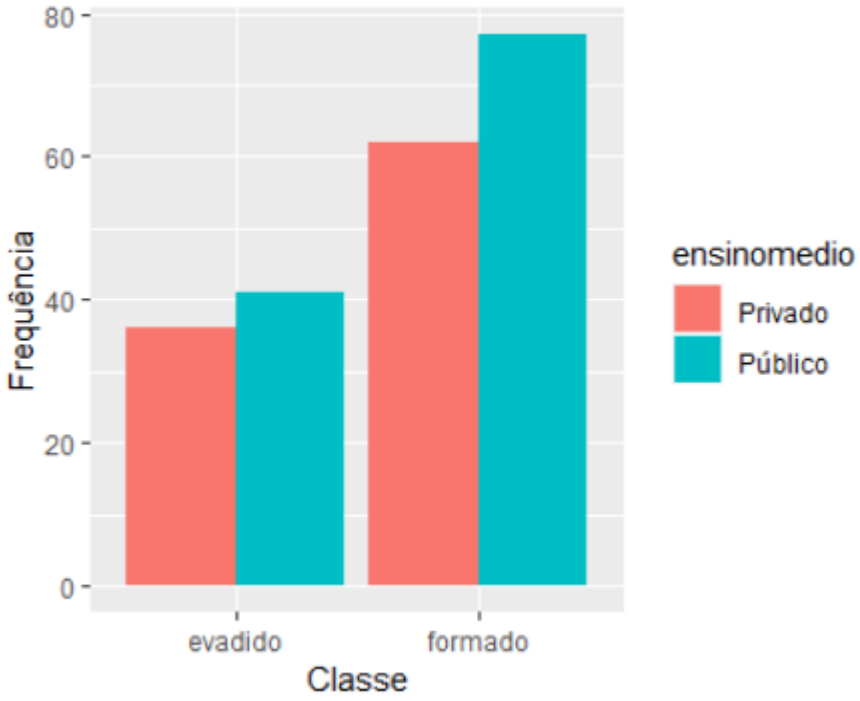
Figura 13 - Gráfico de barras para a variável cor por classe



Fonte: Elaborada pela autora

Observa-se, através da Figura 14, que entre os formados, a proporção de alunos que concluíram todo ou maior parte do ensino médio no ensino público é superior, assim como entre os evadidos, porém com uma diferença menor entre os que concluíram no ensino privado.

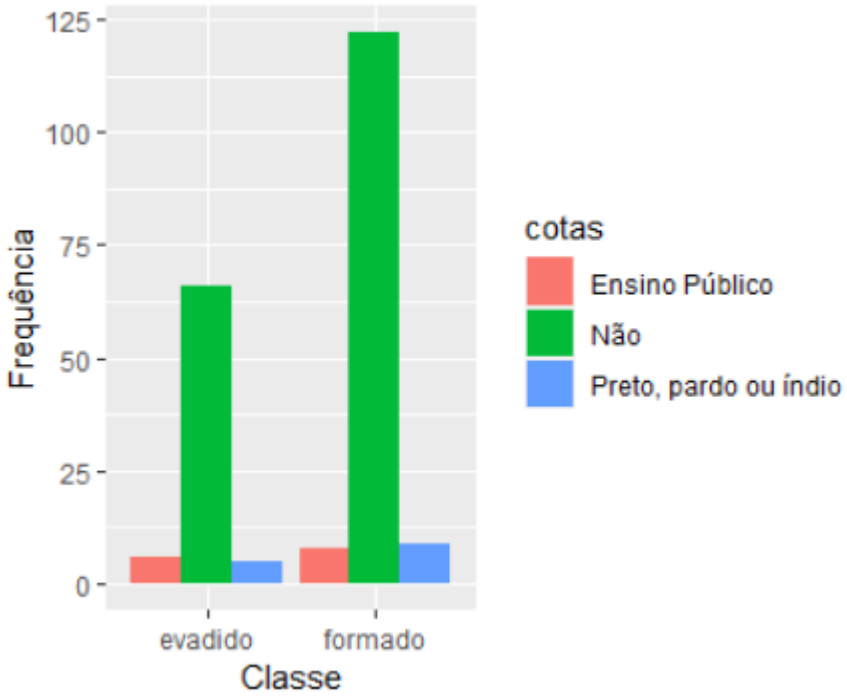
Figura 14 - Gráfico de barras para a variável ensino médio por classe



Fonte: Elaborada pela autora

Na Figura 15 nota-se que, a proporção de cotas entre as classes parece ser igualmente distribuída.

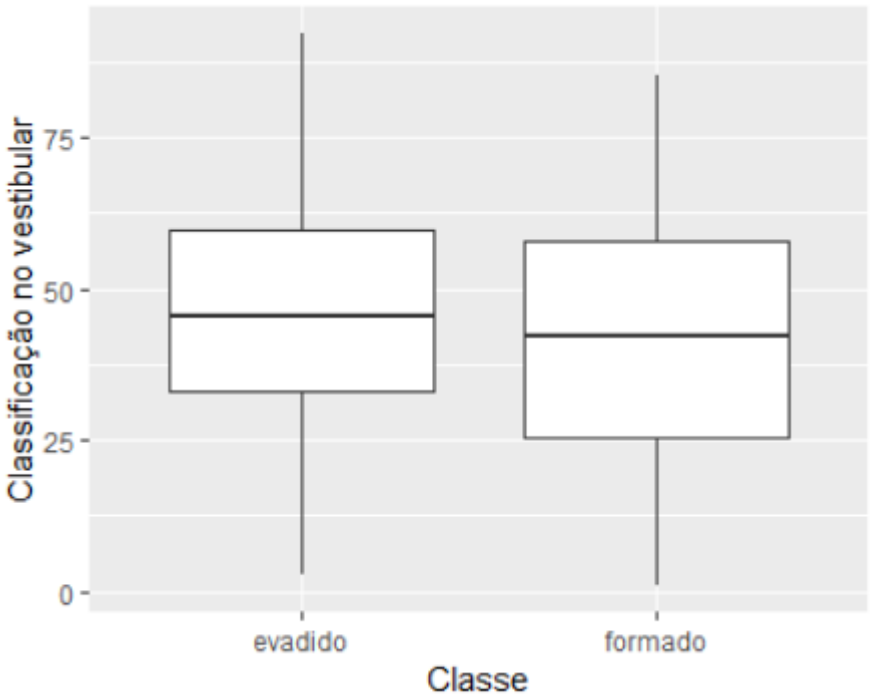
Figura 15 - Gráfico de barras para a variável cotas por classe



Fonte: Elaborada pela autora

Através da Figura 16, observa-se que os alunos que evadem têm uma mediana e amplitude de classificação no vestibular maior que os alunos que se formam.

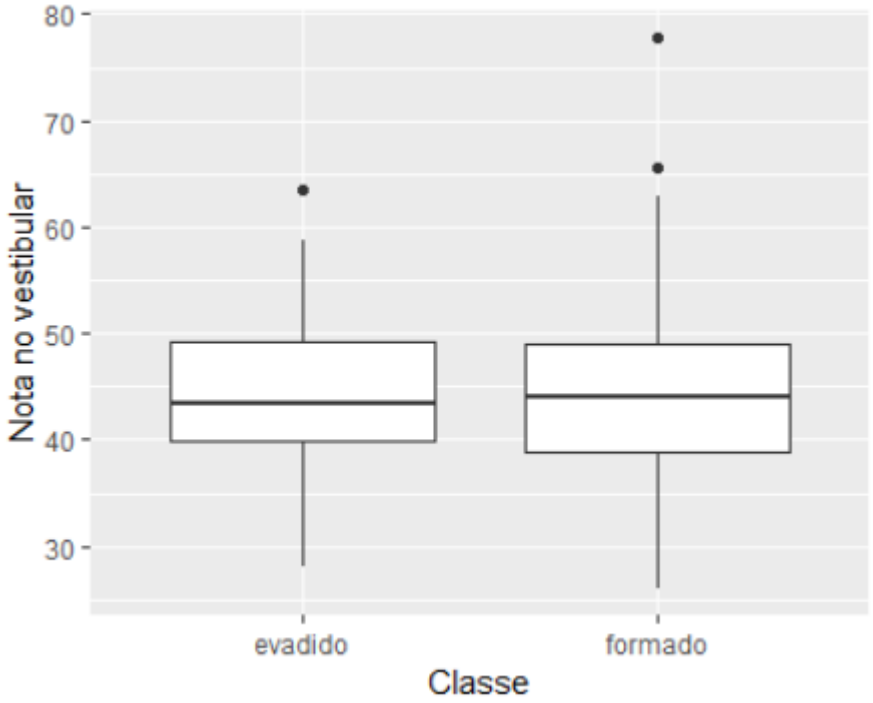
Figura 16 - Boxplot para a variável classificação no vestibular por classe



Fonte: Elaborada pela autora

Observamos na Figura 17 que as duas classes possuem outliers na variável nota do vestibular, porém, a maior nota está nos formados, e diferentemente do gráfico anterior, relacionado a classificação no vestibular, a classe dos formados apresenta maior mediana e amplitude.

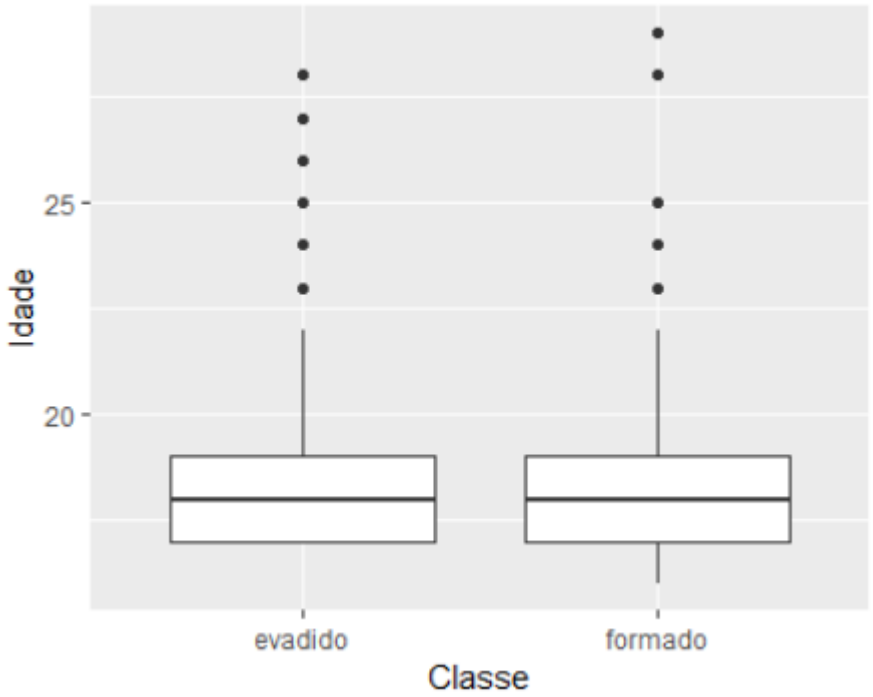
Figura 17 - Boxplot para a variável nota no vestibular por classe



Fonte: Elaborada pela autora

A partir da Figura 18 nota-se que as duas classes possuem outliers na variável idade, suas medianas aparentam ser a mesma, e a maior idade de ingresso pertence a classe dos formados.

Figura 18 - Boxplot para a variável idade por classe



Fonte: Elaborada pela autora

3.2 Análise

Inicialmente, foi realizada a preparação dos dados para a aplicação do modelo de regressão logística. Para isso foi executada a imputação da média nas 11 observações faltantes da variável classvest, já que esse modelo não funciona com missings. Além disso, foram criadas categorias para as variáveis numéricas distKM, classvest, e idade e para a variável cidadeorigem. A variável cidadeorigem se tornou “residente” se a cidade de origem do aluno era Presidente Prudente – SP e “migrante” se não, e as demais variáveis ficaram como apresentado na Tabela 10:

Tabela 10 - Categorização das variáveis

Variável	Quebras	Classes
distKM	0 |- 100 100 |- 600 600 |- 1500	“regiao” ”distante” ”muito_dist”
classvest	0 |- 29 29 |- 58 58 |- 92	“bem” ”regular” ”mal”
idade	0 |- 19 19 |- 24 24 |- 50	“adol_jovem” ”jovem_adulto” ”adulto”

Fonte: Elaborada pela autora

Também foram criadas variáveis dummies para as variáveis categóricas. As variáveis dummies são novas variáveis com o número de categorias menos um, onde cada observação terá valor “um” ou “zero”, caso tenha ou não o atributo, respectivamente, transformando assim, uma variável categórica em numérica. A categoria que não será colocada na criação das dummies é chamada de categoria de referência, isso porque os resultados das demais serão feitos em relação a ela. Para a execução do modelo de regressão logística, foram utilizadas as seguintes categorias como referência:

- sexo: “Masculino”
- cor: “Branca”
- ensinomedio: “Privado”
- cotas: “Não”
- cidadeorigem: “Migrante”

- distKM: “muito_dist”
- classvest: “mal”
- idade: “adulto”

Para aplicação do modelo, a base de dados foi dividida em 70% para treinamento, e 30% para teste, tendo, portanto, 151 e 65 observações, respectivamente. Então foi criado o modelo de regressão logística a partir da base de treino, o resultado do modelo logístico completo pode ser visto na Tabela 11, onde é possível observar que incialmente, pelo teste de Wald, nenhuma variável é significativa para o modelo.

Tabela 11 - Resultados para o modelo logístico completo

Variável	Estimativa	Erro Padrão	P-valor do teste Wald
(Intercept)	0,53901	1,29011	0,676
sexoFeminino	0,06858	0,37137	0,853
corAmarela	-2,05187	1,30020	0,115
corNaoInformada	-0,73634	1,23577	0,551
corParda	0,07217	0,78967	0,927
corPreta	-1,04304	1,48418	0,482
ensinoPublico	-0,11227	0,39976	0,779
cotasPublico	0,37290	0,75237	0,620
cotasPPI	0,49583	1,04480	0,635
cidadeResidente	0,09544	0,62476	0,879
distRegiao	0,35651	0,91253	0,696
distDistante	0,11598	0,78527	0,883
classvestRegular	0,41597	0,45607	0,362
classvestBem	-0,84457	0,56327	0,134
idadeAdoljovem	-1,39327	1,01540	0,170
idadeJovemadulto	-1,11911	1,10011	0,309

Fonte: Elaborada pela autora

Existem algumas formas de dar prosseguimento a essa análise, e o método escolhido foi o stepwise, que é um método de seleção de variáveis que faz vários passos adicionando e retirando variáveis, a fim de, obter o melhor ajuste. Após a aplicação do stepwise, obteve-se os resultados presentes na Tabela 12.

Tabela 12 - Resultados para o modelo logístico pelo método stepwise

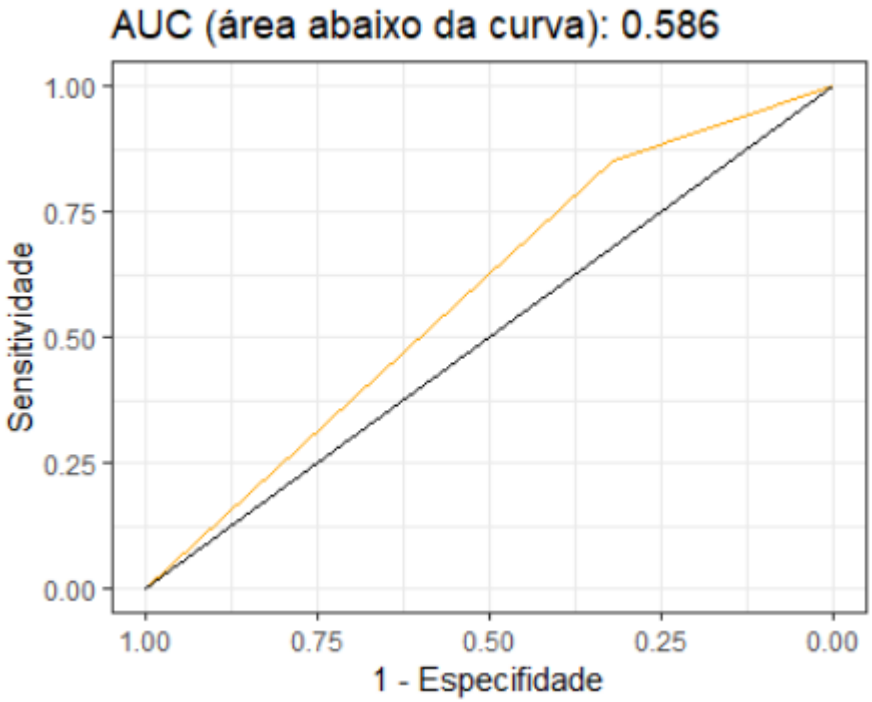
Variável	Estimativa	Erro Padrão	LI	LS	P-valor do teste Z-Wald
(Intercept)	-0,3610	0,1921	-0,7375	-0,0155	0,0602
classvestBem	-0,9935	0,4406	-1,8571	-0,1299	0,0241

Fonte: Elaborada pela autora

Após a aplicação do stepwise, a variável classvestBem foi a única significativa, e que, juntamente ao intercepto, compõe a equação final. A deviance residual do modelo foi 191,25 com 149 graus de liberdade. A interpretação das variáveis do modelo é feita a partir das odds ratio, que para a variável classvestBem foi 0,3703, ou seja, a chance de um aluno que ficou bem classificado no vestibular (ficou entre os 29 primeiros colocados) evadir é 62,97% menor que a chance de alguém que ficou mal classificado (abaixo da 58º colocação) evadir.

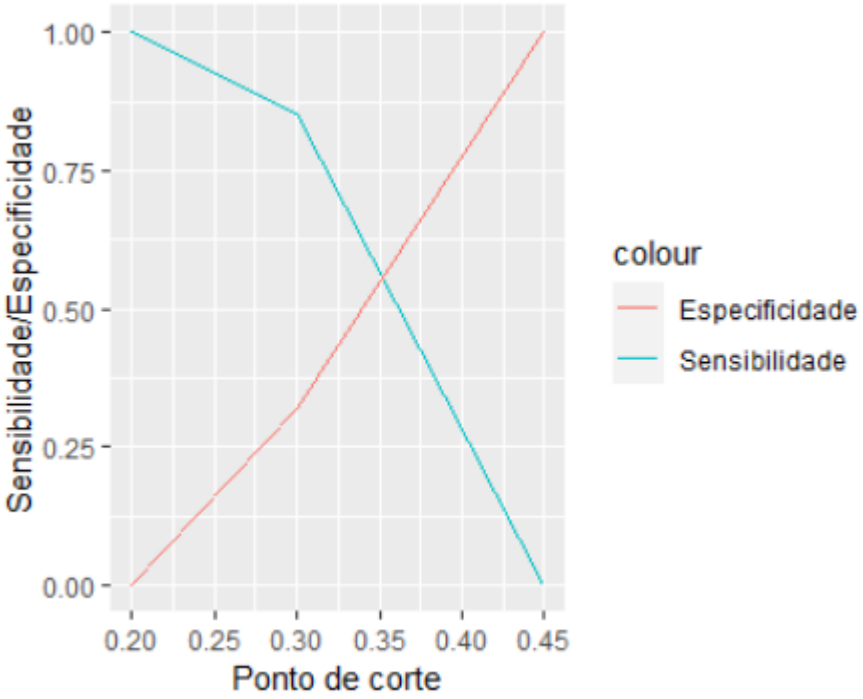
Para encontrar o ponto de corte ótimo, existe a curva ROC, que auxilia a determiná-lo e, além disso, a área sobre essa curva (AUC) varia de 0,5 a 1 e indica se o modelo teve uma boa discriminação, onde, quanto mais próximo de 1 melhor, e mais próximo de 0,5 pior. Também existe um gráfico que plota os resultados da sensibilidade e da especificidade, o ponto onde esses resultados se cruzam é considerado o ponto ótimo. Fazendo a curva ROC, Figura 19, tem-se a área sobre a curva do modelo dada por 0,586, como esse resultado está próximo do 0,5, podemos afirmar que o modelo não está fazendo muito bem a discriminação de quem evadiu ou não do curso. Pela Figura 20, observamos que o ponto ótimo é 0,36. Como o resultado da regressão logística é uma probabilidade, toda vez que o modelo der um resultado igual ou maior que 0,36, será previsto como aluno evadido.

Figura 19 - Curva ROC e AUC do modelo stepwise



Fonte: Elaborada pela autora

Figura 20 - Gráfico para encontrar o ponto ótimo



Fonte: Elaborada pela autora

Após a aplicação do método stepwise, o modelo foi aplicado na base de treino utilizando o ponto ótimo, com o intuito de ver como ele está performando para os dados no qual ele foi treinado, onde obteve-se as métricas presentes na Tabela 13 abaixo.

Tabela 13 - Métricas de qualidade de ajuste do modelo logístico

Acurácia	Precisão	Recall	F1-score
0,51	0,41	0,85	0,55

Fonte: Elaborada pela autora

Como as classes da variável de interesse não estão balanceadas, a acurácia não é uma medida tão interessante de ser analisada nesse caso, mas seu valor indica que o modelo acerta se o aluno irá evadir ou não 51% das vezes. Dentre todas as medidas, o recall, que diz de todos os alunos que evadiram qual a proporção que o modelo previu corretamente, foi a métrica mais alta, com 85%, enquanto a precisão foi a mais baixa, indicando que de todos que o modelo previu como evadidos, ele acertou apenas 41%. O F1-score é uma única medida que sintetiza a precisão e o recall, e seu valor foi de 55%.

A fim de testar o poder de previsão para indivíduos que não estavam contidos na base de treinamento, o modelo foi aplicado na base de teste, onde obteve-se as seguintes métricas vistas a seguir na Tabela 14.

Tabela 14 - Métricas de qualidade de ajuste do modelo logístico na base de teste

Acurácia	Precisão	Recall	F1-score
0,45	0,37	0,78	0,5

Fonte: Elaborada pela autora

Nota-se que todas as métricas sofreram uma diminuição de performance em uma base desconhecida, o recall diminuiu 7%, enquanto as demais diminuíram cinco pontos percentuais.

A Tabela 15 apresenta um exemplo para ilustrar a classificação do modelo logístico na prática. Tem-se o aluno 1, que ficou classificado na 77ª posição, sua estimativa foi de 0,4107, como está acima do ponto de corte, sua vaga será considerada como evasão. O aluno 2 ficou classificado em 23º lugar, sua estimativa

pelo modelo foi de 0,2051, isso é, abaixo do ponto de corte, portanto será considerado como não evasão.

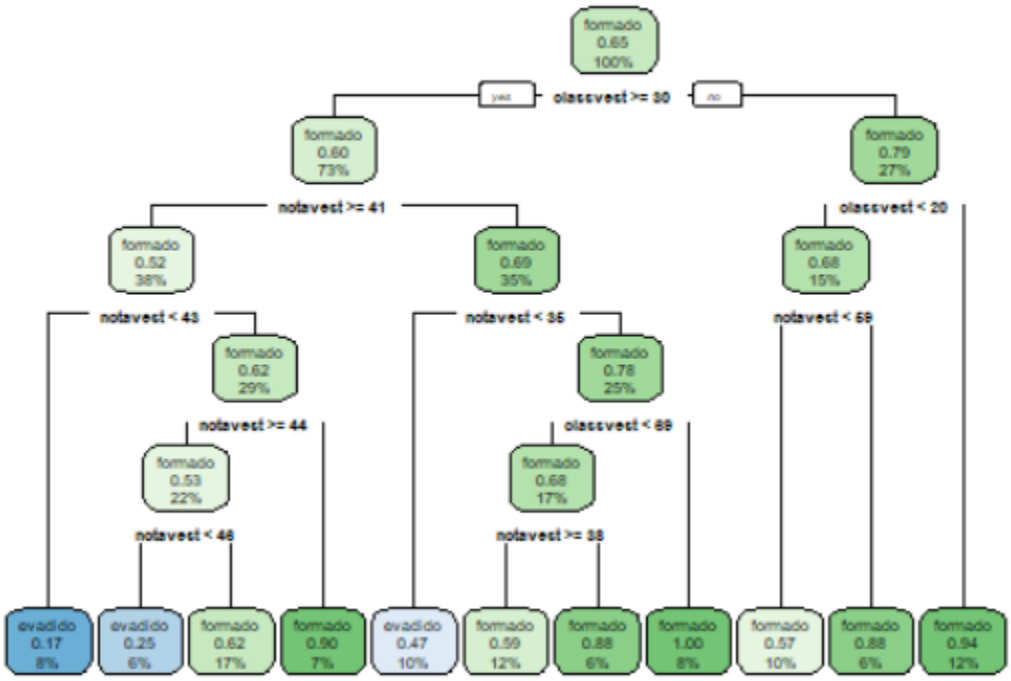
Tabela 15 - Classificação do modelo logístico

Aluno	classvest	Estimativa	Ponto de corte (0,36)
1	77	0,4107	Evasão
2	23	0,2051	Não evasão

Fonte: Elaborada pela autora

A partir do resultado obtido no modelo logístico, foi construída uma árvore de decisão usando a variável classvest que foi significativa, e a variável notavest que para a construção do modelo logístico foi retirada, por ser altamente correlacionada negativamente com a classvest. A árvore de decisão foi feita a fim de ver as quebras que seriam realizadas, tendo o resultado apresentado na Figura 21 a seguir.

Figura 21 - Árvore de decisão para as variáveis classvest e notavest



Fonte: Elaborada pela autora

É possível distinguir a proporção de formados nas folhas pela cor, as verdes indicam maior proporção de formados, e a azul de evadidos, e quanto mais forte a cor, maior as respectivas proporções. Identifica-se que há três regras de decisão baseadas

na classificação do vestibular, enquanto as demais são feitas baseadas na nota que o aluno tirou no vestibular. Assim, as quebras propostas pelo modelo para as variáveis são as presente na Tabela 16.

Tabela 16 - Quebras feitas pela árvore de decisão para as variáveis classvest e notavest

Variável	Quebras
classvest	0 |- 20 20 |- 30 30 |- 69 69 |- +∞
notavest	0 |- 35 35 |- 38 38 |- 41 41 |- 43 43 |- 44 44 |- 46 46 |- 59 59 |- 100

Fonte: Elaborada pela autora

Fazendo então as métricas de qualidade de ajuste para a árvore de decisão temos a Tabela17.

Tabela 17 - Métricas de qualidade de ajuste da árvore de decisão para as variáveis classvest e notavest

Acurácia	Precisão	Recall	F1-score
0,73	0,69	0,44	0,54

Fonte: Elaborada pela autora

Ao comparar com as mesmas métricas de ajuste do modelo logística, notamos que houve uma melhora de 22% na acurácia e 28% na precisão, enquanto o recall diminuiu 41% e o F1-score 1%. Concluímos que, embora esse modelo acerte menos a proporção real de evadidos, ele classifica menos alunos de forma errônea como evadidos, quando são formados. Para a base de teste, obteve-se a Tabela 18.

Tabela 18 - Métricas de qualidade de ajuste da árvore de decisão para as variáveis classvest e notavest na base de teste

Acurácia	Precisão	Recall	F1-score
0,54	0,23	0,13	0,17

Fonte: Elaborada pela autora

Por ser um modelo mais específico, onde as notas são quebradas mais vezes, ele performa muito bem para a base em que foi criado, mas ao aplica-lo em uma base desconhecida sua performance cai bastante, notamos pelo F1-score de 17%.

A árvore de decisão também foi aplicada para os dados sem categorização (variáveis apresentadas no Quadro 2), para analisar quais quebras seriam feitas. Foi utilizado a mesma divisão de treino e teste. Os resultados obtidos são os apresentados a seguir na Tabela 19.

Tabela 19 - Métricas de qualidade de ajuste da árvore de decisão para a base sem categorização

Acurácia	Precisão	Recall	F1-score
0,87	0,84	0,78	0,81

Fonte: Elaborada pela autora

E aplicando na base de teste, teve-se os resultados apresentados na Tabela 20 a seguir.

Tabela 20 - Métricas de qualidade de ajuste da árvore de decisão para a base sem categorização na base de teste

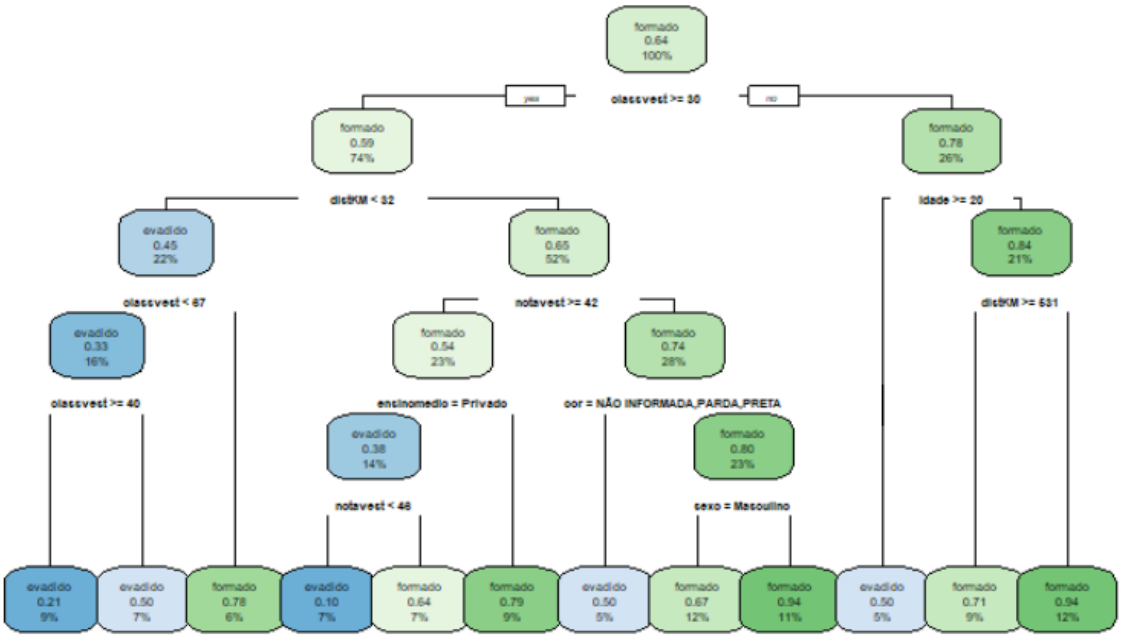
Acurácia	Precisão	Recall	F1-score
0,54	0,27	0,17	0,21

Fonte: Elaborada pela autora

Apesar de um ótimo resultado no ajuste, a visualização dessa árvore ficou muito precária, por conta da variável cidadeorigem, que contém 67 cidades distintas. Pelo alto número de classes na variável cidadeorigem, acredita-se que o ajuste na base de teste não tenha ficado tão bom por essa ser uma variável muito específica, e aparecerem 20 novas cidades. Embora bem ajustado, esse modelo, assim como a árvore com duas variáveis, não ficou bom por ficar específico demais para a base de treino.

Como a árvore com os dados originais ficou de difícil visualização pela variável cidadeorigem, ela foi retirada da base e foi feito uma nova árvore de decisão, como pode ser visto na Figura 22.

Figura 22 - Árvore de decisão para a base sem categorização e sem a variável cidadeorigem



Fonte: Elaborada pela autora

As quebras feitas foram baseadas nas variáveis classvest, idade, distKM, notavest, ensinomedio, sexo e cor. Calculando as métricas a partir da base de treino, tem-se a Tabela 21.

Tabela 21 - Métricas de qualidade de ajuste da árvore de decisão para a base sem categorização e sem a variável cidadeorigem

Acurácia	Precisão	Recall	F1-score
0,75	0,66	0,61	0,64

Fonte: Elaborada pela autora

O modelo apresenta métricas de qualidade de ajuste razoavelmente boas, onde todas as estão acima de 60% e o F1-score é de 64%. Aplicando a base de teste nas regras de decisões criadas, obteve-se a Tabela 22.

Tabela 22 - Métricas de qualidade de ajuste da árvore de decisão para a base sem categorização e sem a variável cidadeorigem na base de teste

Acurácia	Precisão	Recall	F1-score
0,51	0,20	0,13	0,16

Fonte: Elaborada pela autora

Ao aplicado na base de teste, a performance do modelo tem uma queda muito brusca, com F1-score de 16%, sendo o menor de todos os modelos.

3.3 Comparação dos Modelos

Comparando as métricas de qualidade de ajuste dos modelos para a base de treino, através da Tabela 23, onde RL é o modelo de regressão logística, AD1 é a árvore de decisão para as variáveis classvest e notavest e AD2 é a árvore de decisão para a base sem categorização e sem a variável cidadeorigem e desconsiderando a árvore para a base sem categorização, tem-se que o modelo de regressão logística contém o melhor resultado de recall, com 85%, a árvore de decisão para as variáveis classvest e notavest traz a melhor precisão, com 69% e a árvore de decisão para a base sem categorização e sem a variável cidadeorigem contendo a melhor acurácia e F1-score, com 75% e 64% respectivamente.

Tabela 23 - Métricas de qualidade de ajuste dos modelos

	Acurácia	Precisão	Recall	F1-score
RL	0,51	0,41	0,85	0,55
AD1	0,73	0,69	0,44	0,54
AD2	0,75	0,66	0,51	0,64

Comparando as métricas de qualidade de ajuste dos modelos para a base de teste, através da Tabela 24, onde RL é o modelo de regressão logística, AD1 é a árvore de decisão para as variáveis classvest e notavest e AD2 é a árvore de decisão para a base sem categorização e sem a variável cidadeorigem e desconsiderando a árvore para a base sem categorização, tem-se a árvore de decisão para as variáveis classvest e notavest trazendo a melhor acurácia, com 69%, já o modelo de regressão logística contém a melhor precisão, recall e F1-score, com 37%, 78% e 50% respectivamente.

Tabela 24 - Métricas de qualidade de ajuste dos modelos na base de teste

	Acurácia	Precisão	Recall	F1-score
RL	0,45	0,37	0,78	0,50
AD1	0,54	0,23	0,13	0,17
AD2	0,51	0,20	0,13	0,16

4 CONCLUSÃO

Comparando os resultados obtidos após a aplicação das duas técnicas pelo F1-score, o modelo de árvore para a base sem categorização e sem a variável cidadeorigem desempenhou um bom ajuste para a base de treino, porém apresentou dificuldade de previsão para a base de teste, possivelmente por ficar específico demais, entrando em um estado de overfitting, onde o modelo se ajusta com excelência a base em que foi treinado, porém ao ser colocado em produção, em bases de dados desconhecidas, seu poder de previsão é fraco.

Ainda levando o F1-score como métrica de comparação, a regressão logística foi a técnica que obteve melhor performance de previsão para os dados de evasão do curso de Estatística na FCT/UNESP de Presidente Prudente - SP. Porém, ao ser aplicado, esse modelo não se mostra tão preciso, ou seja, ele consegue captar quase 80% dos alunos que irão evadir, porém classifica a maioria dos alunos como propensos a evadir, classificando assim, muitos alunos que não irão evadir como evasão.

Além disso, é importante salientar que existem diversos fatores, como problemas de saúde, problemas familiares, problemas financeiros, etc., que levam um estudante a evadir que não foram levados em consideração para a realização desse estudo.

Referências

CARDOSO, C. B. **Efeitos da Política de Cotas na Universidade e Brasília:** Uma análise do rendimento e da evasão. Dissertação de Mestrado em Educação, Universidade de Brasília, Brasília, 2008.

DAVOK, D. F., BERNARD, R. P. Avaliação dos índices de evasão nos cursos de graduação da Universidade do Estado de Santa Catarina – UDESC. **Avaliação, Campinas; Sorocaba, SP**, v. 21, n. 2, p. 503-521, jul. 2016. DOI: http://dx.doi.org/10.1590/S1414-40772016000200010. Disponível em: https://www.scielo.br/j/aval/a/5VJRg7PrXDTQ5mYXK95rh8r/abstract/?lang=pt#:~:text=A%20pesquisa%20teve%20o%20objetivo,de%20Santa%20Catarina%20(UDESC).&text=O%20%C3%ADndice%20m%C3%A9dio%20de%20evas%C3%A3o,foi%20de%2038%2C2%25. Acesso em: 18 jan. 2022.

HOSMER, D. W., LEMESHOW, S. **Applied Logistic Regression**, New York: Wiley, 3a edição, 2013.

KLEINBAUM, D. G., MITCHEL, K. **Logistic Regression:** a self-learning text, 2a. ed., New York: Springer, 2006.

LAURETTO, M. S. **Árvores de classificação para escolha de estratégias de operação em mercados de capitais**. 1996. Dissertação (Mestrado em Matemática Aplicada) – Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1996.

PINTO, P. S. Universidades federais têm evasão de 15% em 2018. **Poder 360**, 8 out. 2019. Disponível em: https://www.poder360.com.br/governo/universidades-federais-tem-evasao-de-15-em-2018/. Acesso em: 19 jan. 2022.

R CORE TEAM. **R:** A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. Disponível em: https://www.r-project.org/. Acesso em: 23 jan. 2022.

SILVA, L. M. O. **Uma Aplicação de Árvores de Decisão, Redes Neurais e KNN para a Identificação de Modelos ARMA Não-Sazonais e Sazonais**. 2005. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2005.

SIMÕES, A. C. A. **Mineração de dados baseada em árvores de decisão para análise do perfil de contribuintes**. 2008. Dissertação (Mestrado em Ciência da Computação) - Universidade Federal de Pernambuco, Recife, 2008. Disponível em: https://repositorio.ufpe.br/handle/123456789/1476. Acesso em: 13 jul. 2022.

TACHIBANA, V. M. **Notas de aula de aula de Regressão Logística**. 2020. Presidente Prudente: [s.n.], 2020. Apostila disponível no ambiente virtual de apoio disciplina Regressão Logística da FCT/Unesp.