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

GABRIELA FABRINI MEDINA

EVASÃO DE ALUNOS DE ESTATÍSTICA DA UNESP DE PRESIDENTE PRUDENTE: Aplicação de Análise Discriminante e Regressão Logística.

PRESIDENTE PRUDENTE
2023

GABRIELA FABRINI MEDINA

EVASÃO DE ALUNOS DE ESTATÍSTICA DA UNESP DE PRESIDENTE PRUDENTE: Aplicação de Análise Discriminante e Regressão Logística.

Relatório Final do Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso II.
Orientadora: Profa. Dra. Miriam Rodrigues Silvestre.

PRESIDENTE PRUDENTE
2023

M491e

Medina, Gabriela Fabrini
Evasão de alunos de estatística da UNESP de Presidente Prudente : aplicação de análise discriminante e regressão logística / Gabriela Fabrini Medina. -- Presidente Prudente, 2023
45 f. : il., tabs., fotos

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

1. Análise de regressão logística. 2. Análise discriminante. 3. Evasão universitária. 4. Estatística. 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

Gabriela Fabrini Medina

EVASÃO DE ALUNOS DE ESTATÍSTICA DA UNESP DE PRESIDENTE PRUDENTE: Aplicação de Análise Discriminante e Regressão Logística.

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:

Orientadora: _____

Profª. Drª. Miriam Rodrigues Silvestre
Departamento de Estatística

_____

Prof. Dr. Manoel Ivanildo Silvestre Bezerra
Departamento de Estatística

_____

Prof. Dr. Klaus Schlüzen Junior
Departamento de Estatística

Presidente Prudente, 08 de fevereiro de 2023.

RESUMO

O desenvolvimento desse estudo foi realizado utilizando dados de evasão de alunos de estatística da UNESP de Presidente Prudente, tendo como objetivo comparar duas técnicas de modelagem estatística, Regressão Logística e Análise Discriminante. A primeira demonstra uma flexibilidade na aplicação de diferentes tipos de informações existentes em um conjunto de dados. A segunda apresenta algumas restrições no formato de variáveis e outros pressupostos. Como resultado, foi concluído uma melhor performance pela técnica de Regressão Logística para esse conjunto de dados, apesar da previsão do modelo não ser tão eficaz, foi considerado como variáveis significativas para o modelo as informações de sexo do aluno, se ele utiliza de cotas, o ano da primeira matrícula e o ano de nascimento/ idade do estudante.

Palavras-chave: Análise de regressão logística; Análise discriminante; Evasão universitária; Estatística.

ABSTRACT

The development of this study was carried out using dropout data from statistics students at UNESP in Presidente Prudente, comparing two statistical modeling techniques, Logistic Regression and Discriminant Analysis. The first demonstrates flexibility in the application of different types of information existing in a dataset. The second presents some restrictions on the format of variables and other assumptions. As a result, a better performance was concluded by the Logistic Regression technique for this data set, despite the model's prediction not being so effective, it was considered as significant variables for the model the student's gender information, if he uses quotas, the year of first enrollment and the year of birth/ age of the student.

Keywords: Logistic regression analysis; Discriminant analysis**;** University dropout; Statistic.

SUMÁRIO

1 INTRODUÇÃO 7

2 REFERENCIAL TEÓRICO 10

3 METODOLOGIA 11

4 FUNDAMENTAÇÃO TEÓRICA 12

4.1 REGRESSÃO LOGÍSTICA 12

4.1.1 Regressão Logística Binária *13*

4.1.2 Regressão Logística Múltipla *16*

4.1.3 Métodos de Avaliação do Modelo *18*

4.2 ANÁLISE DISCRIMINANTE 20

4.2.1 Análise Discriminante Simples *21*

4.2.2 Análise Discriminante de Fisher *24*

4.2.3 Desempenho de Classificação *26*

4.3 TESTES 27

4.3.1 Teste de Normalidade *27*

4.3.2 Teste da Diferença de Médias (Numéricas) *27*

4.3.3 Teste da Diferença de Médias (Categóricas) *28*

5 DESENVOLVIMENTO 29

5.1 DADOS E ANÁLISE DESCRITIVA 29

5.2 TESTES 33

5.3 REGRESSÃO LOGÍSTICA 35

5.4 ANÁLISE DISCRIMINANTE 42

6 CONCLUSÃO 45

REFERÊNCIAS 46

1 Introdução

A evasão no ensino superior é o ato de abandono ou não conclusão de um curso de graduação após ingresso, rompendo o vínculo de aluno e instituição ao não renovar a matrícula. Está presente em qualquer modalidade de educação, sendo ela presencial ou a distância, de oferta pública ou privada, considerada um fenômeno multifatorial.

Ao entender os motivos pelos quais isso ocorre, segundo Gibson (1998), após a realização de um estudo foram identificadas três categorias de fatores que podem explicar o que leva ao abandono: fatores do estudante, situacionais e fatores do sistema educacional. O primeiro inclui a preparação educacional anterior, atributos de motivação, relacionamentos e persistência, bem como a sua autoconfiança acadêmica. Os fatores situacionais são descritos como apoio da família, além de mudanças em circunstâncias da vida pessoal. Enquanto a última engloba tanto a qualidade e as dificuldades com a didática empregada, como com o suporte oferecido pela instituição.

No Brasil considerando a trajetória dos ingressantes em cursos de graduação de 2010 até 2019, a taxa de desistência acumulada é superior a mais da metade dos alunos, chegando ao valor de 59% (INEP, 2019). No total há 302 IES públicas e 2.306 IES privadas (INEP, 2019), ou seja, 11,6% representam o ensino gratuito o qual fornece menos vagas do que na rede privada, no entanto, possui mais concorrência. A evasão nessas poucas vagas existentes gera um desperdício que poderia estar atendendo outro aluno sem condições de pagar pelo ensino, além de ser considerada como um problema social, acadêmico e econômico (FIGUEIREDO E SALLES, 2017), gerando múltiplos impactos. A evasão acadêmica é vista como um investimento sem retorno principalmente nas instituições públicas, considerando o desperdício de materiais e recursos, podendo levar até ao fechamento de cursos caso haja uma taxa muito alta. Ela não representa só uma frustração para o aluno – ela também pode afetar a sua família (por conta de gastos realizados, de sonho de graduação desfeito), para o docente (por não conseguir cumprir seu papel de educador), para a instituição de ensino (pelo não cumprimento de sua missão institucional) e para a própria sociedade (por não poder contar com mais mão-de-obra qualificada).

Assim como diversas outras instituições, a UNESP vem sofrendo com a evasão de alunos, apesar de sua taxa média ser de 6,9%, determinadas áreas acabam tendo taxas de evasão altas, que podem levar a possíveis cortes, como a exclusão de cursos nas faculdades nas quais estão vinculadas. Em 2018, a própria UNESP propôs avaliar a pertinência de manutenção de graduações com baixa procura e alta evasão. A Faculdade de Ciências e Tecnologia do Campus de Presidente Prudente oferece 12 cursos de graduação em licenciatura e bacharelado, sendo um deles o de Estatística. A graduação em Estatística teve várias mudanças desde sua implementação no ano de 1984, principalmente em relação à grade curricular e a quantidade de vagas disponibilizadas. Inicialmente eram 20 e atualmente são 30. Segundo Marin (2018), considerando os matriculados no período de 2000 a 2015, cerca de 42% dos alunos evadiram da graduação, e, desses evadidos, 41,83% ocorreram já no primeiro ano. Portanto, reduzir esse número é um grande desafio de todos os envolvidos com o curso de Estatística da FCT/UNESP, professores, estudantes e gestores.

Objetivo Geral

O principal objetivo deste trabalho é analisar o comportamento do status de matrícula (Formado, Evadido) dos alunos de Estatística na UNESP de Presidente Prudente, identificando os principais fatores que levam a evasão por meio de técnicas de Análise Discriminante e Regressão Logística.

Objetivos Específicos

Os objetivos específicos são listados a seguir:

- Estabelecer um procedimento para classificar os alunos em grupos, com base em seus scores em um conjunto de variáveis independentes;
- Dividir o conjunto de dados dos alunos em partes denominadas conjunto de treinamento e teste.
- Determinar quais das variáveis independentes explicam o máximo de diferenças entre os grupos;
- Analisar se existem diferenças estatisticamente significantes entre os grupos de alunos;
- Fazer um comparativo entre os procedimentos estatísticos de Análise Discriminante e Regressão Logística;

Hipóteses

A taxa de evasão permanece maior no primeiro ano de curso.

A taxa de evasão é maior aos alunos que cursaram o Ensino Médio na rede pública.

A evasão dos alunos cotistas é maior que a evasão de alunos não cotistas.

A nota do vestibular é uma variável significativa nas evasões.

Discentes que vieram de outras cidades evadem mais do que discentes que já residem na cidade.

O modelo de análise discriminante é mais eficiente.

Justificativa

O Curso de Estatística da UNESP possui um número instável de alunos formados por ano, chegando a apenas dois formandos no ano de 2016 enquanto em 2018 concluíram dezesseis. Este resultado é decorrente tanto por reprovações, como evasão de estudantes. Este fato leva a uma constante observação pela universidade e, junto de outros cursos, vem sofrendo ameaças de ser fechado ou até mesmo haver uma junção de graduações.

Sendo assim, com esse estudo, busca-se detectar as possíveis razões que mais influenciam na evasão para que futuramente seja possível tomar ações preventivas para diminuir esses casos e ajudar o curso de Estatística a diminuir a sua taxa de evasão. Por se tratar de uma variável resposta situação dicotômica (y = evasão ou conclusão), existem estudos específicos na Estatística que abordam essa análise, entre eles está a Análise discriminante e a Regressão logística, onde em busca de uma classificação mais eficiente dos grupos de alunos, optou-se por fazer a aplicação das duas técnicas no conjunto de dados e após isso uma comparação entre elas.

2 Referencial Teórico

Nesse estudo serão analisados dados a respeito da evasão de alunos do curso de estatística matriculados na Universidade Estadual Paulista, Faculdade de Ciências e Tecnologia do Campus de Presidente Prudente. Outros trabalhos realizados anteriormente para a instituição serviram de base ao atual projeto.

Utilizando *Educational Data Mining*, Aragão (2020) aplicou técnicas de Regressão Logística, Análise de Agrupamento, K-means, entre outras apenas nos alunos de estatística de 2008 a 2020, construindo dois modelos para aplicação. Após separar a base de dados em 70% treino e 30% teste, apenas três variáveis apresentaram ser significantes para a predição, sendo elas a classificação no vestibular, se cursou um Ensino Médio público e se o aluno entrou através de cotas. No primeiro modelo obtido a predição não se mostrou eficaz, enquanto no segundo modelo (reduzido) foram apresentados bons resultados.

No projeto de Alves (2018), utiliza-se uma base de dados considerando todos os alunos matriculados em curso de Graduação na UNESP no ano de 2000 a 2015, na qual é realizada a aplicação de técnicas de Modelos Lineares Generalizados (MLG) e Regressão Logística. Sete variáveis que se demonstraram importantes (área do curso, gênero, como se mantém no curso, tipo de escola no Ensino Médio, idade de ingresso, tempo vinculado e coeficiente de rendimento) foram compostas nos modelos de abordagem Clássica e Bayesiana, entretanto apenas o primeiro se mostrou possível a aplicação. Para futuras análises é deixado como sugestão a inclusão de efeitos aleatórios no modelo afim de comparar a eficiência das extensões dos MLG.

Para predição da evasão dos discentes do curso de Estatística, em busca de possíveis resultados diferentes e modelos mais acurados, além da utilização de um período de ingresso maior, nesse estudo serão abordadas técnicas estatísticas de Análise Multivariada comparando Análise Discriminante com Regressão Logística.

3 Metodologia

O objetivo do estudo desse projeto consiste em analisar como se comportam os matriculados no curso de Estatística e identificar quais fatores possuem mais influência no modelo de previsão da evasão dos estudantes. Os dados a respeito das matrículas dos alunos são fornecidos diretamente pela Universidade e por serem dados sensíveis é necessário a autorização da mesma para uso e análise. Nele se encontram informações desde dados básicos como RA, sexo, idade, até dados mais relacionados com o desempenho dos estudantes do curso.

Serão incluídos no estudo, a princípio, todo estudante que já fez parte do curso de Estatística da UNESP de Presidente Prudente durante o período de 2008 a 2017, sendo eles já formados ou evadido. Nesse trabalho serão utilizados os softwares para análise estatística Minitab e RStudio.

Com o banco de dados já obtido, será observado de forma mais precisa as variáveis existentes e avaliada a inclusão de novas que possam complementar o estudo. Serão utilizadas técnicas de Análise Multivariada e Regressão Logística para observar o comportamento das matrículas dos alunos. Das técnicas, será utilizada Análise Discriminante Múltipla e Regressão Logística para dois grupos. Os dados serão divididos em treinamento e teste para que após a criação do processo haja uma forma de validar o quão eficiente ele está e assim avaliar qual procedimento estatístico se mostrou mais acurado.

4 Fundamentação Teórica

Neste capítulo serão abordados o conceito das duas técnicas estatísticas que terão aplicação nesse trabalho: Regressão Logística e Análise Discriminante, assim como seus métodos de avaliação das estimações.

4.1 Regressão Logística

A regressão logística é uma técnica estatística multivariada desenvolvida na década de 1960 em resposta ao desafio de realizar predições e explicar a ocorrência de determinados eventos quando a variável de seu interesse fosse binária ou dicotômica, diferente de um modelo linear, onde a variável resposta é contínua. Essa diferença reflete na escolha do modelo paramétrico e nas suposições

Através dessa técnica é possível aferir a probabilidade de ocorrência de um fenômeno e identificar características dos elementos pertencentes a cada grupo determinado pela variável resposta, em outras palavras, podemos identificar as variáveis mais significativas para previsão da ocorrência de um evento de interesse, permitindo estimar diretamente sua probabilidade.

Um dos primeiros estudos que contribuíram para conferir notoriedade à técnica foi o Framingham Hearth Study, iniciado em 1948 com a colaboração da Universidade de Boston, que tinha como objetivo identificar fatores que propiciam doenças cardiovasculares em uma amostra de 5.209 indivíduos com idade entre 30 e 60 anos residentes na cidade de Framingham, Massachusetts. Com a utilização da regressão logística, vários fatores foram identificados como sendo de risco, como hipertensão arterial, taxas altas de colesterol, tabagismo, obesidade, diabetes e sedentarismo (FÁVERO, BELFIORE, SILVA, CHAN, 2009).

Trata-se de uma técnica muito difundida em diversos campos do conhecimento humano, como ciências médicas, mercado financeiro e nas ciências sociais, principalmente em função da facilidade de sua aplicação e da flexibilização de seus pressupostos, se comparado a outras técnicas, como regressão múltipla e análise discriminante.

Apesar do modelo de resposta binário ser o mais conhecido, ela também é capaz de descrever a relação no caso multinominal, ou seja, quando a variável resposta possui mais de duas categorias. Quanto às explicativas, tanto podem ser

categóricas ou não. Há então, uma separação de dois tipos de variáveis no modelo de Regressão logística:

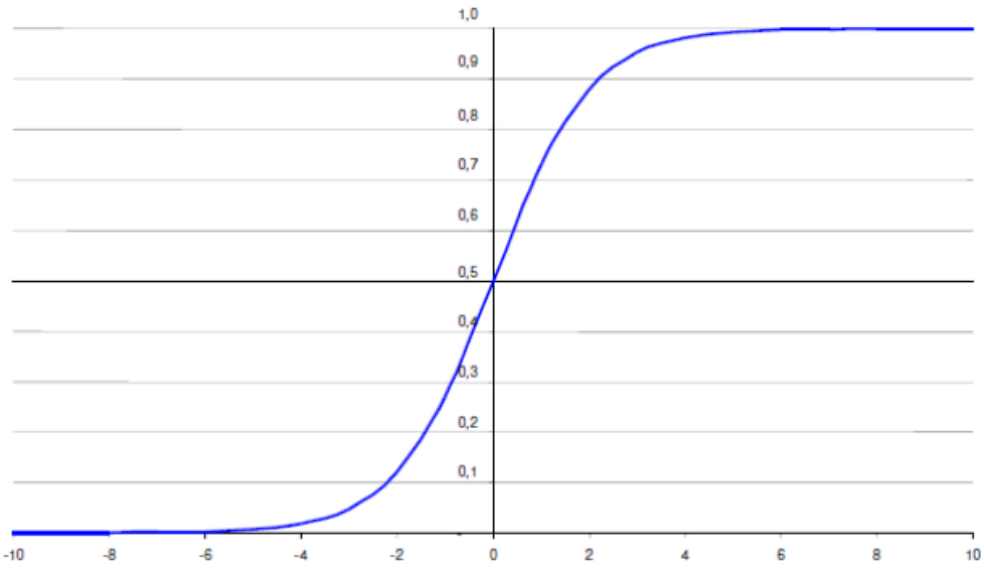
- Variável de interesse (Y): Variável resposta ou dependente;
- Variável auxiliar (X): Variável explicativa, preditora, covariáveis ou independentes.

4.1.1 Regressão Logística Binária

Nesta seção apresenta-se o contexto em que a variável resposta possui apenas duas categorias, ou seja, natureza binária ou dicotômica, e apenas uma variável independente envolvida.

A regressão logística, como dito anteriormente, é um recurso que nos permite estimar a probabilidade associada à ocorrência de determinado evento em face de um conjunto de variáveis explicativas. Uma característica marcante dessa técnica é a de sua função que apresenta comportamento probabilístico no formato da letra “S”, cujos valores se situam entre 0 e 1, representando a probabilidade de ocorrência do evento de interesse. Tal função pode ser observada na Figura 1.

Figura 1 – Gráfico da Função Inversa



Fonte: Gonzalez (2018)

Para simplificar a notação, usamos a quantidade $\pi\,(x) = E\,(Y|\,X)$, quando usamos distribuição logística. A forma do modelo é:

$$\pi\,(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{1+e^{(\beta_0 + \beta_1 x)}}, \tag{4.1}$$

ou ainda, é comum usar:

$$\pi\,(x) = \frac{1}{1+e^{-(\beta_0 + \beta_1 x)}}. \tag{4.2}$$

Uma transformação de π(x) que é central para o estudo da regressão logística é a transformação Logito, ou razão de desigualdades. O Logito, $g(x)$, tem muitas das propriedades que queremos de um modelo de regressão linear. É linear em seus parâmetros que devem ser contínuos, pode variar de −∞ a ∞, dependendo da variação de x. Também é fundamental para a estimação dos coeficientes das variáveis independentes.

Para chegar a essa transformação, temos que:

$$1 - \pi\,(x) = \frac{1+e^{(\beta_0 + \beta_1 x)} - e^{(\beta_0 + \beta_1 x)}}{1+e^{(\beta_0 + \beta_1 x)}} = \frac{1}{1+e^{(\beta_0 + \beta_1 x)}}. \tag{4.3}$$

Então, através da razão de desigualdades,

$$\frac{\pi\,(x)}{1-\pi\,(x)} = \frac{\frac{e^{(\beta_0 + \beta_1 x)}}{1+e^{(\beta_0 + \beta_1 x)}}}{\frac{1}{1+e^{(\beta_0 + \beta_1 x)}}} = \frac{e^{(\beta_0 + \beta_1 x)}}{1+e^{(\beta_0 + \beta_1 x)}} \cdot \frac{1+e^{(\beta_0 + \beta_1 x)}}{1} = e^{(\beta_0 + \beta_1 x)}. \tag{4.4}$$

Portanto,

$$g(x) = \ln\left[\frac{\pi\,(x)}{1-\pi\,(x)}\right] = \ln[e^{(\beta_0 + \beta_1 x)}] = \beta_0 + \beta_1 x. \tag{4.5}$$

Uma das diferenças entre a Regressão Logística e a Linear se deve ao comportamento dos erros (ε), a parte aleatória do modelo. Com o modelo em questão sendo binário, ou seja, a variável resposta y assumindo dois valores (0 ou 1), implica que o ε pode assumir duas possibilidades de valores também, conforme ilustrado no quadro abaixo.

Quadro 1 - Comportamento dos erros (ε).

y	1	0
ε	$1 - \pi\,(x)$	$-\pi\,(x)$
$P(x)$	$\pi\,(x)$	$1 - \pi\,(x)$

Sendo assim, o erro possui uma distribuição com média zero e variância igual a $\pi\,(x)[1 - \pi\,(x)]$. A distribuição condicional $Y|x$ é binomial com a probabilidade dada por $\pi\,(x)$.

Concluindo então alguns pressupostos para a aplicação de Regressão Logística:

1. A variável dependente é dicotômica;
2. $0 \leq E(Y|X) \leq 1$;
3. $\varepsilon \sim Binomial.$

Estimação dos Coeficientes (RLB)

Com a equação $\pi\,(x)$ definida, é necessário realizar o ajuste do modelo de regressão nela, e através do conjunto de dados, fazer a estimação dos parâmetros desconhecidos.

A estimação dos coeficientes da regressão logística, ao contrário da regressão múltipla que utiliza o método dos mínimos quadrados, é efetuada pelo uso da máxima verossimilhança. Esta, por sua vez, busca encontrar as estimativas mais prováveis dos coeficientes e maximizar a probabilidade de que um evento ocorra. A qualidade do ajuste do modelo é avaliada pelo "pseudo" R_2 e pelo exame da precisão preditiva (matriz de confusão).

Com Y assumindo uma distribuição Bernoulli, temos sua distribuição de probabilidade de cada observação da amostra dada por:

$$f_i(y_i) = \pi\,(x_i)^{y_i}\,.\,[1 - \pi\,(x_i)]^{1-y_i}. \qquad (4.6)$$

Desde que as observações são assumidas independentes, constrói-se a função de verossimilhança. A demonstração a seguir foi retirada de (TACHIBANA, 2021):

$$l(y_i, \dots, y_i; \beta) = \prod_{i=1}^{n} f_i(y_i) = \prod_{i=1}^{n} \pi\,(x_i)^{y_i}\,.\,[1 - \pi\,(x_i)]^{1-y_i} \qquad (4.7)$$

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

$$= \sum_{i=1}^{n} \{ y_i [ln(\pi(x_i)) - ln[1-\pi(x_i)]] + ln[1-\pi(x_i)] \} =$$

$$= \sum_{i=1}^{n} \left\{ y_i \left[ln \frac{\pi(x_i)}{1-\pi(x_i)} \right] + ln[1-\pi(x_i)] \right\} =$$

$$= \sum_{i=1}^{n} \{ y_i [\beta_0 + \beta_1 x_i] + ln[1-\pi(x_i)] \} =$$

$$\therefore L(y,\beta) = \sum_{i=1}^{n} \{ y_i [\beta_0 + \beta_1 x_i] + ln[1 + e^{\beta_0 + \beta_1 x_i}] \}. \quad (4.8)$$

E a partir de (3.8), obtém-se os seguintes resultados fazendo as derivadas parciais:

$$\frac{\partial L(y,\beta)}{\partial \beta_0} = 0 \rightarrow \sum_{i=1}^{n} \left\{ y_i - \frac{e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \rightarrow \sum_{i=1}^{n} \{ y_i - \hat{\pi}(x_i) \} = 0 \quad (4.9)$$

$$\frac{\partial L(y,\beta)}{\partial \beta_1} = 0 \rightarrow \sum_{i=1}^{n} \left\{ y_i x_i - \frac{x_i e^{\beta_0 + \beta_1 x_i}}{1 + e^{\beta_0 + \beta_1 x_i}} \right\} = 0 \rightarrow \sum_{i=1}^{n} \{ x_i [y_i - \hat{\pi}(x_i)] \} = 0 \quad (4.10)$$

Na regressão logística as equações (3.9) e (3.10) não são lineares em β0 e β1, e assim requerem métodos especiais para suas soluções, como softwares, que facilitam o cálculo.

Assim, obtém-se os modelos e os parâmetros estimados:

$$\hat{\pi}(x) = \frac{e^{(\hat{\beta}_0 + \hat{\beta}_1 x)}}{1 + e^{(\hat{\beta}_0 + \hat{\beta}_1 x)}}. \quad (4.11)$$

4.1.2 Regressão Logística Múltipla

A regressão múltipla representa o contexto da regressão logística, em que a variável dependente Y é binária ou dicotômica e que há mais de uma variável independente x. A força de uma técnica de modelagem consiste na capacidade em modelar muitas variáveis, algumas das quais podem ser em escalas de medidas diferentes.

Há uma semelhança muito grande entre o caso binário e o múltiplo. Podemos considerar então, a Regressão Logística Múltipla como uma generalização da Regressão Logística Binaria, para o caso de mais de uma variável explicativa.

Neste contexto, temos um conjunto de variáveis independentes denotados por $X = (x_1, x_2, x_3, \dots, x_{p)}$. Então, a forma do Logito da RLM é dado pela equação:

$$g(x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p. \quad (4.12)$$

Substituindo em $\pi\,(x)$, temos:

$$\pi\,(x) = \frac{e^{g(x)}}{1+e^{g(x)}} = \frac{e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}{1+e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p}}. \quad (4.13)$$

As variáveis auxiliares podem ser tanto contínuas quanto discretas nominais, como nível de escolaridade, sexo etc., sendo inapropriado inclui-las no modelo como se fossem de escala intervalar. Isto porque os números usados para representar os vários níveis são simplesmente identificadores e não têm significado numérico. Neste caso, usa-se um grupo de variáveis de planejamento (ou variável dummy).

Em geral, se uma variável nominal tem k categorias, então são necessárias $k - 1$ variáveis dummy, por exemplo, se temos a variável nível de escolaridade com três classes (ensino médio completo, ensino superior incompleto e ensino superior completo) então a codificação ficaria igual ao quadro abaixo.

Quadro 2 – Representação do Uso de variáveis Dummy.

Nível de Escolaridade	Variável Dummy	
	D_1	D_2
EM completo	0	0
ES incompleto	1	0
ES completo	0	1

Neste caso, com o uso de variáveis dummy, a equação do Logito é representada da seguinte forma:

$$g(x) = \beta_0 + \beta_1 x_1 + \cdots + \sum_{u=1}^{k_j - 1} \beta_{ju} D_{ju} + \beta_p x_p, \quad (4.14)$$

sendo β_{ju} os coeficientes da variável e D_{ju} as variáveis de planejamento.

Estimação dos Coeficientes (RLM)

Assim como no caso univariado, a estimação dos coeficientes é feita através do método de verossimilhança, entretanto, no modelo de Regressão Logística Múltipla, tem-se um vetor $\beta' = \beta_0, \beta_1, \dots, \beta_p$ para serem estimados.

Após a derivação da função log de verossimilhança, é obtido $p + 1$ equações referentes aos $p + 1$ coeficientes. Sendo elas:

$$\sum_{i=1}^{n}\{y_i - \hat{\pi}(x_i)\} = 0 \text{ e } \sum_{i=1}^{n}\{x_{ij}[y_i - \hat{\pi}(x_i)]\} = 0, \text{ para } j = 1,2,\dots,p \quad (4.15)$$

Então, realizado o cálculo da estimação desses parâmetros através de softwares, têm se o Logito e a equação de $\pi(x)$, conforme representada abaixo.

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} \quad (4.16)$$

$$\text{e } \hat{g}(x) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p. \quad (4.17)$$

4.1.3 Métodos de Avaliação do Modelo

Após estimar os coeficientes, temos interesse em assegurar a significância das variáveis no modelo. Isto geralmente envolve formulação e teste de uma hipótese estatística para determinar se as variáveis independentes no modelo são significativamente relacionadas com a variável dependente. Para isto, há testes para avaliar o modelo logístico. Os testes mais utilizados são os testes da Razão da Verossimilhança, o teste de Wald e Pseudo R_2 de Cox e Snell (HOSMER, LEMESHOW, STURDIVANT, 2013).

Teste de Significância

Na regressão logística vamos comparar os valores preditos da variável resposta obtidos dos modelos com e sem a variável explicativa. A comparação é baseada na função do log da verossimilhança. A comparação entre os valores observados e esperados usando a função de verossimilhança, é expressa da seguinte forma:

$$G = D[modelo\ sem\ a\ variável] - D[modelo\ com\ a\ variável], \qquad (4.18)$$

em que,

$$D = -2ln\left[\frac{verossimilhança\ do\ modelo\ atual}{verossimilhança\ do\ modelo\ saturado}\right]. \qquad (4.19)$$

O modelo é dito saturado se contém todas as variáveis, enquanto o modelo ajustado corresponde ao modelo apenas com as variáveis desejadas para o estudo. Esta função D, também chamada de deviance (desvio), sempre é positiva e quanto menor, melhor é o ajuste do modelo.

Sob a hipótese nula que os p coeficientes de inclinação para as covariáveis no modelo são iguais a zero, a distribuição de G será qui-quadrado com p graus de liberdade.

$$\begin{cases} H_0: \beta = 0 \\ H_1: \beta \neq 0\,, \quad (\exists\ B_j \neq 0) \end{cases}.$$

E então, a partir de (4.18), têm se que:

$$G = 2\left\{\sum_{i=1}^{n}\left[y_i\ ln(\hat{\pi}_i) + (1-y_i)\ ln[1-\hat{\pi}_i]\right] - \{n_1\ ln(n_1) + n_0\ ln(n_0) - n\ ln(n)\}\right\} \qquad (4.20)$$

Teste de Wald

O teste de Wald é obtido comparando a estimativa do parâmetro com o desvio padrão. Após concluir que pelo menos um parâmetro é diferente de zero, calcula-se a

estatística do teste univariado de Wald para testar se cada parâmetro é diferente de zero individualmente. Deste modo, o teste de Wald verifica se uma determinada variável independente possui uma relação estatisticamente significativa com a variável dependente (GONZALEZ, 2018).

Admitindo as hipóteses:

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

Sob a hipótese nula que o coeficiente é igual a zero, a distribuição de W tem distribuição Normal (0,1). A estatística do teste é dada por:

$$W_j = \frac{\hat{\beta}_j}{D.P(\hat{\beta}_j)}. \qquad (4.21)$$

Curva ROC e AUC

A curva ROC, ou Receiver Operating Characteristic, traça a relação entre a sensibilidade (verdadeiro positivo) e a especificidade (verdadeiro negativo) de cada valor encontrado.

E a métrica AUC, Area Under the ROC, representa uma maneira de resumir a curva ROC em um único valor. Esse valor pertence ao intervalo de 0 a 1. Quando mais próximo de 1 melhor é o desempenho do modelo, pois demonstra ter mais previsões corretas.

Através dessas duas métricas é possível comparar os modelos propostos e ter auxílio na escolha do com o melhor desempenho.

4.2 Análise Discriminante

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. Considerada exploratória por natureza, é geralmente empregada para investigar diferenças observadas quando relações causais não são bem compreendidas.

A técnica estuda a separação de objetos de uma população em duas ou mais classes. A discriminação ou separação é a primeira etapa, sendo a parte exploratória da análise e consiste em se procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos. A classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos (JOHNSON e WICHERN, 2007).

Normalmente, discriminação e classificação se sobrepõem na análise, e a distinção entre separação e alocação é confusa. O problema da discriminação entre dois ou mais grupos, visando uma classificação, foi inicialmente abordado por Fisher (1936). O método foi introduzido em sua forma inicial para problemas com duas classes. Tem aplicações em diversos campos do conhecimento, como na biologia, nas finanças e na tecnologia, entre outros.

Como objetivo principal, a análise discriminante oferece ao pesquisador a possibilidade de elaborar previsões a respeito de a qual grupo certa observação (por exemplo, um produto, uma pessoa ou uma empresa) pertencerá, uma vez que se caracteriza como uma técnica de previsão e classificação. [...]. Em resumo, a análise discriminante ajuda o pesquisador que deseja criar um estudo com grupos diferenciados por meio das variáveis independentes que possui. (FÁVERO, BELFIORE, SILVA, CHAN, 2009).

4.2.1 Análise Discriminante Simples

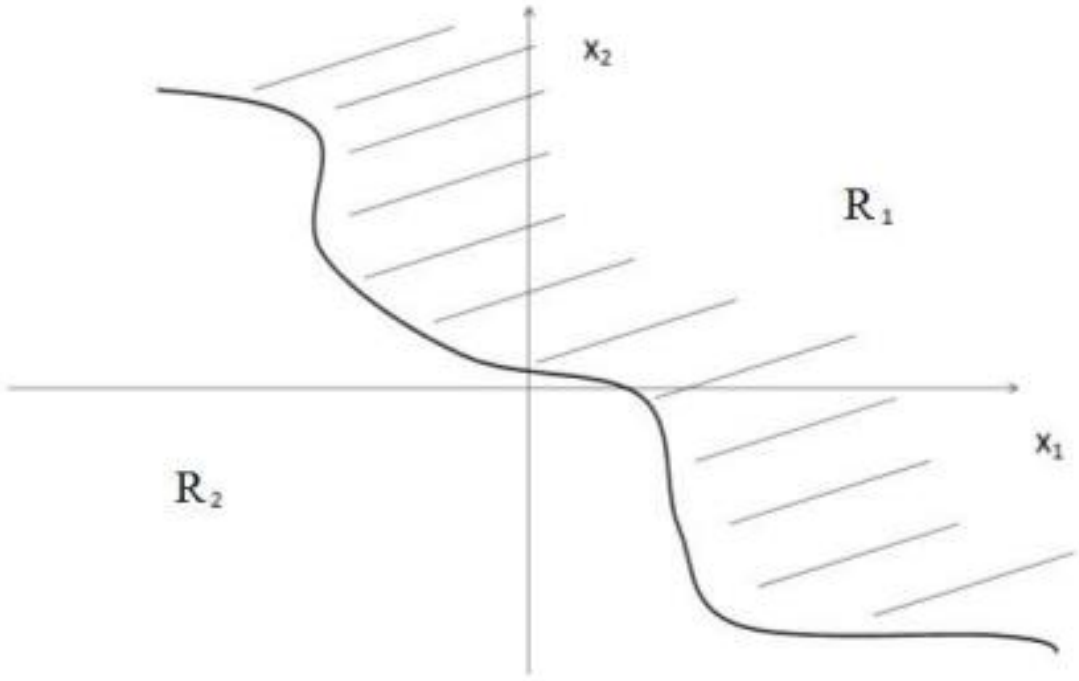
Sendo $\pi_1, \pi_2, \dots, \pi_g$ um grupo de g populações. Para se obter as funções discriminantes para esse grupo são necessárias algumas pressuposições:

- as populações apresentam algum tipo de distribuição;
- existe uma probabilidade de ocorrência a priori para cada população no grupo;
- existe um custo de má classificação.

Seja $f_1(x)\ e\ f_2(x)$ as funções densidade de probabilidade associadas com o vetor de variáveis aleatórias $px1$ para as populações $\pi_1\ e\ \pi_2$, respectivamente. Um objeto com medidas x associadas deve ser atribuído a uma das duas populações. Seja Ω o espaço amostral, isto é, a coleção de todas as observações possíveis x, seja R_1 o conjunto de todos os valores x para os quais classificamos objetos como

pertencente à população π_1 e seja $R_2 = \Omega - R_1$ o conjunto dos valores x remanescentes para os quais classificamos o objeto como pertencente à π_2. Desde que cada objeto deve ser atribuído a uma e apenas uma das duas populações, os conjuntos R_1 *e* R_2 são mutuamente exaustivos e exclusivos. Para $p = 2$, podemos ter um caso como o da Figura 2. (SILVESTRE, 2021)

Figura 2 – Regiões de classificação para duas populações.



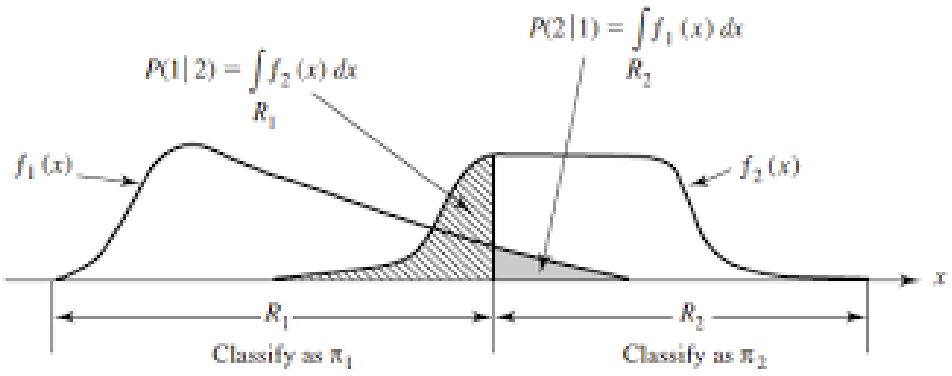
Fonte: Silvestre (2021)

Seja p_1 a probabilidade a priori de π_1 e p_2 a probabilidade a priori de π_2, tal que $p_1 + p_2 = 1$. As probabilidades envolvendo os erros e acertos na classificação são dadas no Quadro 3 e Figura 3, respectivamente.

Quadro 3 – Probabilidade de classificação.

População verdadeira	Classificado como	
	π_1	π_2
π_1	$P(1\|1)p_1$	$P(2\|1)p_1$
π_2	$P(1\|2)p_2$	$P(2\|2)p_2$

Figura 3 - Probabilidade de classificação incorreta para regiões hipotéticas



Fonte: Johnson e Wichern (2007).

Com objetivo de diminuir os erros de classificação, o estudo de características adicionais é fundamental. A probabilidade a priori de ocorrência observa diferenças entre as populações, em alguns casos a probabilidade de ocorrência de uma classe pode ser maior que a de outra devido a discrepância do tamanho de suas populações. O custo pode ser considerado como um peso a má classificação daquele objeto, em certas ocasiões o erro é mais sério necessitando de uma maior cautela na hora da atribuição.

O *Expected Cost of Misclassification* (ECM) ou custo de classificação incorreta esperado é fornecido pré-multiplicando os o custo de uma classificação incorreta por suas probabilidades de ocorrência. Para uma melhor classificação é importante que seu valor seja baixo.

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \qquad (4.22)$$

As regiões $R1\ e\ R2$ que minimizam o ECM são definidas pelos valores de *x* para os quais as seguintes desigualdades acontecem:

E

$$R_1: \frac{f_1(x)}{f_2(x)} \geq \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1}$$

$$R_2: \frac{f_1(x)}{f_2(x)} < \frac{c(1|2)}{c(2|1)}\frac{p_2}{p_1} \qquad (4.23)$$

Há casos especiais de regiões de custos esperados minimos que se destacam na forma de uso das desigualdades apresentadas acima, quando probabilidades a priori são iguais, o custo de classificação incorreta são iguais ou ainda um terceiro caso contemplando os dois anteriores.

Quando os custos de classificação incorreta são iguais, uma outra maneira de minimizar o ECM é através do *total probability of misclassification* (TPM).

Escolher as regiões R1 e R2 que minimizassem ou seja, a probabilidade total de classificaçãoincorreta através da seguinte equação:

$$TPM = P(classificar\ incorretamente) = P(2|1) + P(1|2)$$
$$= p_1 \int_{R_2} f_1(x)dx \ + p_2 \int_{R_1} f_2(x)dx \qquad (4.24)$$

Também poderia ser alocada a nova observação $\boldsymbol{x}_0$ à população com a maior probabilidade a posteriori:

$$P(\pi_1|\boldsymbol{x}_0) = \frac{p_1 f_1(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}$$

E (4.25)

$$P(\pi_2|\boldsymbol{x}_0) = \frac{p_2 f_2(\boldsymbol{x}_0)}{p_1 f_1(\boldsymbol{x}_0) + p_2 f_2(\boldsymbol{x}_0)}$$

4.2.2 Análise Discriminante de Fisher

A função discriminante de Fisher é uma combinação linear de características a qual se caracteriza por produzir separação máxima entre duas populações, através de funções mais simples. Segundo Johnson e Wichern (2007), sua ideia era transformar observações multivariadas x em observações univariadas y tal que os $y's$ derivados da população $\pi 1$ e $\pi 2$ fossem tão separados quanto possível. A abordagem de Fisher não requer que as populações sejam normais. Porém, assume

implicitamente que as matrizes de covariância são iguais, porque a estimativa da matriz de covariância agrupada é usada.

A combinação linear é dada por:

$$\hat{y} = \hat{\boldsymbol{a}}\boldsymbol{x} = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_{agrup}^{-1}\boldsymbol{x}. \tag{4.26}$$

E a separação:

$$\frac{|\hat{y}_1 - \hat{y}_2|}{s_y}, \tag{4.27}$$

e a variância agrupada:

$$s_y^2 = \frac{\sum_{j=1}^{n_1}(y_{1j}-\bar{y}_1)^2+\sum_{j=1}^{n_2}(y_{2j}-\bar{y}_2)^2}{n_1+n_2-2}. \tag{4.28}$$

Como o objetivo desse método é maximizar a separação dos y's derivados da população π_1 *e* π_2. Temos através de:

$$\frac{(\bar{y}_1-\bar{y}_2)^2}{s_y^2} = \frac{(\hat{\boldsymbol{a}}'\bar{\boldsymbol{x}}_1-\hat{\boldsymbol{a}}'\bar{\boldsymbol{x}}_2)^2}{\hat{\boldsymbol{a}}' S_{agrup}\hat{\boldsymbol{a}}} = \frac{(\hat{\boldsymbol{a}}'\boldsymbol{d})^2}{\hat{\boldsymbol{a}}' S_{agrup}\hat{\boldsymbol{a}}}, \tag{4.29}$$

tal que, $d = (\bar{x}_1 - \bar{x}_2)$ e o máximo valor da razão, a qual maximiza a separação das duas populações será:

$$D = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_{agrup}^{-1}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2). \tag{4.30}$$

As regiões de alocação definidas por:

$$R_1: \hat{y}_0 \geq \hat{m} \quad ou \quad R_2: \hat{y}_0 < \hat{m}\,. \tag{4.31}$$

Sendo,

$$\hat{y}_0 = (\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_{agrup}^{-1}\,\boldsymbol{x}_0 \quad e \quad \hat{m} = \frac{1}{2}(\bar{\boldsymbol{x}}_1 - \bar{\boldsymbol{x}}_2)' S_{agrup}^{-1}(\bar{\boldsymbol{x}}_1 + \bar{\boldsymbol{x}}_2) \tag{4.32}$$

*Sendo possível apenas se $p(número\ de\ variáveis) \leq n_1 + n_2 - 2$, fazendo com que a matriz S_{agrup}^{-1} exista.

4.2.3 Desempenho de Classificação

Existe uma medida de desempenho que não depende da forma de populações e pode ser calculada por algum procedimento de classificação. Esta medida, chamada de *apparent error rate* (APER), razão do erro aparente, é definida como a fração de observações em amostras de treinamento que são erroneamente classificadas pela função de classificação amostral. A razão do erro aparente será facilmente calculada pela matriz de confusão, que mostra os valores atuais dos membros preditos dos grupos (JOHNSON & WICHERN, 2007).

Para n_1 observações da população $\pi 1$ e n_2 observações da população $\pi 2$, a matriz de confusão tem a forma descrita no Quadro 3.

Quadro 4 – Matriz de Confusão do APER

		Membro predito		
		π_1	π_2	
Membro atual	π_1	n_{1C}	$n_{1M} = n_1 - n_{1C}$	n_1
	π_2	$n_{2M} = n_2 - n_{2C}$	n_{2C}	n_2

n_{1C} = número de itens de π_1 corretamente classificados como itens de π_1.
n_{1M} = número de itens de π_1 incorretamente classificados como itens de π_2.
n_{2C} = número de itens de π_2 corretamente classificados como itens de π_2.
n_{2M} = número de itens de π_2 incorretamente classificados como itens de π_1.

A equação abaixo, mostra a forma de mensurar a Razão do Erro Aparente, o qual é obtido uma proporção das classificações incorretas.

$$APER = \frac{n_{1M}+n_{2M}}{n_1+n_2}. \tag{4.33}$$

4.3 Testes

Para verificar as hipóteses pré-definidas e entender o comportamento dos dados, assim como os próximos passos a serem seguidos com o estudo, foram realizados três testes estatísticos.

4.3.1 Teste de Normalidade

O teste de Normalidade ou aderência, é utilizado para testar se um conjunto de observações provém de uma população com distribuição normal. Existem diversos métodos diferentes para realização das hipóteses.

$$H_0: A\ distribuição\ das\ observações\ é\ normal.$$
$$H_1: A\ distribuição\ das\ observações\ não\ é\ normal.$$

Um dos mais conhecidos é o Teste de Shapiro-Wilk, esse teste baseia-se em regressão e correlação, utilizando a razão entre duas estimativas obtidas de estatísticas de ordem.

Definido por:

$$W = \frac{1}{(n-k)s^2}\left(\sum_{i=1}^{n} a_i \widehat{z_i}\right)^2 \tag{4.34}$$

com

$$a' = (a_1, \dots, a_n) = \frac{c'V^{-1}}{(c'^{V^{-2}}c)^{1/2}} \tag{4.35}$$

Dos quais c' representa o vetor de valores esperados e V a matriz de covariâncias das estatísticas de ordem normal padrão.

4.3.2 Teste da Diferença de Médias (Numéricas)

Quando os dados não apresentam normalidade e é de interesse testar as médias, pode-se usar um teste não paramétrico. O teste de Wilcoxon ou teste de Soma das Ordens, baseia-se na ordenação das observações de duas amostras independentes. Geralmente sua utilização é dada para comparação de um grupo

controle e tratamento, e através das hipóteses determina se um grupo possui diferença sobre o outro.

$$H_0: \Delta = 0.$$
$$H_1: \Delta \neq 0\ .$$

Considerando duas amostras $X_i\ e\ Y_i$. O teste é calculado sendo:

$$X_i = e_i \quad \text{e} \quad Y_j = \Delta + e_{m+j}\ , \quad \text{onde} \quad i = 1,2,\dots,m \quad e \quad j = 1,2,\dots,n$$

Após a ordenação crescente da variável de $X_i\ e\ Y_i$, a estatística para decisão de rejeição ou não de H_0 é dada na comparação de W com $W_\propto$.

$$W = \sum_{j=1}^{n} O_j \tag{4.36}$$

Sendo O_j somente a ordem no conjunto de observações ordenados dos dois grupos.

4.3.3 Teste da Diferença de Médias (Categóricas)

Um método de comparação de médias para variáveis categóricas é o teste exato de Fisher. Esse teste é recomendado para casos em que há duas amostras, denotadas por A e B, de tamanho amostral pequeno, e através de tabelas de contingência é realizado o cálculo de probabilidade da existência de um valor igual ou mais extremo do que o menor valor existente na tabela.

$$H_0: P(A)\ = P(B).$$
$$H_1: P(A)\ \neq P(B)\ .$$

Tabela 1 – Tabela de Contingência Teste Exato de Fisher.

	I	II	
A	a	b	N_1
B	c	d	N_2
	$a + c$	$b + d$	N

Fonte: Campos (1983)

5 Desenvolvimento

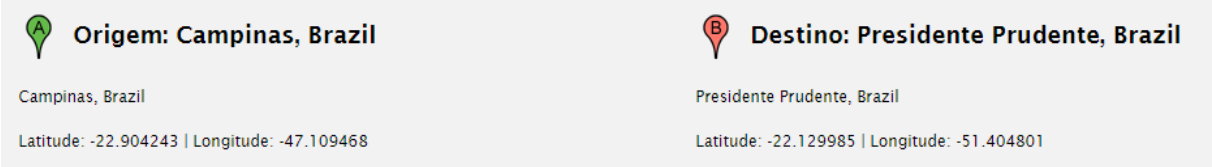
Para facilidade e mais exatidão nos cálculos, a parte relacionada ao desenvolvimento do estudo e obtenção dos resultados foram realizadas pelo software de programação RStudio. Durante este capítulo será descrito a aplicação dos dados nos modelos gerados para estimação das classes, assim como testes e análises necessárias para o andamento.

5.1 Dados e Análise Descritiva

Os dados são referentes aos alunos de Estatística da UNESP de Presidente Prudente matriculados em 2008 a 2017. No total a tabela possui 216 linhas e 27 colunas com informações dos estudantes, para garantir a segurança e privacidade dos alunos, os dados foram disponibilizados sem o número de Registro do Aluno (RA) e sim com um ID apenas para identificação de cada linha.

Duas das variáveis existentes, *DistKM* e *DisRetaKM*, foram construídas a partir de um site que realiza o cálculo em quilômetros da distância do ponto central da Cidade origem, informada pelo estudante, ao ponto central da cidade de Presidente Prudente, a qual está localizada a Instituição de Ensino. Abaixo está exemplificado a utilização do site, onde a informação rosa representa o valor da distância em linha reta entre as cidades e, o valor azul representa a distância por estrada.

Figura 4 – Representação da Informação de Distância



Muitas variáveis possuíam a formatação em data (dd/mm/aaaa) e somente ano, sendo assim essas colunas no formato data foram desconsideradas como possíveis informações ao modelo pela complexidade de serem utilizadas no mesmo. Abaixo está a lista com as descrições de cada variável que permaneceram no banco de dados.

Quadro 5 – Descrição das Variáveis do Conjunto de Dados.

Variável	**Descrição**
sexo	Feminino Masculino
disretaKM	Distância em linha reta do ponto central de Presidente Prudente ao ponto central da cidade origem do aluno (1 a 972,43 quilômetros).
distKM	Distância percorrida por estrada do ponto central de Presidente Prudente ao ponto central da cidade origem do aluno (1 a 1386 quilômetros).
tipoingresso	Concurso Vestibular Portador de Diploma de Curso Superior Transferência Interna
cor	Branca Preta Parda Amarela Não Informado
ensinomedio	Privado Público
cotas	Ensino Público Preto, Pardo ou índio Não
classvest	Classificação do aluno na prova de ingresso da unesp (1 a 92)
notavest	Nota do aluno na prova de ingresso da unesp (26 a 77,76)
classe	Formado Evadido
ano_nasc	Ano de nascimento do aluno (1979 a 1999)
ano_mat	Ano da primeira matrícula do aluno (2008 a 2017)
ano_saida	Ano de saída do curso (2011 a 2022)
idade	Idade do aluno em anos ao ingressar no curso (16 a 29)

idade_saida	Idade do aluno em anos ao sair do curso, finalizado ou evadido (18 a 35)
tempo_cursado	Tempo em anos que o aluno permaneceu no curso (1 a 11)

Observando as informações contidas no banco de dados, foi notada uma inconsistência na variável *notavest* e *classvest*, possuindo campos faltantes em aproximadamente 5% do conjunto, sendo esses casos de alunos de transferência interna ou que entraram como portador de diploma. Considerando que o conjunto já possui uma população pequena, com a intenção de não perder essas informações foi realizado a estimação desses NA's e preenchidos através da mediana.

Tabela 2 – Mediadas de tendencia Central das variáveis *notavest* e *classvet.*

	Min	1st Qt	Mediana	Média	3rd Qt	Max	NA's
notavest	26.00	39.00	43.83	44.45	49.18	77.76	11
classvest	1.00	28.00	43.00	43.35	58.00	92.00	11

Considerando as variáveis que estavam relacionadas com as hipóteses estabelecidas no estudo, foi realizada a análise descritiva no conjunto de dados. Nos gráficos abaixo é realizada uma comparação entre o grupo de alunos evadidos e formados através do gráfico boxplot, sendo o grupo evadido e formado compostos por 35.6% e 64.4% respectivamente, dos 216 alunos totais.

Figura 5 – Gráfico Boxplot da Distância em Relação a Classe dos Alunos.

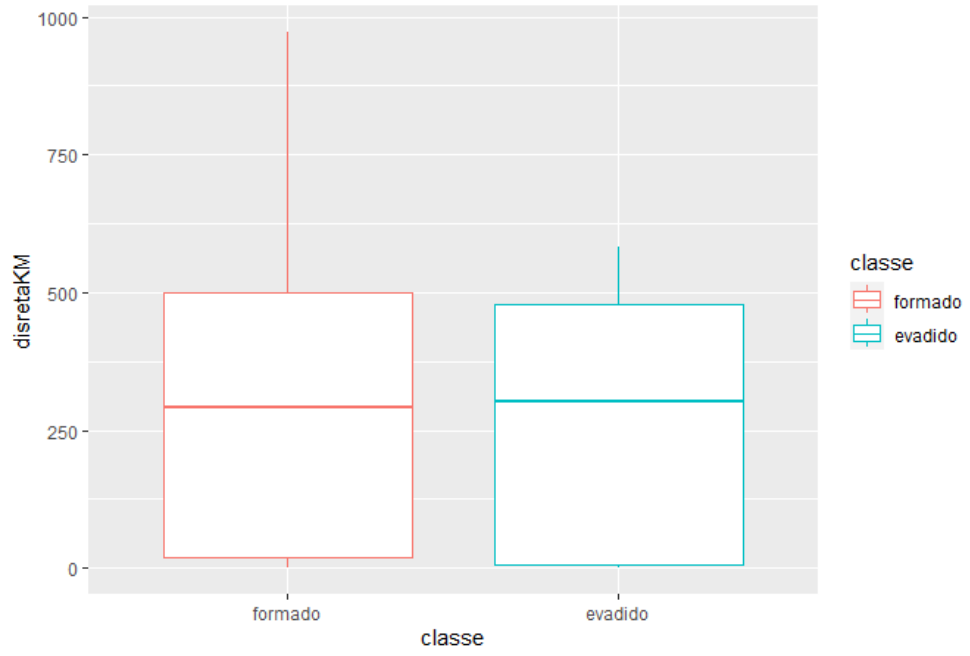


Figura 6 – Gráfico Boxplot da Nota em Relação a Classe dos Alunos.

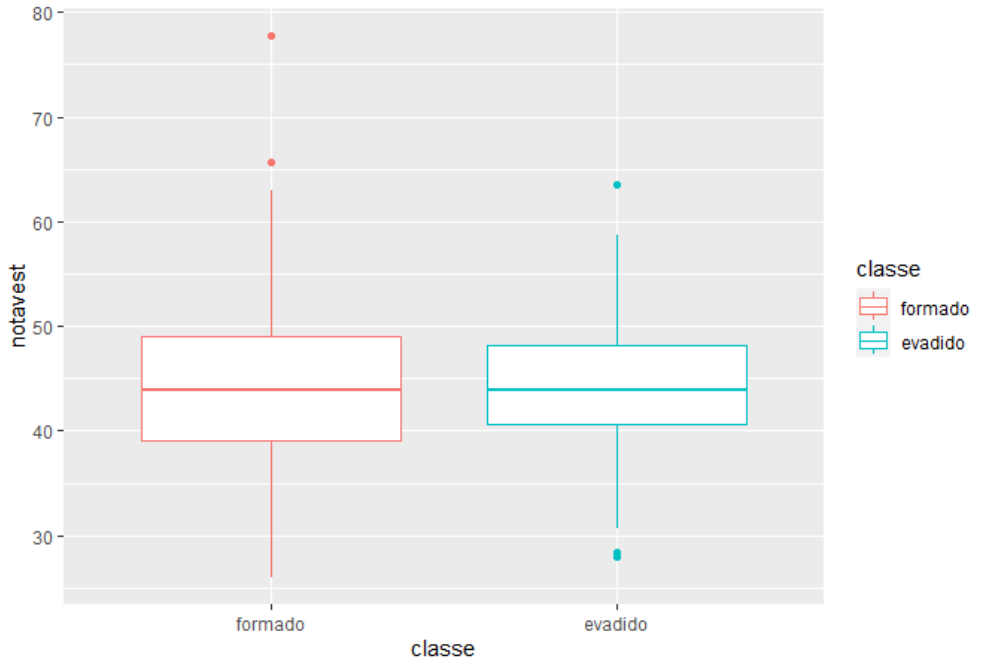


Figura 7 – Gráfico Boxplot da Idade em Relação a Classe dos Alunos.

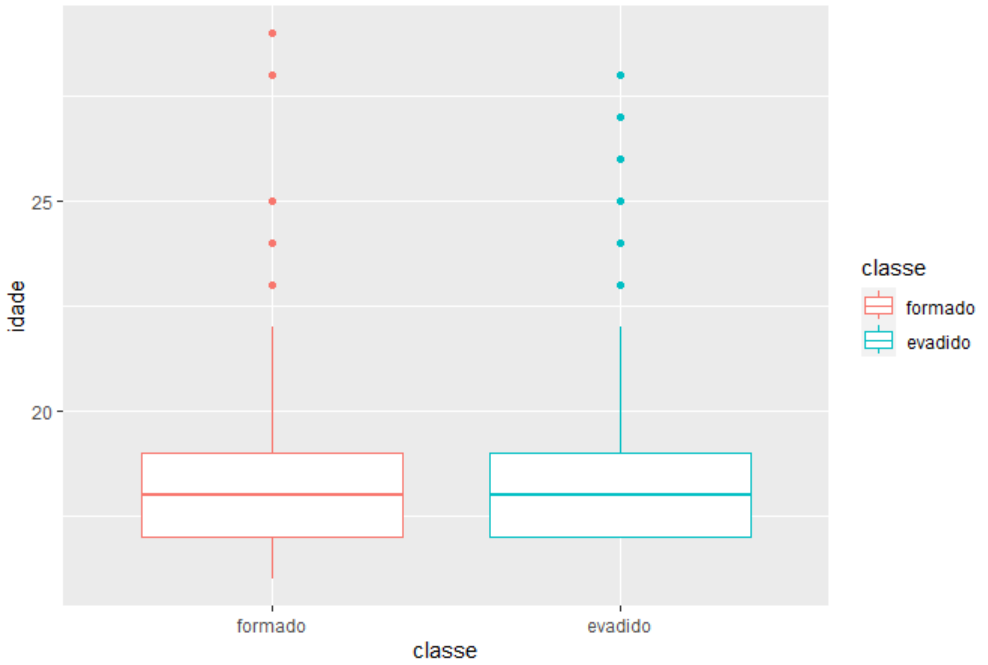
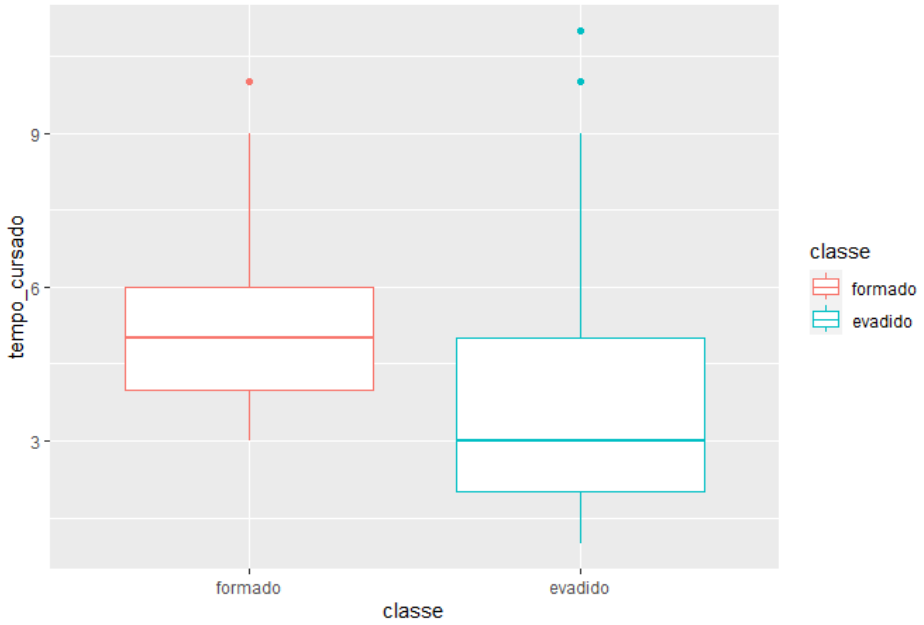


Figura 8 – Gráfico Boxplot do tempo cursado em Relação a Classe dos Alunos.



Nas figuras (5-7) os gráficos sugerem que apesar dos quartis, em alguns casos, apresentarem valores diferentes, ao observar a linha da média, aparentemente não há uma diferença do seu valor entre os dois grupos para as variáveis de distância entre cidade origem do aluno e Presidente Prudente, nota no vestibular e idade. Diferente da figura (8) em que a variável tempo de curso, apresenta um afastamento entre as duas médias.

Para validar essa diferença ou não diferença entre os dois grupos, foi aplicado testes estatísticos nas variáveis e pontuados posteriormente nesse estudo .

5.2 Testes

O primeiro teste a ser aplicado nas variáveis numéricas foi o de Shapiro Wilk para verificar a normalidade dos dados.

Tabela 3 – Valores Obtidos do Teste de Shapiro Wilk.

Variável	Todo	Apenas Evadidos	Apenas Formados
NotaVest	0.015	0.354	0.009
ClassVest	0.011	0.327	0.024
DisretaKM	0.000	0.000	0.000
DistKM	0.000	0.000	0.000
Idade	0.000	0.327	0.024

Através do p-valor é possível concluir que os dados não possuem normalidade, apesar de três das variáveis do grupo evadido terem um p-valor > 0.05, o grupo dos formados não possui normalidade, sendo inviável uma comparação de ambos pela abordagem paramétrica.

O teste de Soma das Ordens, ou Wilcoxon foi aplicado nas características que não possuem normalidade. Esse teste é uma alternativa não paramétrica considerando os grupos como independentes ou não pareados, quando os requisitos do teste T não são cumpridos.

Tabela 4 – Valores Obtidos do Teste de Wilcoxon.

Variável	**p-valor**
NotaVest	0.9565
ClassVest	0.2716
DisretaKM	0.7221
DistKM	0.7264
Idade	0.1913

Em nenhum dos resultados foi apresentado um valor significativo abaixo de 0.05 para não rejeitar a hipótese nula e concluir que a média dos grupos difere, o que pode apresentar ser um problema na parte de treinamento do modelo, pois o grupo de evadidos e formados não demonstram haver uma segregação entre eles.

No caso das variáveis categóricas foram testadas a diferença das médias pelo Teste Exato de Fisher para as informações de *sexo*, *ensinomedio*, *cor* e *cotas* dos estudantes.

Tabela 5 – Valores Obtidos no Teste Exato de Fisher

Variável	**p-valor**
sexo	0.476
ensinomedio	0.777
cor	0.946
cotas	0.858

Em nenhum dos casos foi observado uma diferença significativa entre as variáveis categóricas em relação ao grupo de evadidos e formados.

5.3 Regressão Logística

Para codificação das características categóricas em dummy foi utilizado a forma default do software R, apenas modificando quais categorias seriam consideradas de referência.

Os quadros abaixo representam a codificação das variáveis, onde para cada k categorias, eram criadas $k - 1$ variáveis auxiliares.

Quadro 7 – Codificação da Variável classe para o modelo.

classe	$\boldsymbol{D1}$
Formado	0
Evadido	1

Quadro 8 – Codificação da Variável sexo para o modelo.

sexo	$\boldsymbol{D1}$
Feminino	0
Masculino	1

Quadro 9 – Codificação da Variável Ensino Médio para o modelo.

ensino medio	$\boldsymbol{D1}$
E.M. Publico	0
E.M. Privado	1

Quadro 10 – Codificação da Variável cotas para o modelo.

cotas	$\boldsymbol{D1}$	$\boldsymbol{D2}$
Não	0	0
Ensino Publico	1	0
Preto, Pardo e Índio	0	1

Quadro 11 – Codificação da Variável tipo de ingresso para o modelo.

tipo ingresso	$\boldsymbol{D1}$	$\boldsymbol{D2}$
Vestibular	0	0
Diploma	1	0
Transferência	0	1

Quadro 12 – Codificação da Variável cor para o modelo.

cor	$D1$	$D2$	$D3$	$D4$
Branca	0	0	0	0
Amarela	1	0	0	0
Não Informada	0	1	0	0
Parda	0	0	1	0
Preta	0	0	0	1

A variável *classe* está como referência “formado”, a *sexo* possui “feminino”, *cotas* é referenciada através da categoria “Não”, seguindo com “Ensino Público” “Preto, pardo ou índio” e *cor* está ordenada por “BRANCA” “AMARELA” “NÃO INFORMADA” “PARDA” e “PRETA”.

Para criação do modelo e validação das estimações foi realizado a separação do conjunto de dados em Treino e Teste utilizando o método *Holdout.* Esse método consiste na separação de forma aleatória de 70% dos dados para treinamento do modelo e 30% para teste.

Utilizando a semente 1234 no R foi gerado a amostra para o treino, na qual contém os seguintes ID’s da população ordenados abaixo.

Quadro 6 – Linhas Utilizadas no conjunto de Treino.

2	4	6	8	9	10	11	14	16	17	19	20	21	22
23	25	26	28	29	30	32	35	36	38	40	41	42	43
46	48	49	51	53	57	58	60	61	62	63	66	67	68
70	71	72	73	74	77	79	80	82	83	85	86	87	88
90	91	93	94	96	97	98	100	101	102	103	105	106	108
109	111	112	113	115	116	117	119	122	123	124	125	126	127
129	130	131	132	133	134	135	136	137	138	139	142	144	145
147	149	150	153	154	155	156	157	158	159	160	161	162	163
165	166	167	168	170	171	172	173	174	175	176	178	179	180
181	182	184	185	187	188	191	192	193	194	195	198	200	201
202	203	204	205	207	208	210	214	215	216				

As 65 linhas restantes, identificadas pelo ID, foram separadas para o grupo de teste.

Com intuito de observar quais variáveis apresentam indícios de serem importantes para o modelo, foi realizado uma análise univariada para cada fator utilizando os dados separados para treino.

Tabela 6 – Valores Estimados no Modelo de Regressão Logística Univariado.

Variável	$\widehat{\beta}$	$DP(\widehat{\beta})$	OR	IC 95%	Log-Ver	p-valor
Constante	-0.557	0.169			-99,029	
Sexo M	0.455	0.341	1.57	0.81 ; 3.10	-98.131	0.182
DisRetaKM	0.00001	0.001	1.00	0.99 ; 1.00	-99.028	0.985
DistKM	-0.00003	0.001	0.99	0.99 ; 1.00	-99.027	0.959
TipoIngresso PD	15.147	882.743	0.00	0.00 ; 0.00	-97.995	0.986
TipoIngresso TI	0.176	0.929	1.19	0.15 ; 7.41		0.849
CorA	-0.387	0.858	0.67	0.09 ; 3.30	-98.624	0.652
CorN	-0.569	1.171	0.56	0.02 ; 4.57		0.626
CorPA	0.173	0.529	1.18	0.40 ; 3.32		0.744
CorPR	-0.387	0.858	0.67	0.09 ; 3.30		0.652
EnsinoMedio PB	-0.284	0.343	0.75	0.38 ; 1.47	-98.684	0.407
Cotas EP	0.151	0.671	1.16	0.28 ; 4.27	-98.976	0.821
Cotas PPI	-0.137	0.639	0.87	0.22 ; 2.92		0.830
ClassVest	0.0002	0.008	1.00	0.98 ; 1.01	-99.028	0.977
NotaVest	0.025	0.023	1.02	0.98 ; 1.07	-98.444	0.282
Ano_Nasc	0.092	0.052	1.09	0.99 ; 1.21	-97.378	0.075
Ano_mat	0.214	0.067	1.23	1.09 ; 1.41	-93.464	0.001
Tempo_cursado	-0.348	0.107	0.70	0.56 ; 0.86	-92.611	0.001
Idade	0.091	0.077	1.09	0.94 ; 1.28	-98.315	0.234

É possível observar que a Razão de Chances na maioria dos casos está bem próxima do valor 1, indicando que conforme é acrescentado uma unidade na variável explicativa, essa característica possui uma vez mais chance de estar na classe de interesse. Ou seja, como os valores estão próximos a 1, elas não influenciam tanto. O Log de Verossimilhança é outra informação importante a pontuar, na maioria dos casos ele não apresentou sofrer tanta alteração ao acrescentar as variáveis no modelo comparado com o valor do que possui somente o B0. Exceto no caso de *ano_mat* e *tempo_cursado*.

A informação que demandavam o aluno cursar estatística por um período foi uma das quais se mostraram a princípio significativas para a variável de interesse *classe*, entretanto a utilização dela no modelo não faz sentido por demandar essa espera. Outra variável que mostrou significância através do p-valor foi a *ano_mat*.

Realizar a verificação de interações entre as variáveis também é fundamental para garantir que o modelo consiga ter o melhor desempenho com os dados utilizados. Foram testadas interações entre as variáveis *disret_KM* x *cotas*, *cor* x *cotas*, *ensinomedio* x *cotas*, *nota* x *cotas*, *ano_mat* x *ano_nasc*. Em nenhum dos casos essas interações demonstraram ser estatisticamente significativas para o modelo.

O modelo com todas as colunas exceto tempo cursado, pois como foi dito, essa variável não se aplicaria no contexto do modelo.

Tabela 7 - Valores Estimados no Modelo de Regressão Logística.

Variável	$\widehat{\beta}$	$DP(\widehat{\beta})$	Log-Ver	p-valor
Constante	-929.474	240.017	-83.364	
Sexo	1.027	0.421		0.014
DisRetaKM	0.004	0.007		0.483
DistKM	-0.004	0.006		0.457
TipoIngresso PD	13.288	882.744		0.987
TipoIngresso TI	-0.666	1.069		0.533
CorA	-1.976	1.0444		0.058
CorN	0.697	1.3222		0.597
CorPA	0.005	0.737		0.993
CorPR	-0.532	1.108		0.630
EnsinoMedio	0.127	0.468		0.785
Cotas EP	-1.969	0.941		0.036
Cotas PPI	-2.013	1.035		0.051
ClassVest	0.001	0.019		0.947
NotaVest	-0.023	0.567		0.679
Ano_nasc	-1.466	0.607		0.015
Ano_mat	1.926	0.659		0.003
Idade	-1.256	0.596		0.035

Diferente da análise univariada, outras variáveis demonstraram ser importantes para o modelo também. Para obter um bom modelo com menos características e que consiga descrever melhor ou igualmente a classe dos alunos, foi aplicado no RStudio o método stepwise, onde sugeriu um novo modelo de regressão logística, contendo as varáveis *sexo*, *cotas*, *ano_nasc*, *ano_mat* e *idade*. Entretanto, foi optado por testar separadamente *idade* e *ano_nasc*, visto que estão relacionadas.

O primeiro modelo gerado a partir do conjunto de treino, obteve os seguintes valores estimados:

Tabela 8 - Valores Estimados no Modelo Sugerido.

Variável	$\widehat{\beta}$	$DP(\widehat{\beta})$	Log-Ver	p-valor
Constante	-669.423	170.322	-88.117	
Sexo M	0.875	0.391		0.025
Cotas EP	-1.374	0.775		0.076
Cotas PPI	-1.394	0.722		0.053
ano_mat	0.469	0.124		0.000
ano_nasc	-0.138	0.082		0.092

A partir do modelo de estimação de probabilidade, temos:

$$\hat{\pi}(x) = \frac{e^{\hat{g}(x)}}{1 + e^{\hat{g}(x)}} \tag{5.1}$$

onde $\hat{g}(x)$ é dado por:

$$\hat{g}(x) = -669.423 + 0.875\, sexoM - 1.374\, cotasEP - 1.394\, cotasPP + 0.469\, ano\_mat - 0.138\, ano\_nasc \tag{5.2}$$

Um segundo modelo foi utilizado para comparação trocando a variável *ano_nasc* com a i*dade*, apesar de serem similares é possível ter uma pequena diferença entre elas devido ao mês de nascimento do aluno.

Tabela 9 - Valores Estimados no Segundo Modelo

Variável	$\widehat{\boldsymbol{\beta}}$	$\boldsymbol{DP(\widehat{\beta})}$	Log-Ver	p-valor
Constante	-654.529	168.818	-88.584	
Sexo M	0.836	0.387		0.030
Cotas EP	-1.290	0.766		0.092
Cotas PPI	-1.364	0.721		0.058
Ano_mat	0.323	0.083		0.000
Idade	0.112	0.080		0.161

$$\hat{g}(x) = -654.529 + 0.836\, sexoM - 1.290\, cotasEP - 1.364\, cotasPPI + 0.323\, ano\_mat + 0.112\, idade \tag{5.3}$$

No software R o método de resposta da previsão utilizado indicava a probabilidade daquele aluno evadir e foi definido como ponte de corte caso o valor seja maior que 72%, o aluno era classificado como “evadido”, caso a probabilidade fosse menor que esse valor seria classificado como “formado”.

Como exemplo do cálculo e classificação serão utilizados dois alunos aplicados ao primeiro modelo, com a variável ano de matrícula.

Quadro 13 – Informações de alunos criados para exemplo.

Aluno	Sexo	Cotas	Ano Matrícula	Ano Nascimento
1	Masculino	Não	2015	1993
2	Masculino	E. Publico	2014	1995

Utilizando as codificações das variáveis nos quadros (7-12), a probabilidade de cada aluno evadir, com base no modelo de regressão logística proposto é dado por:

$$\hat{g}(1) = -669.423 + 0.875\,(1) - 1.374\,(0) - 1.394\,(0) + 0.469\,(2015) - 0.138\,(1993)$$

$$\hat{g}(1) = 1.453$$

$$\hat{\pi}\,(1) = \frac{e^{\hat{g}(1)}}{1 + e^{\hat{g}(1)}} = \frac{e^{1.453}}{1 + e^{1.453}} = 0.8143$$

Podemos concluir que o aluno 1 seria classificado como “evadido” devido a sua probabilidade ser de aproximadamente 81% de chance de ser um aluno evadido, valor que é superior ao de corte (72%) estabelecido.

Para o segundo aluno têm-se:

$$\hat{g}(2) = -669.423 + 0.875\,(1) - 1.374\,(1) - 1.394\,(0) + 0.469\,(2014) - 0.138\,(1995)$$

$$\hat{g}(2) = -0.666$$

$$\hat{\pi}\,(2) = \frac{e^{\hat{g}(2)}}{1 + e^{\hat{g}(2)}} = \frac{e^{\hat{g}(1)}}{1 + e^{\hat{g}(1)}} = 0.3393$$

Em contraste com o primeiro, esse segundo aluno possui 33% de chance de evadir, sendo então classificado no grupo de “formado”. É notável que apesar da categoria da variável sexo ser a mesma nos dois alunos, as outras variáveis possuem informações diferentes, que unidas já modificaram o valor da probabilidade calculada.

Com o conjunto de teste foi realizado a previsão e comparação dos resultados estimados com os verdadeiros através da tabela de contingência.

Tabela 10 – Tabela de Preditos X Reais do Modelo 1.

	Evadido	**Formado**	**T**
Evad (P)	4	6	10
For (P)	18	37	55
T	22	43	65

O modelo acertou aproximadamente 63%, considerando o método de APER, temos a partir da tabela 10, como 24 (18+6) o total de alunos classificados erroneamente, dividindo esse valor pelo total da amostra de teste, têm-se que 37% representa a taxa de erro do modelo.

Tabela 11 – Tabela de Preditos X Reais do Modelo 2.

	Evadido	**Formado**	**T**
Evad (P)	4	5	9
For (P)	18	38	56
T	22	43	65

O segundo modelo, possuindo a variável *idade* ao invés de *ano_matricula,* não apresentou tanta diferença ao primeiro, acertando uma classe de aluno a mais e com taxa de acerto de 64,5% aproximadamente.

Com objetivo de comparar o desempenho dos modelos propostos, foi realizado a plotagem da curva ROC com os dados de treino e obtido a métrica de AUC.

Figura 9 – Gráfico da Curva ROC para o modelo 1.

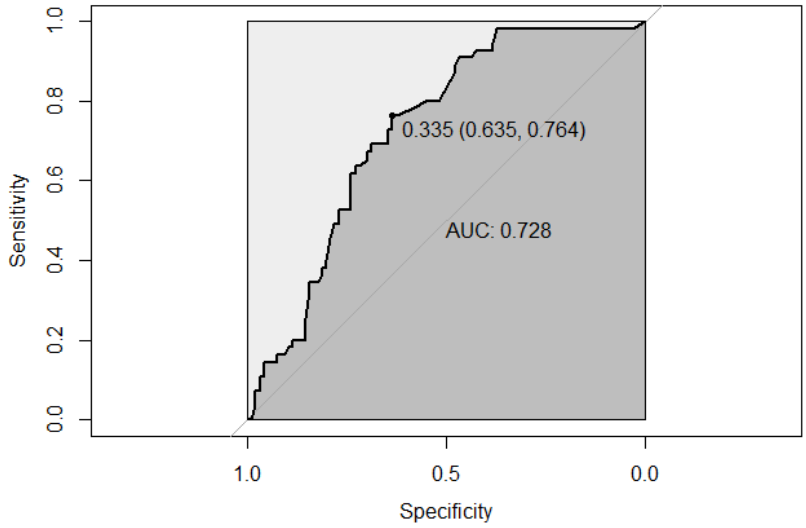
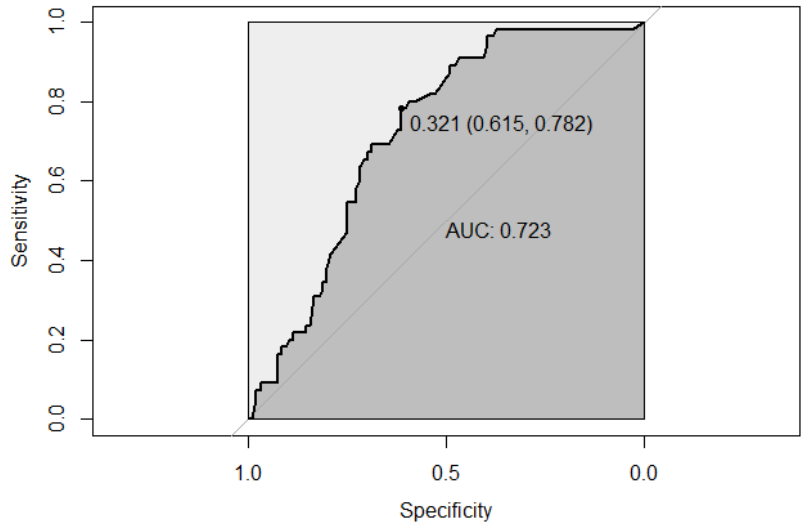


Figura 10 – Gráfico da Curva ROC para o modelo 2.



Os gráficos apresentam grande similaridade, com valores bem próximos em ambos. Através do valor de AUC pode-se concluir que o modelo tem um bom desempenho, pois os valores obtidos para o modelo 1 e 2 são 0,728 e 0,723 respectivamente, são próximos de 1. Sendo a partir dessa métrica definido o ponte de corte na classificação para 0,72.

Um terceiro modelo de regressão logística foi considerado devido a variável *tempo_cursado* se mostrar importante em resultados anteriores.

$$\hat{g}(x) = -482.899 + 0.894\ sexo - 1.139\ cotasEP - 1.240\ cotasPPI + 0.322\ idade + 0.240\ ano\_nasc - 0.319\ tempo\_cursado \quad (5.4)$$

Aplicando no conjunto de teste foi obtido os seguintes resultados sobre as previsões:

Tabela 12 – Tabela de Preditos X Reais do Modelo 3.

	Evadido	**Formado**	**T**
Evad (P)	5	2	7
For (P)	17	41	58
T	22	43	65

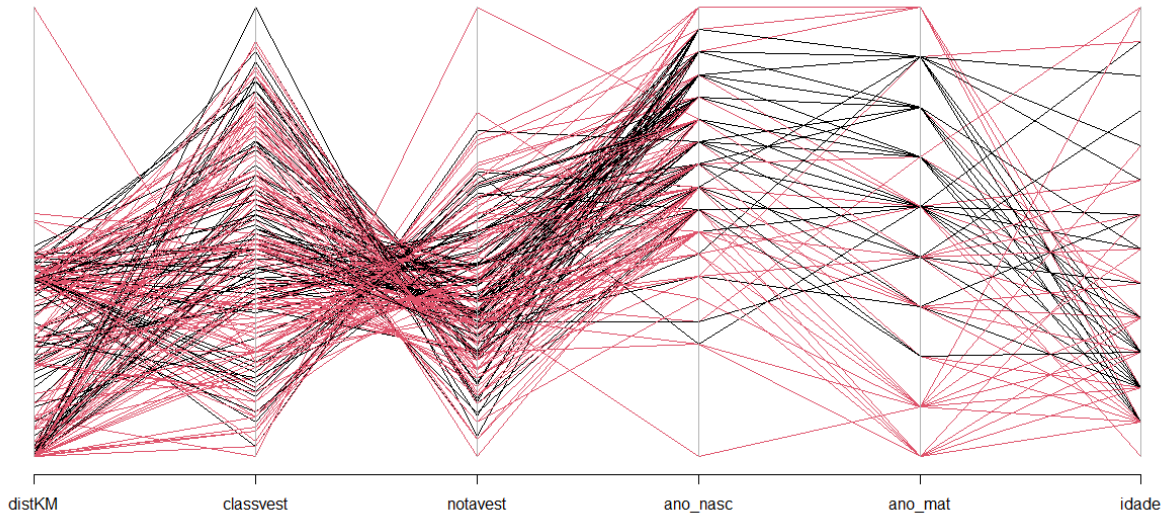
Utilizando o mesmo ponto de corte nota-se uma melhora na classificação dos alunos com taxa de 70% de acerto, entretanto devido à complexidade de usar uma variável que demanda um tempo de retorno é preferível continuar com o primeiro modelo sendo considerado o melhor dos apresentados.

5.4 Análise Discriminante

A técnica multivariada de Análise Discriminante só é possível ser aplicada em variáveis numéricas, sendo reduzido o número de colunas a serem trabalhadas na criação desse modelo. Serão consideradas seis variáveis *distKM*, *classvest*, *notavest*, *ano_nasc*, *ano_mat* e *idade*.

Através das linhas do Gráfico de perfis é possível observar as ligações entre uma característica e outra, além de pela cor identificar o grupo de interesse.

Figura 11 – Gráfico de perfis.



Com as informações que temos dos estudantes, não é possível notar uma separação muito clara entre o grupo de evadidos e formandos, apenas a variável *ano_mat* demonstra ter uma concentração do grupo de formados mais concentrado nas extremidades.

Apesar da técnica não pressupor a normalidade dos dados, ela implica esse requisito ao necessitar testar a significância do modelo depois. Como relato na tabela (3), através do teste de Shapiro Wilk, conclui-se que os dados não possuem uma distribuição normal.

Existem transformações que podem ser aplicadas nas variáveis para conseguir essa normalidade. No total foram testadas três técnicas, Normalize e Box-Cox.

Tabela 13 – P-valor do teste de Aderência nos Dados Transformados.

	Normalize	**Box-Cox**
distKM	0.000	0.000
notavest	0.004	0.438
classvest	0.018	0.030
ano_nasc	0.000	0.000
ano_mat	0.000	0.000
idade	0.000	0.000

O método de cálculo de cada transformação é dado pelas seguintes fórmulas:

$$z = \frac{x-\mu}{\sigma} \quad \text{e} \quad y^{(\lambda)} = \begin{cases} \frac{y^{\lambda}-1}{\lambda}, & \lambda \neq 0 \\ \log y, & \lambda = 0 \end{cases} \tag{5.5}$$

A única variável que apresentou normalidade após a transformação foi a de nota dos alunos no vestibular, utilizando o método de Box-cox. Os valores de λ estimados para cada variável contemplavam o intervalo de [-5;5].

Devido ao resultado obtido de não normalidade, mesmo após a aplicação de métodos de transformações, acaba sendo inviável continuar com um modelo de Análise Discriminante para a estimação da classe dos estudantes de estatística.

6 Conclusão

Em relação ao resultado de projetos anteriores como o de Aragão (2020) e Alves (2018), neste estudo foi apresentado diferentes variáveis que se mostraram significativas na previsão da situação do aluno. No estudo de Aragão as características de classificação no vestibular, tipo de ensino médio cursado e a utilização de cotas estavam no modelo, entretanto, no estudo atual elas não foram consideradas importantes para a estimação da classe com exceção de cotas. Sendo significativas as variáveis de sexo, cotas, ano de primeira matrícula, idade/ano de nascimento.

Considerando as hipóteses testadas foi obtido na realização do teste de diferença de médias que não há diferença significante entre os grupos de alunos que evadem e os formandos em relação a rede de ensino médio, utilização de cotas no vestibular e a cidade de origem dos alunos. Também foi obtido que a nota do vestibular não demonstrou ser significante na previsão da classe dos alunos.

Com a técnica de Regressão Logística foi possível ter um aproveitamento melhor das informações disponibilizadas para compor o banco de dados, apesar de não demonstrar um desempenho tão acurado, foi permitido sua aplicação. Em contrapartida, a Análise Discriminante apresentou diferentes impeditivos, tanto na escassez de informações, quanto na transformação delas em dados com normalidade. Apesar do esforço não foi possível realizar a classificação pela técnica.

É possível concluir que devido a flexibilidade do modelo logístico, além da utilização de diferentes tipos de variáveis, comparado a análise discriminante, a técnica de regressão logística se mostrou mais eficiente para esse conjunto, utilizando as cinco variáveis já citadas.

Para estudos futuros creio que seja importante a busca de outras informações que possam enriquecer a análise, é possível que outros fatores consigam descrever melhor a evasão dos alunos, visto que há inconstância nas variáveis consideradas importantes entre os estudos realizados até o momento.

Referências

ALVES, M.A. Modelo de Regressão Logística para Dados de Evasão em Cursos de Graduação. 2018. Trabalho de Conclusão de Curso (Graduação em Estatística) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, 2018.

ANUÁRIO ESTATÍSTICO 2021. Universidade Estadual Paulista “Júlio de Mesquita Filho”. vol. 21 – São Paulo: Unesp, APE, 2021. Disponível em: https://www2.unesp.br/portal%23!/anuario/. Acesso em: 20 jan. 2022.

ARAGÃO, P.H. de. Estudo da Evasão de Alunos de Graduação Utilizando Educational Data Mining. 2020. Trabalho de Conclusão de Curso (Graduação em Estatística) - Faculdade de Ciências e Tecnologia, Universidade Estadual Paulista, Presidente Prudente, 2020.

CAMPOS, H. Estatística experimental não-paramétrica. 4 ed. Piracicaba: Esalq (USP), 1983.

FÁVERO, L.P.L.; BELFIORE, P.P.; SILVA, F.L. da; CHAN, B.L. Análise de dados: modelagem multivariada para tomada de decisões. [S.l: s.n.], 2009.

FIGUEIRA, C.V. **Modelos de regressão logística**. Universidade Federal do Rio Grande do Sul. Porto Alegre; 2006. Disponível em: https://www.lume.ufrgs.br/handle/10183/8192. Acesso em: 22 jan. 2022.

FILHO, R.L.L. e S.; MOTEJUNAS, P.R.; HIPÓLITO, O.; LOBO, M. B. de C.M. A Evasão no Ensino Superior Brasileiro. **Cadernos de Pesquisa**, [*s. l.*], v. 37, ed. 132, p. 641-659, set/dez 2007. Disponível em: https://www.scielo.br/j/cp/a/x44X6CZfd7hqF5vFNnHhVWg/?lang=pt.%20. Acesso em: 20 jan. 2022.

GIBSON (1998) apud TEIXEIRA, R.P.; MENTGES, M.J.; KAMPFF, A.J.C. Evasão no Ensino Superior: Um Estudo Sistemático. **Apresentação em Evento,** out 2019. Disponível em: https://repositorio.pucrs.br/dspace/bitstream/10923/15080/2/Evasao_no_Ensino_Superior_um_Estudo_Sistematico.pdf. Acesso em: 20 jan. 2022

GONZALEZ, L. de A. **Regressão Logística e suas Aplicações**. 2018. Trabalho de Conclusão de curso (Curso de Graduação em Ciência da Computação) - São Luís, 2018. Disponível em: https://monografias.ufma.br/jspui/bitstream/123456789/3572/1/LEANDRO-GONZALEZ.pdf. Acesso em: 20 jul. 2022.

HOSMER, D.W. JR; LEMESHOW, S; STURDIVANT, R.X. **Applied Logistic Regression**. 3. ed. [*S. l.*: *s. n.*], 2013

INEP. **Censo da Educação Superior 2019** – Notas Estatísticas. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2020/N

otas_Estatisticas_Censo_da_Educacao_Superior_2019.pdf. Acesso em: 20 jan. 2022.

JOHNSON, R.A.; WICHERN, D.W. **Applied multivariate statistical analysis**. 6th ed. Upper Saddle River, New Jersey: Prentice-Hall, 2007, 773 p.

LUZ, M.R. da; KUFNER, M.F. Comparação entre Analise Discriminante e Regressão Logística como método de melhor resultado na verificação de fatores que influenciam a evasão de alunos do Curso de Estatística da UFPR. Trabalho de Conclusão de Curso - UFPR, Curitiba, 2008. Disponível em: http://www.leg.ufpr.br/lib/exe/fetch.php/disciplinas:ce229:tcc2008_marcelo_marcos.pdf. Acesso: em 20 jan. 2022.

MARIN, V.P. Estudo sobre o rendimento dos alunos do curso de Estatística da FCT/UNESP. 2018. Trabalho de Conclusão de Curso (Graduação em Estatística) - Faculdade de Ciências e Tecnologias, Universidade Estadual Paulista, Presidente Prudente, 2018.

MORAES, R.B.N. de; MELO, C.G. Evasão no Ensino Superior: uma revisão de literatura em psicologia e educação. **Psicologia - Saberes & Práticas**, v.1, n.2, p. 83-91, 2018. Disponível em: https://www.unifafibe.com.br/revistasonline/arquivos/psicologiasaberes&praticas/sumario/64/16012019154350.pdf. Acesso em: 20 jan. 2022.

SILVESTRE, M.R. Notas de aula de Análise Multivariada II. 2021. Presidente Prudente: [s. n.], 2021. Apostila disponível no ambiente virtual de apoio disciplina Análise Multivariada II da FCT/Unesp.

TACHIBANA, V.M. Notas de aula de Regressão Logística. 2021. Presidente Prudente: [s. n.], 2021. Apostila disponível no ambiente virtual de apoio disciplina Regressão Logística da FCT/Unesp.

VARELLA, C.A.A. Análise Discriminante. Análise Multivariada Aplicada as Ciências Agrárias Pós-Graduação em Agronomia Ciência do Solo: CPGA-CS. Tese. Disponível em: http://www.ufrrj.br/institutos/it/deng/varella/Downloads/multivariada%20aplicada%20as%20ciencias%20agrarias/Aulas/ANALISE%20DISCRIMINANTE.pdf. Acesso em: 20 jan. 2022.