UNIVERSIDADE FEDERAL DE OURO PRETO
INSTITUTO DE CIÊNCIAS EXATAS E BIOLÓGICAS
DEPARTAMENTO DE ESTATÍSTICA
BACHARELADO EM ESTATÍSTICA

Modelo de detecção de risco de evasão nos programas de pós-graduação da UFOP

Tiago Alves de Morais

Ouro Preto-MG

2025

Tiago Alves de Morais

Modelo de detecção de risco de evasão nos programas de pós-graduação da UFOP

Monografia de Graduação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas e Biológicas da Universidade Federal de Ouro Preto como requisito parcial para a obtenção do grau de bacharel em Estatística.

Orientador: Helgem de Souza Ribeiro Martins

Coorientador: Anderson Ribeiro Duarte

Ouro Preto

2025

SISBIN - SISTEMA DE BIBLIOTECAS E INFORMAÇÃO

M827m Morais, Tiago Alves de.
Modelo de detecção de risco de evasão nos programas de pós-graduação da UFOP. [manuscrito] / Tiago Alves de Morais. - 2025.
23 f.: il.: gráf., tab..

Orientador: Prof. Dr. Helgem de Souza Ribeiro Martins.
Coorientador: Prof. Dr. Anderson Ribeiro Duarte.
Monografia (Bacharelado). Universidade Federal de Ouro Preto. Instituto de Ciências Exatas e Biológicas. Graduação em Estatística .

1. Educação- Estatística. 2. Aprendizado do computador. 3. Evasão universitária. 4. Controle preditivo. 5. Modelos Logísticos. I. Martins, Helgem de Souza Ribeiro. II. Duarte, Anderson Ribeiro. III. Universidade Federal de Ouro Preto. IV. Título.

CDU 519.2

Bibliotecário(a) Responsável: Soraya Fernanda Ferreira e Souza - SIAPE: 1.763.787

MINISTÉRIO DA EDUCAÇÃO
UNIVERSIDADE FEDERAL DE OURO PRETO
REITORIA
INSTITUTO DE CIENCIAS EXATAS E BIOLOGICAS
COLEGIADO DO CURSO DE ESTATISTICA

FOLHA DE APROVAÇÃO

Tiago Alves de Morais

Modelo de detecção de risco de evasão nos programas de pós-graduação da UFOP

Monografia apresentada ao Curso de Estatística da Universidade Federal
de Ouro Preto como requisito parcial para obtenção do título de Bacharel em Estatística

Aprovada em 28 de março de 2025

Membros da banca

Dr. Helgem de Souza Ribeiro Martins - Orientador (Universidade Federal de Ouro Preto)
Dr. Anderson Ribeiro Duarte - Coorientador (Universidade Federal de Ouro Preto)
Dr. Josino José Barbosa (Universidade Federal de Ouro Preto)
Dr. Tiago Martins Pereira (Universidade Federal de Ouro Preto)

Professor Dr. Helgem de Souza Ribeiro Martins, orientador do trabalho, aprovou a versão final e autorizou seu depósito na Biblioteca Digital de Trabalhos de Conclusão de Curso da UFOP em 28/03/2025

Documento assinado eletronicamente por **Helgem de Souza Ribeiro Martins**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 31/03/2025, às 16:36, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

Documento assinado eletronicamente por **Tiago Martins Pereira**, **PROFESSOR DE MAGISTERIO SUPERIOR**, em 02/04/2025, às 14:29, conforme horário oficial de Brasília, com fundamento no art. 6º, § 1º, do Decreto nº 8.539, de 8 de outubro de 2015.

A autenticidade deste documento pode ser conferida no site http://sei.ufop.br/sei/controlador_externo.php?acao=documento_conferir&id_orgao_acesso_externo=0 , informando o código verificador **0886039** e o código CRC **C33AF160**.

R. Diogo de Vasconcelos, 122, - Bairro Pilar Ouro Preto/MG, CEP 35402-163
Telefone: (31)3559-1312 - www.ufop.br

Agradecimentos

Em primeiro lugar, gostaria de expressar minha gratidão à minha família e amigos, que me apoiaram incondicionalmente durante toda a minha trajetória acadêmica. Sua presença foi fundamental para fortalecer minha determinação e resiliência nos momentos desafiadores.

Aos professores Helgem e Anderson, direciono meus sinceros agradecimentos pela orientação técnica, paciência e contribuições essenciais ao desenvolvimento desta Monografia em Estatística. Seu conhecimento e disponibilidade foram decisivos para a concretização deste projeto.

Reconheço, ainda, todos os professores e colegas do curso de Estatística que, direta ou indiretamente, enriqueceram minha formação acadêmica. As discussões, colaborações, aprendizados e momentos compartilhados ao longo desta jornada foram indispensáveis para meu crescimento profissional e pessoal.

Resumo

Nas últimas duas décadas, a pós-graduação no Brasil, inclusive a Universidade Federal de Ouro Preto (UFOP), testemunhou um crescimento notável. Na UFOP, o número de programas e cursos de pós-graduação mais do que dobrou nos últimos 10 anos. Esse aumento em programas também trouxe desafios administrativos e acadêmicos, como um aumento na taxa de evasão de estudantes. Este estudo tem como objetivo analisar o comportamento da evasão de alunos de pós-graduação na UFOP, por meio de modelos de *machine learning* para prever o risco de evasão no momento da matrícula. A intenção é fornecer *insights* que permitam a adoção de medidas para reduzir a evasão.

Palavras-chave: Evasão acadêmica, Machine Learning, Modelo Preditivo, Regressão Logística .

Abstract

Over the past two decades, graduate education in Brazil, including at the Federal University of Ouro Preto (UFOP), has experienced significant growth. At UFOP, the number of graduate programs and courses has more than doubled in the last 10 years. However, this expansion has introduced administrative and academic challenges, such as an increase in graduate student dropout rates. This research project aims to analyze the dropout behavior of graduate students at UFOP by employing *machine learning* models to predict dropout risk at the time of enrollment. The objective is to provide actionable insights that enable the implementation of targeted measures to mitigate dropout rates.

Keywords: Student Attrition , Machine Learning, Predictive Model, Logistic Regression.

Lista de ilustrações

Sumário

1 Introdução

A pós-graduação brasileira passou por um processo de expansão e consolidação a partir do término do século passado. Segundo Ambiel *et al.* (2020) [1], o número de programas de pós-graduação no país passou de 1237 no ano de 1998 para um total de 4177 em 2016, ou seja, em pouco menos de 20 anos, o número de programas foi praticamente triplicado. No ano de 2024, segundo a Plataforma Sucupira da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) [2], são 4662 programas de pós-graduação, que totalizam 7317 cursos. Destes, são 3727 cursos de mestrado acadêmico, 855 cursos de mestrado profissional, 2534 cursos de doutorado acadêmico e 127 cursos de doutorado profissional. Em pouco mais de 20 anos, o número de cursos aumentou em aproximadamente 6 vezes, o que obviamente causou crescimento significativo no número de opções de cursos e consequentemente de vagas ofertadas.

No âmbito da Universidade Federal de Ouro Preto, a expansão dos cursos de pós-graduação apresentou forte crescimento na última década. Em relação à 2011, em termos comparativos, houve um crescimento de 66% no número de programas oferecidos, bem como um crescimento de 75% na oferta de cursos de pós-graduação, entre mestrados e doutorados acadêmicos e mestrados profissionais.

A evasão discente na pós-graduação é um problema que afeta a instituição em diversas instâncias. Do ponto de vista institucional, a evasão discente reduz qualidade da avaliação dos programas de pós-graduação frente à CAPES, indicadores que são avaliados a cada quatro anos pela referida agência de fomento. Existe também influência na esfera financeira, pois a redução do número de alunos impacta no montante de recursos recebidos pelas instituições, por meio da matriz de Orçamento e Custeio (OCC), que leva em consideração o número de alunos matriculados e concluintes da graduação e pós-graduação das universidades federais, bem como outros fatores. Estudos anteriores demonstram que o abandono por parte do discente ocorre por questões pessoais e acadêmicas (Castelló *et al*., 2017 [3]). Dentre as questões pessoais, destacam-se questões sócio demográficas, como a necessidade de trabalhar para manter seus custos, a distância de casa, dentre outros motivos (Gardner, 2009 [4]).

Diversos estudos tentam avaliar questões associadas à evasão na pós-graduação brasileira, sobre diversos aspectos. O trabalho de Fernandes *et al.* (2017) [5] teve como objetivo identificar o índice de evasão com base nos dados do Sistema de Informações Georreferenciadas (GEOCAPES) e encontrou resultados expressivos de evasão entre os anos 2000 e 2016. O trabalho de Ambiel *et al.* (2020) [1] utilizou análise fatorial para identificar os principais fatores associados à evasão discente no âmbito da pós-

graduação.

Junior *et al.* (2020) [6] discutiram a temática da evasão avaliada do ponto de vista gerencial. No trabalho foram relacionadas as principais ações governamentais desenvolvidas no Brasil com o intuito de minimizar os índices de evasão na pós-graduação stricto sensu. Outros trabalhos apresentaram avaliações locais dos índices de evasão, com foco em uma única instituição e seus programas e cursos, como os trabalhos de Alves (2018) [7], Pereira *et al.* (2021) [8], dentre outros estudos.

Com a crescente oferta de informações geradas por meio de sistemas de gestão e acompanhamento em diversas instâncias, bem como a constante atualização de dados, a modelagem estatística tem se tornado a cada dia mais dinâmica, por meio de modelos do tipo *data-driven*, associados à ferramentas de Estatística e Ciências de Dados, com destaque para modelos de aprendizado de máquina (*machine learning*).

Neste sentido, a expansão de metodologias estatísticas desenvolvidas para modelar dados na era do *big data* pode contribuir sobremaneira para a compreensão do fenômeno da evasão em programas de pós-graduação, de acordo com a atual disponibilidade de dados, sobretudo por meio da utilização de métodos de aprendizado de máquina, tanto com o objetivo de descrever o fenômeno, quanto prever potenciais evasões do sistema de pós-graduação, para mitigar os efeitos negativos desse fenômeno.

1.1 Objetivos

O objetivo principal é desenvolver um modelo de aprendizado de máquina para detectar discentes com maiores chances de evasão. Os objetivos específicos são listados a seguir:

- Realizar uma análise descritiva bivariada.
- Estudar as técnicas de aprendizado de máquina.
- Desenvolver um modelo preditivo.
- Avaliar os resultados e desempenho do modelo.

2 Material e Métodos

2.1 Dados relacionados à evasão na pós-graduação

Os dados em estudo foram obtidos por meio de solicitação formal à Pró-Reitoria de Pesquisa, Pós-Graduação e Inovação da Universidade Federal de Ouro Preto. As informações foram cedidas devido ao interesse institucional na compreensão dos processos geradores de evasão na pós-graduação.

A base de dados em estudo é composta por dois blocos distintos. O primeiro conta com 16.937 observações e 29 variáveis, e abrange informações demográficas dos estudantes à respeito de dados sobre o curso, informações sobre o processo de admissão, bolsas de estudo, evasão além de outras informações também relevantes. A base de dados compreende informações relacionadas aos estudantes que ingressaram na pós-graduação da UFOP a partir do ano de 1983, até o ano de 2022. Este primeiro bloco será utilizado como base de treinamento e conta apenas com alunos cujo desfecho (concluinte ou evadido) em 2022 já estava definido.

O segundo bloco contém 1304 observações de alunos que em 2022 estavam ativos no sistema e no ano de 2024 já apresentam desfecho, ou seja, já se sabe se concluíram seus respectivos cursos de pós-graduação, ou se evadiram.

Os dados em estudos se classificam como dados estruturados. Segundo Daróczi (2015) [9], dados estruturados são aqueles que apresentam uma estrutura e formato pré-definido antes de ser armazenado em uma base de dados. Geralmente são dados apresentados em um formato tabular, em arquivo único, ou em estruturas de dados mais complexas, como bancos de dados relacionais.

Para a extração das informações de forma eficiente, é necessária a realização de um pré-processamento dos dados. Esta etapa compreende uma série de procedimentos que tornam a estrutura dos dados adequada para a modelagem estatística dos dados. Nessa etapa, são sanados problemas corriqueiros em bancos de dados, tais como a remoção de duplicatas, correção de dados faltantes, dados com ruídos, *outliers*, dentre outros.

Devido à natureza dos dados, não foram encontrados dados duplicados, uma vez que cada aluno possui apenas um registro. Também não foram detectados *outliers*. Em relação à dados faltantes, algumas variáveis os apresentaram. De modo geral, os dados faltantes foram detectados em variáveis que foram criadas no decorrer do tempo, ou seja, foram inseridas após a existência de dados registrados em versões anteriores. Um exemplo ilustrativo é apresentado nos dados relacionados à declaração étnico-racial,

que não era solicitada no ato da matrícula até o ano de 1998. Esses dados faltantes, por estarem ausentes devido a uma razão estrutural (não registrados intencionalmente no período anterior), são classificados como MNAR (*missing not at random*). Embora técnicas de imputação sejam aplicáveis em alguns cenários de MNAR, sua utilização exige cautela, pois a ausência está ligada a mecanismos não observáveis ou a decisões sistemáticas (como a não coleta da informação em determinado período). Nesses casos, imputações podem introduzir vieses se não forem acompanhadas de análises robustas sobre o contexto da ausência ou de modelos que incorporem a natureza do mecanismo de missing.

Outro tipo de dado faltante observado é proveniente de registros nulos. Por exemplo, o principal objeto deste estudo, a evasão escolar, apresenta dados nulos quando o evento não ocorre. Neste caso, os dados, na verdade, não são faltantes, e sim apresentam problemas de classificação. Neste caso, para as variáveis que apresentaram este tipo de comportamento, os dados nulos foram interpretados de acordo com sua natureza. Como exemplo, é possível citar a variável tipo de evasão, cujos níveis são **desligamento**, **jubilamento**, **cancelamento** e **transferência externa**. Para situações em que não ocorre evasão, um valor nulo é atribuído. Para esta variável, os valores nulos foram substituídos pelo valor "não evadido".

Para evitar distorções relacionadas à alterações regimentais e estatutárias relacionadas à pós-graduação, foram considerados dados a partir do ano de 2010 até 2022. Esta segmentação dos dados, além de evitar as referidas distorções tornam o posterior processo de modelagem mais acurado, uma vez que torna a base de dados mais homogênea. A Tabela 1 apresenta as variáveis que compõe o conjunto de dados.

Tabela 1 – Variáveis componentes do banco de dados.

Variável	Descrição
sexo	Variável categórica que indica o gênero do aluno.
cor_pele	Representa a cor da pele do aluno.
data_nascimento	Representa a data de nascimento dos alunos.
pais_nascimento	País de nascimento dos alunos.
estado_nascimento	Estado de nascimento dos alunos.
cidade_nascimento	Cidade de nascimento dos alunos.
cod_curso	Código do curso em que o aluno está matriculado.
curso	Curso que o aluno está matriculado.
nivel	Nível de ensino em que o aluno se encontra ou se formou.
area	Área de conhecimento do curso que o aluno está matriculado.
ano_ingresso	Ano em que o aluno ingressou na instituição.
inicio_atv	Data de início das atividades do aluno na instituição.

Continuação na próxima página

Tabela 1 – Variáveis componentes do banco de dados.

Variável	Descrição
modo_admissao	Tipo de ingresso do aluno no programa de pós-graduação
curso_admissao	Curso no qual o aluno foi matriculado
nivel_ingresso	Nível no qual o aluno foi matriculado
pais	País em que o aluno reside.
estado	Estado em que o aluno reside.
cidade	Cidade em que o aluno reside.
bolsa	Bolsa recebida pelo aluno durante sua trajetória acadêmica.
descricao_bolsa	Descrições das bolsas de estudo concedidas aos alunos.
data_defesa	Data em que o aluno defendeu sua dissertação ou tese.
x_data_evasao	Data em que o aluno evadiu do programa de pós-graduação.
reserva_vaga	Indica se o aluno participa do programa de reservas de vagas.
descricao_nec_esp	Indica se o aluno possui alguma necessidade especial e, se sim, qual a descrição dessa necessidade.
x_meses_de_curso	Número de meses que o aluno ficou matriculado no curso.
ano_evasao	Ano que o aluno evadiu do programa de pós-graduação.
data_evasao	Data que o aluno evadiu do programa de pós-graduação.
motivo_evasao	Motivo da evasão do aluno no programa de pós-graduação.

Neste estudo, foram criadas novas variáveis a partir dos dados originais para enriquecer a análise. Essas variáveis estão apresentadas na Tabela 2, e fornecem informações adicionais que foram úteis para uma compreensão mais abrangente do conjunto de dados.

Tabela 2 – Variáveis criadas no banco de dados.

Variável	Descrição
nacionalidade	Origem do aluno.
idade	Idade do aluno.
tempo_evasao	Indica o período de permanência, em anos, no programa de pós-graduação até a ocorrência de evasão.
evadiu	Indica se o aluno evadiu ou não.
recebeu_bolsa	Indica se o aluno recebeu bolsa ou não.
tipo_bolsa	Indica o tipo de bolsa recebida pelo aluno.
reservou_vaga	Indica se o aluno reservou vaga ou não.
necessidade_especial	Indica se o aluno possui necessidade especial ou não.

2.2 Metodologias Aplicadas

Os objetivos do presente estudo compreende a obtenção de padrões que permitam identificar eventuais fatores que influenciam na decisão de evasão por parte dos discentes de pós-graduação da Universidade Federal de Ouro Preto. Essa identificação se dará por meio de análises bivariadas, onde serão avaliadas as relações entre variáveis, duas a duas. Como o estudo prevê a avaliação de fatores que podem influenciar a ocorrência de evasão, serão consideradas como variáveis resposta aquelas relacionadas à ocorrência de evasão. As demais variáveis sociodemográficas serão consideradas variáveis explicativas, ou seja, serão utilizadas para tentar identificar alguma espécie de padrão amostral que propicie variações significativas na ocorrência de evasão.

Em seguida, será utilizado um modelo de classificação para prever a evasão dos alunos. Modelos de classificação são abordagens que visam prever respostas qualitativas e estimam a probabilidade de uma observação pertencer a cada uma das categorias de uma variável qualitativa. Para este estudo, foi escolhida a Regressão Logística como técnica adequada, uma vez que o objetivo é prever a probabilidade de uma resposta binária, como a ocorrência ou não de evasão, com base em variáveis preditoras. A Regressão Logística utiliza a função logística para garantir que as previsões estejam sempre no intervalo entre 0 e 1.

Para realizar as previsões, o modelo de Regressão Logística será ajustado aos dados de treinamento, com a estimação dos coeficientes por meio do método da máxima verossimilhança. Com os coeficientes estimados, será possível calcular a probabilidade de ocorrência de evasão para cada aluno.

A avaliação do desempenho do modelo será feita por meio da validação cruzada *k-fold*, uma técnica amplamente utilizada para garantir uma estimativa robusta da capacidade de generalização dos modelos preditivos. Na validação *k-fold*, os dados disponíveis são divididos em k subconjuntos (ou *folds*) de tamanho aproximadamente igual. Em cada uma das k iterações, um desses subconjuntos é utilizado como conjunto de teste, enquanto os outros $k - 1$ subconjuntos são utilizados para treinar o modelo. Esse processo é repetido k vezes, de forma que cada subconjunto seja utilizado exatamente uma vez como conjunto de teste.

Ao final das k iterações, as métricas de desempenho obtidas em cada rodada são agregadas, resultando em uma estimativa mais precisa da eficácia do modelo. Essa abordagem ajuda a reduzir o viés e a variância associados a uma única divisão de treino/teste, o que proporciona uma avaliação mais estável e confiável do modelo. Neste estudo, foi utilizada a validação *k-fold* com k igual a 10, o que garante que todas as observações fossem utilizadas tanto no treino quanto no teste. Essa metodologia permitirá a determinação do melhor ponto de decisão para prever a evasão, por meio

da consideração diferentes valores de probabilidade de corte.

2.3 Métricas utilizadas

Nessa fase da pesquisa, foram discutidas algumas métricas para avaliar e comparar a performance do modelo. A ideia é definir qual indicador é mais apropriado para nosso modelo.

Acurácia (ACC): calcula a proporção de previsões corretas. Considera tanto as classificações verdadeiras positivas quanto as verdadeiras negativas. Embora amplamente utilizada, sua eficácia é limitada em conjuntos de dados desbalanceados, pois pode mascarar a incapacidade do modelo em identificar corretamente a classe minoritária.

Precisão (PREC): reflete a proporção de verdadeiros positivos (TP) entre todas as classificações positivas. Um alto valor de precisão indica uma baixa taxa de falsos positivos, o que é fundamental em cenários onde é importante evitar alarmes falsos.

Recall (REC) ou Sensibilidade: mede a capacidade do modelo em identificar corretamente todos os exemplos positivos. Altos valores de *recall* são preferíveis em contextos nos quais a detecção de casos positivos é crítica, como em diagnósticos médicos.

Especificidade (SPEC): avalia a proporção de verdadeiros negativos (TN) corretamente classificados. Isso é especialmente relevante quando a minimização de falsos positivos é priorizada.

F1-Score: média harmônica entre precisão e *recall*. É útil para equilibrar essas duas métricas em cenários em que ambas são igualmente importantes. O *F1-Score* é particularmente vantajoso em situações de classes desbalanceadas, em que é necessário avaliar o desempenho do modelo ao considerar tanto a capacidade de identificar a classe positiva quanto a de evitar falsos alarmes.

Coeficiente de Correlação de Matthews (MCC): métrica robusta para avaliar o desempenho de classificadores binários. Ele mede a correlação entre as classificações previstas e reais. A principal vantagem do MCC em relação a outras métricas é que ele leva em consideração todos os elementos da matriz de confusão (verdadeiros positivos - TP, verdadeiros negativos - TN, falsos positivos - FP e falsos negativos - FN), de maneira equilibrada. Diferentemente de métricas que dependem da classe positiva, o MCC trata de forma igualitária ambas as classes, sendo, portanto, menos sensível a decisões arbitrárias de qual classe é definida como positiva.

Esses métodos permitem uma análise abrangente do desempenho de modelos. Oferecem diferentes perspectivas de acordo com as características do conjunto de

dados e as necessidades específicas do problema em estudo. A escolha do indicador apropriado deve considerar o equilíbrio entre as diferentes métricas, com especial atenção ao contexto da aplicação e às implicações dos erros de classificação.

3 Resultados Alcançados

Nos dados analisados, observou-se uma prevalência de evasão de 15%, o que corresponde a aproximadamente quinze em cada cem estudantes que abandonaram o curso antes da conclusão. A Figura 1 ilustra a distribuição proporcional entre estudantes que evadiram e os que permaneceram no curso, destacando um desbalanceamento na nossa base de dados, indicando um possível problema que pode impactar negativamente no resultado do modelo preditivo.

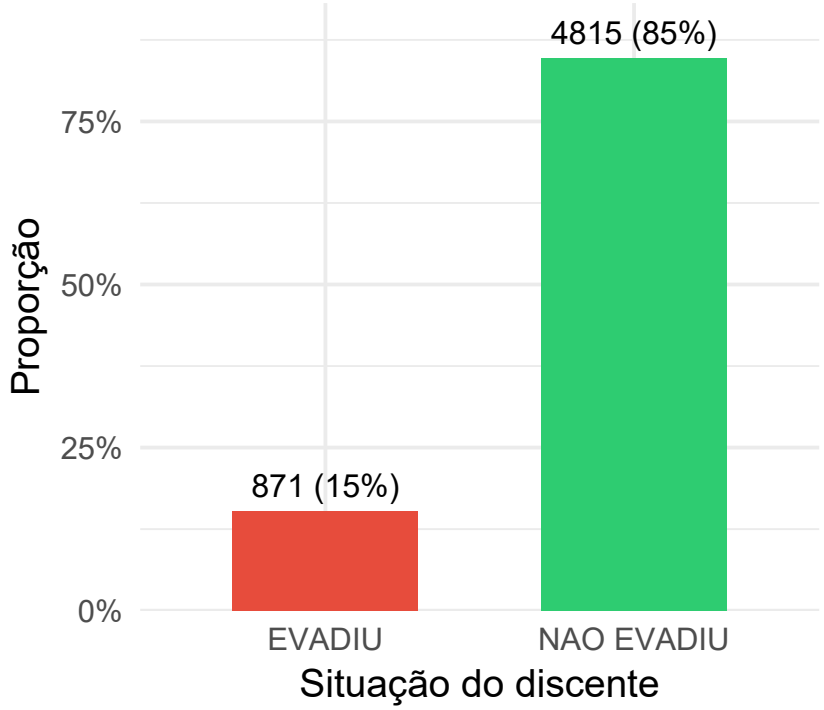

Figura 1 – Proporção geral de evasão

A Figura 2 compara a proporção de evasão entre os dois blocos de dados, de treinamente e de teste, revelando uma distribuição equilibrada: 85% no treinamento e 80% no teste. Essa semelhança sugere que a divisão estratificada preservou a representatividade da variável resposta, evitando vieses na avaliação do modelo.

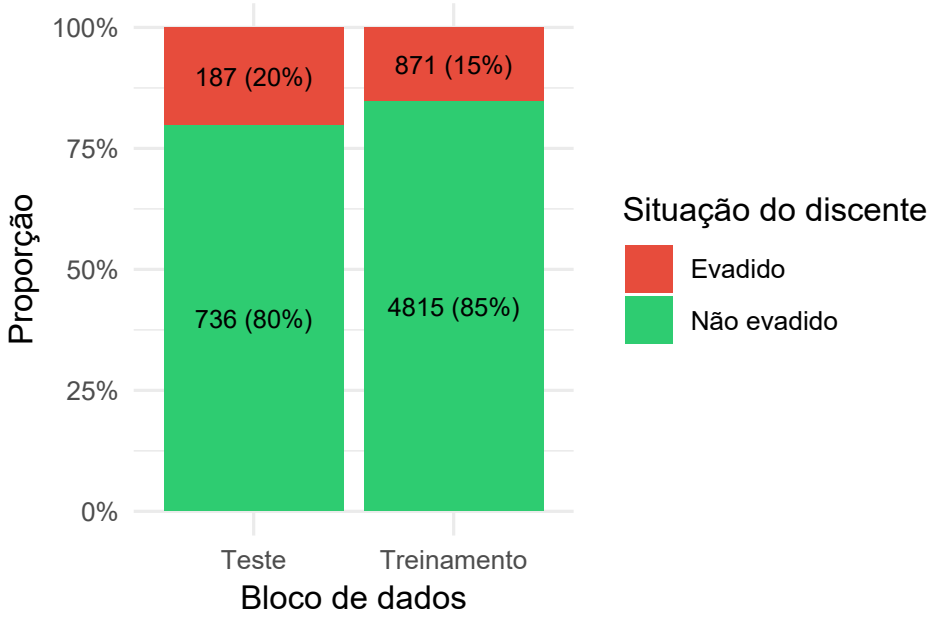

Figura 2 – Proporção de evasão nos blocos de dados

3.1 Análise Descritiva com Cruzamento entre Variáveis

Nessa etapa do estudo, foi realizada a análise descritiva bivariada em relação às principais características dos discentes no que diz respeito à evasão ou não. A análise descritiva univariada dessas variáveis, que explora a distribuição individual de cada característica (como frequências absolutas, proporções e medidas de tendência central), foi detalhada no estudo de Dias *et al.* (2023) [10], oferecendo uma compreensão preliminar da amostra e das variáveis envolvidas. Enquanto a abordagem univariada permite identificar padrões isolados, por exemplo, a proporção de estudantes que evadiram, a análise bivariada adotada neste trabalho busca examinar associações entre as variáveis independentes, como gênero, faixa etária ou renda familiar, e o desfecho de evasão, visando identificar possíveis relações ou disparidades significativas.

Por meio de gráficos de barras que estabelecem relações entre cada variável e a ocorrência de evasão é possível identificar os resultados mais significativos para a análise. Esse processo permite a busca por uma compreensão mais aprofundada das circunstâncias e motivações subjacentes às evasões observadas.

Ao considerar a análise dos fatores de evasão, é importante concentrar-se nos motivos mais prevalentes a fim de obter resultados estatisticamente significativos e embasados. As linhas pontilhadas apresentadas nos gráficos possuem a função de delimitar os valores esperados em caso de ausência de tendência. O primeiro cruzamento associa as variáveis sexo e motivo de evasão.

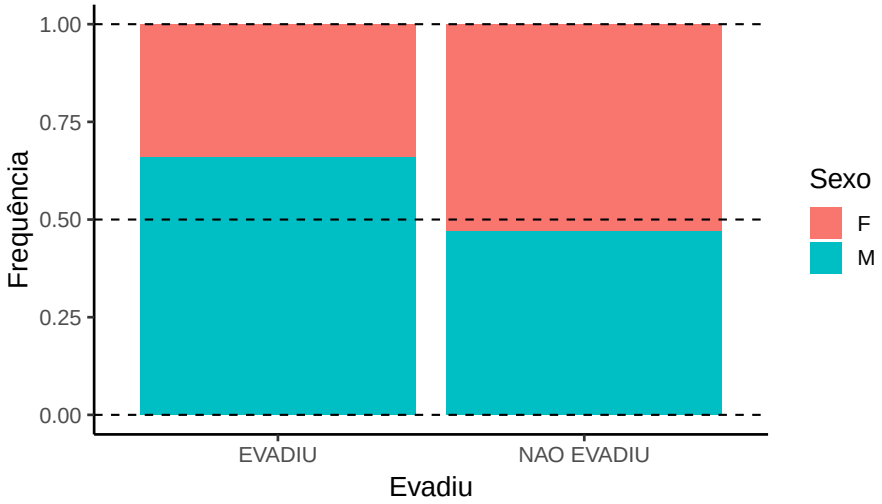

Figura 3 – Frequências absolutas da evasão por sexo

Ao observar a Figura 3, constata-se que o sexo masculino tem destaque como o gênero com a maior evasão. Isso é evidenciado pela altura da barra masculina. Nota-se que entre os não evadidos, existe um equilíbrio bastante evidente entre os dois gêneros, o que reforça a existência de um maior risco de evasão associado ao sexo masculino, de modo geral.

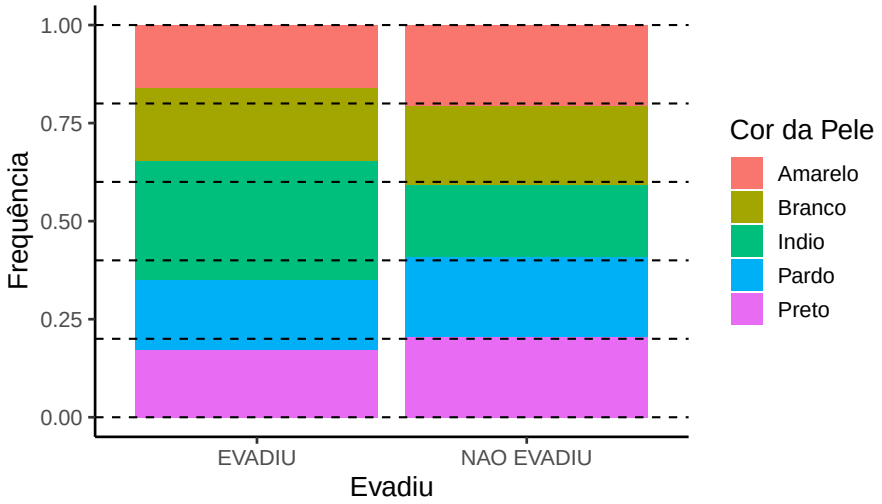

Figura 4 – Frequências absolutas da evasão por cor da pele.

Ao analisar as proporções de evasão por cor da pele, observa-se padrões particulares. A proporção de alunos com cor de pele indígena está acima do esperado. Entre os não evadidos, as proporções de diferentes cores de pele estão equilibradas, o que ressalta um maior risco de evasão associado aos alunos de cor de pele indígena. Este efeito pode ser observado na Figura 4. Cabe ressaltar que a prevalência de alunos indígenas é inferior às demais.

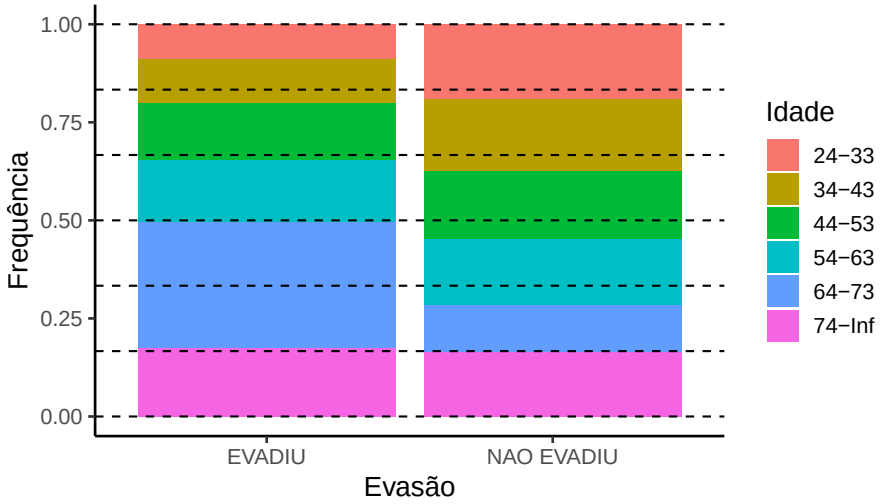

Figura 5 – Frequências absolutas da evasão por idade.

Em relação à faixa etária, apresentada na Figura 5, percebe-se uma tendência de evasão mais elevada em alunos de faixas etárias mais avançadas, sobretudo a partir dos 64 anos. No caso dos alunos que não evadiram, as proporções são mais homogêneas, o que sugere que à medida em que as faixas etárias aumentam, os alunos se tornam mais propensos a evadir.

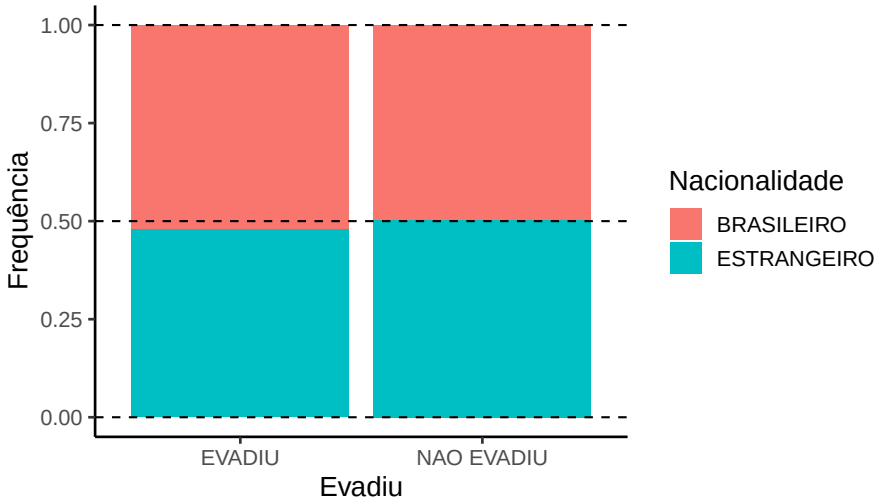

Figura 6 – Frequências absolutas da evasão por país

A Figura 6 indica que existe um equilíbrio perceptível relacionado à evasão em relação à nacionalidade.

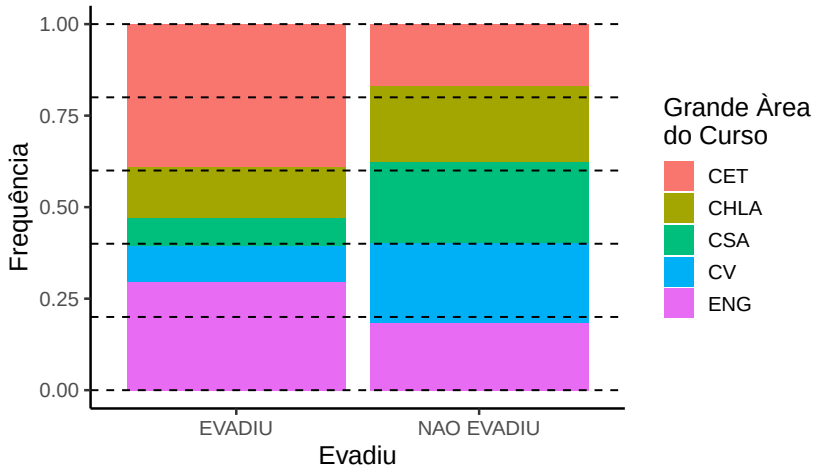

Figura 7 – Frequências absolutas da evasão pela grande área do curso.

Ao analisar a Figura 7, torna-se evidente que as áreas de Ciências Exatas e da Terra e de Engenharias possuem proporções de evasão significativamente maiores do que o esperado. Por outro lado, as áreas de Ciências Humanas, Letras e Artes, e Ciências da Vida exibem proporções menores do que o esperado.

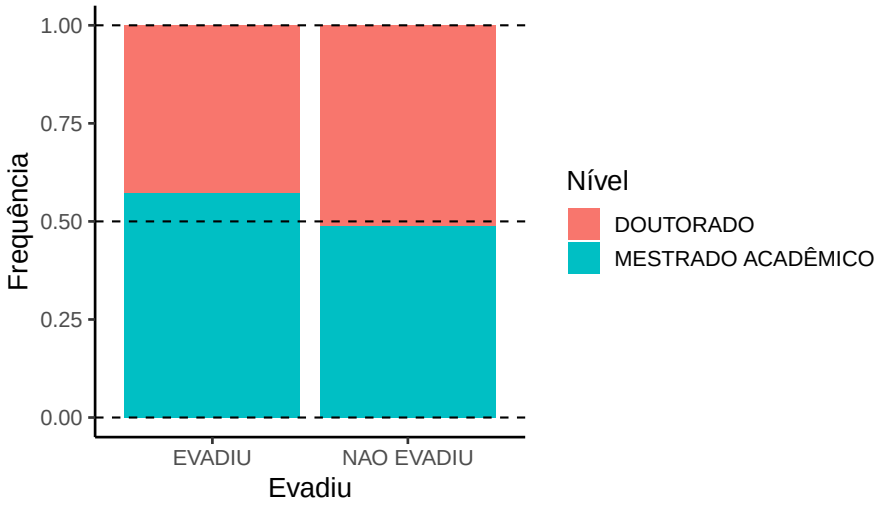

Figura 8 – Frequências absolutas da evasão por nível.

Por meio da Figura 8, pode-se observar que os alunos de Mestrado Acadêmico se destacam como o nível com a maior evasão. É notável que entre os alunos que não evadiram, existe um equilíbrio bastante evidente entre os dois níveis, o que reforça a existência de um maior risco de evasão associado ao nível de Mestrado Acadêmico, de modo geral.

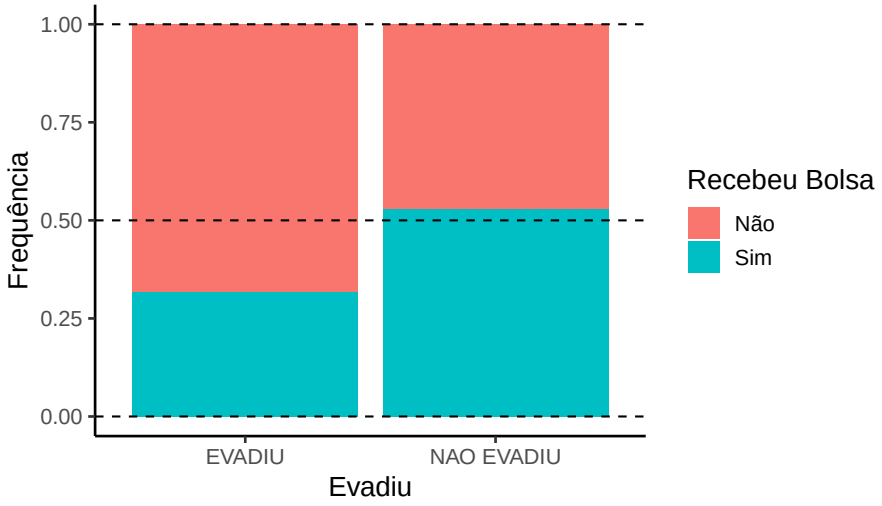

Figura 9 – Frequências absolutas da evasão por alunos que não receberam bolsa.

Ao analisar o cruzamento das variáveis por meio da Figura 9, pode se constatar que os alunos que não receberam bolsas se destacam como os mais propensos à evasão. Por outro lado, entre os alunos que não evadiram, existe um equilíbrio bastante evidente entre aqueles que receberam bolsas e aqueles que não receberam, o que reforça a existência de um maior risco de evasão associado àqueles que não receberam bolsas.

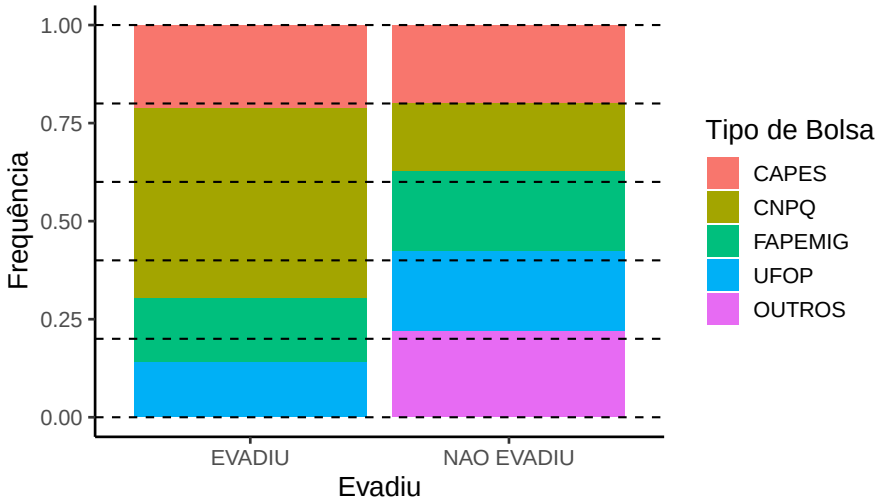

Figura 10 – Frequências absolutas da evasão por tipo de bolsa.

A Figura 10 indica que entre os alunos que receberam bolsas, é interessante observar que a maior parte das evasões ocorreu entre os beneficiários da bolsa CNPq, seguidos pela bolsa CAPES. Os beneficiários da bolsa CNPq apresentaram uma proporção maior que o dobro em evasão em comparação com os outros grupos. Em relação aos alunos que não evadiram é perceptível o equilíbrio entre os tipos de bolsas.

3.2 Regressão Logística

Com a conclusão das análises bivariadas, o passo subsequente é o desenvolvimento de um modelo de Regressão Logística com intuito de possibilitar a predição de eventuais candidatos à evasão. Entretanto, com objetivo de alcançar o ajuste do modelo, da maneira mais parcimoniosa possível para atender ao problema, são retiradas as variáveis não significativas para o estudo.

A análise descritiva inicial demonstrou que as variáveis cor da pele e nacionalidade não apresentaram significância, conforme os testes estatísticos. Os resultados obtidos dessas duas categorias podem não ser estatisticamente suficientes para generalizar padrões ou tendências. Identificou-se também que uma pequena proporção, 1,25% dos alunos presentes nos dados, estão presentes na categoria reserva de vaga. Além disso, apenas duas pessoas apresentam necessidade especial, com 2,82% da reserva de vaga. As baixas representações destas categorias são insuficientes para contribuir com o modelo de Regressão Logística.

Destaca-se que, para o ajuste do modelo de regressão logística, a variável faixa etária foi recategorizada. Os dados foram agrupados em faixas de discentes até 29 anos, de 30 a 39 anos, de 40 a 49 anos e alunos com 50 anos ou mais.

A Tabela 3 apresenta os detalhes de ajuste do modelo.

Variáveis	Estimativa	Razão de Chances	2,5 %	97,5 %	P-valor
Sexo: Masculino	0,4633	1,5893	1,3552	1,8658	$0{,}00\times10^{0}$
Faixa Etária: 30 - 39 anos	0,6326	1,8825	1,3928	2,5941	$6{,}46\times10^{-5}$
Faixa Etária: 40 - 49 anos	0,8060	2,2390	1,6154	3,1534	$2{,}20\times10^{-6}$
Faixa Etária: 50+ anos	1,1150	3,0496	2,0303	4,6084	$1{,}00\times10^{-7}$
Área: Ciências Humanas e Linguísticas e Artes	-1,2481	0,28701	0,2278	0,3604	$0{,}00\times10^{0}$
Área: Ciências Sociais Aplicadas	-1,9455	0,1429	0,0921	0,2136	$0{,}00\times10^{0}$
Área: Ciências da Vida	-1,3090	0,2701	0,2045	0,3536	$0{,}00\times10^{0}$
Área: Engenharias	-0,4350	0,6472	0,5355	0,7824	$6{,}80\times10^{-6}$
Nível: Mestrado Acadêmico	0,6025	1,8267	1,4895	2,2540	$0{,}00\times10^{0}$
Recebeu Bolsa: Sim	-0,7431	0,4756	0,3919	0,5743	$0{,}00\times10^{0}$

Tabela 3 – Estimativas e demais informações para o ajuste do Modelo Geral.

Os resultados apontam que os discentes do sexo masculino têm uma chance aproximadamente 59% maior de evadir, se comparados com discentes do sexo feminino, a categoria de base. Também é perceptível que a probabilidade de evasão aumenta significativamente com a idade. Alunos mais velhos (particularmente aqueles com mais de 50 anos) são mais propensos a abandonar os estudos, com 3.04 vezes mais chance de evasão, seguidos pela faixa de 40 a 49 anos com chance 2.23 vezes maior e por aqueles

de 30 a 39 anos que possuem uma chance de evadir 88% maior, sempre comparados com a categoria de referência, que é a faixa etária de 20 a 29 anos.

Em relação à grande área, se comparadas à área de Ciências Exatas e da Terra, que é a categoria de referência, estudantes nas áreas de Ciências Sociais Aplicadas têm uma chance de 86% menor de evadir, seguido por Ciências da Vida com 73% e Ciências Humanas e Linguísticas e Artes 72%. Já os alunos de Engenharias apresentam uma chance 36% menor de evadir.

Em relação ao nível do curso, alunos matriculados em programas de Mestrado Acadêmico têm uma chance 82% maior de evadir, em relação aos estudantes de Doutorado. Isso pode indicar a necessidade de maior suporte para este nível, que frequentemente enfrentam maiores pressões acadêmicas e financeiras. O recebimento de bolsa de estudos pode ser percebido como um fator protetor contra a evasão. Discentes bolsistas tem uma chance 53% menor de abandono. Isso destaca a importância de apoio financeiro na retenção de estudantes.

3.3 Modelo Preditivo

Uma vez ajustado o modelo de regressão logística, este serviu como base para um modelo preditivo. Para a determinação de um ponto de corte eficiente, foi utilizada a metodologia de validação cruzada *k-fold*, com 10 dobras, conforme descrito na metodologia. Com o objetivo de ajustar o critério de decisão, foi implementado um loop que verifica todos os valores de probabilidade (p) no intervalo entre 0 e 0,5, em incrementos de 0,01, com potenciais candidatos a ponto de corte. Esse processo permitiu identificar o melhor ponto de corte para maximizar a performance do modelo. Para cada valor de p, as métricas de desempenho do modelo foram calculadas.

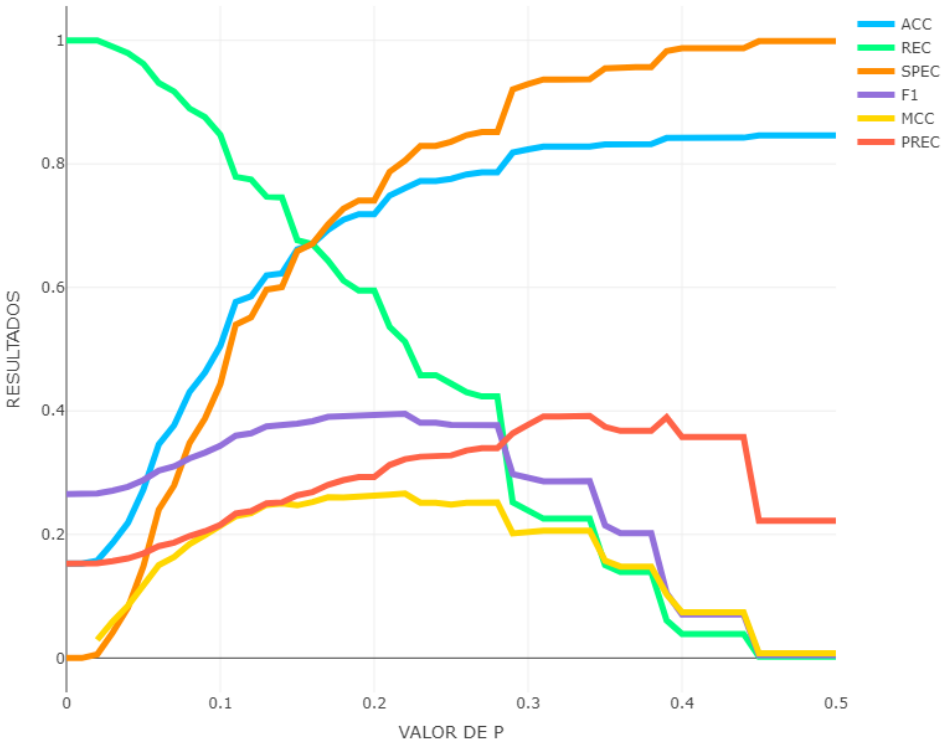

Figura 11 – Métricas

A Precisão foi a métrica escolhida para avaliar o desempenho do modelo preditivo. Essa escolha se baseia na importância de maximizar os verdadeiros positivos. Ao priorizar a precisão, asseguramos que os alunos identificados como propensos à evasão realmente apresentem uma alta probabilidade de deixar o curso. Isso é crucial, pois é preferível alertar um aluno com baixa chance de evasão do que deixar de identificar um aluno com alta chance de desistir. Um alerta incorreto pode ser abordado de forma construtiva, enquanto uma falha em identificar um aluno em risco pode resultar em consequências negativas significativas, como a evasão real. Assim, a ênfase na precisão contribui para intervenções mais eficazes e direcionadas.

A precisão média foi obtida ao longo das 10 dobras, de acordo com os resultados

apresentados no gráfico a seguir, o valor da probabilidade (p) que maximizou a precisão do modelo foi de 0,32.

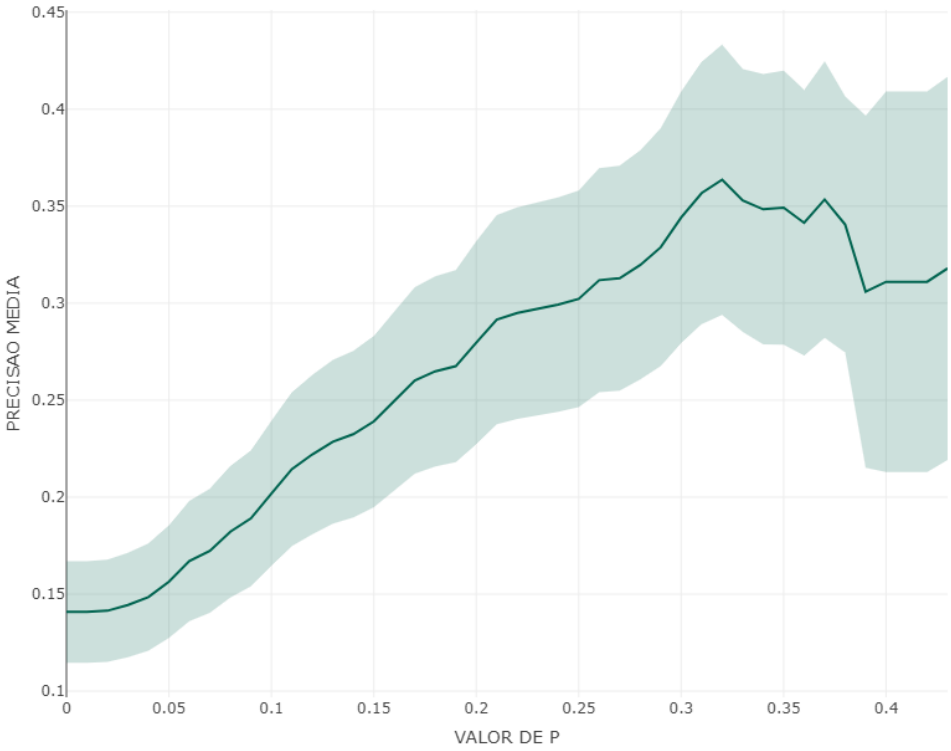

Figura 12 – Precisão Média com Intervalo de Confiança

Após a definição do ponto de corte ideal (p = 0,32), faz-se necessário testar a capacidade preditiva do modelo em novos dados para avaliar sua eficácia. Para esse fim, o modelo foi ajustado utilizando dados de discentes que, no momento do treinamento, ainda estavam ativos no sistema e, portanto, não participaram da análise inicial. Esses alunos ingressaram até 2022 e tiveram sua situação acadêmica definida até o final de 2024, permitindo assim a avaliação do modelo. A Tabela 4 apresenta os detalhes de ajuste do modelo.

Variáveis	Estimativa	Razão de Chances	2,5 %	97,5 %	P-valor
Sexo: Masculino	-0,0715	0,2440	-0,2930	0,7694	
Faixa Etária: 30 - 39 anos	0,2328	0,2950	0,7890	0,4300	
Faixa Etária: 40 - 49 anos	1,1130	0,3830	2,9059	0,0036	
Faixa Etária: 50+ anos	0,9200	0,7250	1,2689	0,2044	
Área: Ciências Humanas, Linguísticas e Artes	-1,0122	0,4539	-2,2299	0,0257	
Área: Ciências Sociais Aplicadas	-1,7773	0,5730	-3,1016	0,0019	
Área: Ciências da Vida	-2,8832	1,0963	-2,6297	0,0085	
Área: Engenharias	-0,3435	0,3753	-0,9151	0,3600	
Nível: Mestrado Acadêmico	0,0741	0,3294	0,2251	0,8218	
Recebeu Bolsa: Sim	-0,8122	0,3467	-2,3424	0,0191	

Tabela 4 – Modelo ajustado para os novos dados.

Após o ajuste do modelo aos novos dados, foi realizada a estimativa da probabilidade de evasão para cada aluno individualmente. Com base nessas estimativas, aplicou-se o ponto de corte previamente definido, que serviu como limite para a classificação dos alunos, o que permitiu categorizar cada indivíduo como evadiu ou não evadiu. Com os novos dados classificados, foi gerada uma tabela de contingência, na qual se compararam as predições realizadas com os resultados observados.

	PREVISÃO		
	Evadiu	Não evadiu	
Evadiu	28	159	187
Não evadiu	25	711	736
	53	870	923

Tabela 5 – Tabela de acertos

A análise da Tabela 5 demonstrou que o modelo obteve um bom desempenho. Com 77,03% nas predições de permanência e 3,03% nas evasões, o modelo foi capaz de avaliar corretamente cerca de 80,06% dos alunos. Observou-se, no entanto, algumas limitações, especialmente na capacidade de prever corretamente os alunos que evadiram, mas que foram classificados como não evadiu e no número de falsos positivos gerados. Especificamente, os falsos negativos representam 17,23% do resultado total, enquanto os falsos positivos chegaram a 2,71%. Esses resultados indicam que, embora o modelo seja eficaz na identificação dos alunos que permaneceram, ele pode subestimar o risco de evasão, e portanto gerar predições incorretas de evasão e permanência.

MÉTRICAS			
Acurácia	Precisão	Recall	Especificidade
0,8006	0,5283	0,1497	0,9660

Tabela 6 – Tabela de Métricas

Com a tabela de contingência, foram calculadas as métricas e avaliamos o desempenho do modelo. A Acurácia é uma métrica que reflete a proporção de predições corretas, tanto positivas quanto negativas, em relação ao total de observações. No caso do modelo em estudo, uma acurácia de 0,8006 indica que ele foi capaz de classificar corretamente cerca de 80% dos casos, ou seja, a soma dos verdadeiros positivos (VP) e verdadeiros negativos (VN) em relação ao total de predições. A Precisão é a proporção de verdadeiros positivos entre todas as predições positivas feitas pelo modelo. Diante disso, dos casos classificados como “evasão”, aproximadamente 52,83% foram de fato evasões reais.

Esse valor de precisão, ainda que superior a 50%, sugere que o modelo comete uma quantidade relativamente alta de falsos positivos (FP). Isso indica que o modelo classifica como "evadiu"um número significativo de estudantes que, na verdade, não evadiram. O *recall*, também conhecido como sensibilidade, mede a proporção de verdadeiros positivos, evasões corretamente identificadas, em relação ao total de casos reais de evasão. O valor de *recall* de 14,97% sugere que o modelo tem uma baixa capacidade de identificar os alunos que realmente evadiram. A Especificidade, por sua vez, indica a capacidade do modelo de identificar corretamente os casos negativos, ou seja, os alunos que não evadiram. Um valor de 96,60% demonstra que o modelo é altamente eficaz em classificar corretamente os alunos que permanecem na pós-graduação.

4 Considerações Finais

A partir da análise de diversas variáveis relacionadas ao perfil dos estudantes e às características dos cursos, foi desenvolvido um modelo de regressão logística que identificou padrões associados à evasão, que possibilitou a construção de um modelo preditivo.

Os resultados mostram que o modelo preditivo teve uma boa performance ao prever a permanência e a efetiva evasão dos alunos, com uma taxa de acerto de 80,06% nesse grupo. Entretanto, a análise da tabela de contingência revelou algumas limitações, e apresentou uma taxa de falsos positivos de 2,71% e falsos negativos de 17,23%.

Apesar da alta acurácia e especificidade, a baixa sensibilidade e precisão moderadamente indicam que, embora o modelo tenha um desempenho geral satisfatório, há espaço para melhorar sua capacidade de prever a evasão corretamente. Isso indica a necessidade de ajustes, como a otimização dos limites de decisão ou o uso de outras técnicas para melhorar a detecção de alunos em risco de evasão.

Este estudo representa um passo importante na compreensão dos fatores que influenciam a evasão em programas de pós-graduação, mas ainda há caminhos a serem explorados. Futuras pesquisas poderiam ampliar a análise incluindo, por exemplo, entrevistas ou questionários com os estudantes, para capturar aspectos subjetivos, como suas percepções sobre a orientação recebida ou a adequação da estrutura do curso. Testar o modelo em outras universidades brasileiras também seria relevante para verificar se os padrões identificados na UFOP se repetem em contextos institucionais distintos. Além disso, seria interessante adotar técnicas que tornem as decisões do modelo mais transparentes, facilitando a interpretação dos resultados por gestores e docentes. Por fim, a universidade poderia usar os alertas gerados pelo modelo para propor ações concretas, como programas de mentoria adaptados às necessidades individuais dos estudantes, criando assim um ciclo virtuoso entre pesquisa e prática educacional.

Referências

[1] Ambiel, Rodolfo Augusto Matteo, Ariela Raissa Lima Costa, Ana Deyvis Santos Araújo Jesuíno, Camila Cardoso Camilo e Samanta Romanin Zuchetto: *Motivos de evasão na pós-graduação no Brasil: um instrumento de medida*. Interação em Psicologia, 24(1), 2020. Citado na página 1.

[2] CAPES, COORDENAÇÃO DE APERFEIÇOAMENTO DE PESSOAL DE NÍVEL SUPERIOR.: *Plataforma Sucupira*. <https://sucupira.capes.gov.br/sucupira/>, Acesso em: 30 jun. 2022. Citado na página 1.

[3] Castelló, Montserrat, Marta Pardo, Anna Sala-Bubaré e Núria Suñé-Soler: *Why do students consider dropping out of doctoral degrees? Institutional and personal factors*. Higher Education, 74:1053–1068, 2017. Citado na página 1.

[4] Gardner, Susan K: *Student and faculty attributions of attrition in high and low-completing doctoral programs in the United States*. Higher Education, 58:97–112, 2009. Citado na página 1.

[5] Fernandes, Eduardo Francisco, Andressa Sasaki Vasques Pacheco, Fernanda Cristina da Silva, Thiago Luiz de Oliveira Cabral e Viviane Santos Círio de Azevedo: *Panorama do fenômeno da evasão discente na pós-graduação: uma análise a partir do Geocapes*. Em *Anais do XVII Colóquio Internacional de Gestão Universitária - Mar del Plata/ARG*, páginas 1–16, 2017. Citado na página 1.

[6] Santos Junior, José da Silva, Ana Maria da Silva Magalhães e Giselle Cristina Martins Real: *A gestão da evasão nas políticas educacionais brasileiras: Da graduação à pós-graduação stricto sensu*. ETD Educação Temática Digital, 22(2):460–478, 2020. Citado na página 2.

[7] Alves, José Eduardo Viana: *Evasão e permanência dos alunos nos cursos de pós-graduação Lato Sensu online e presenciais da Fundação Getulio Vargas-FGV*. PhD dissertation, Fundação Getulio Vargas-FGV, 2018. Citado na página 2.

[8] Pereira, Marcelo Almeida de Camargo, Luciane de Fátima Giroto Rosa e Vera Lucia Felicetti: *Evasão do curso de pós-graduação em gestão de negócios de Universidade corporativa: percepções do estudante/empregado*. Educação: Teoria e Prática, 31(64), 2021. Citado na página 2.

[9] Daróczi, Gergely: *Mastering data analysis with R*. Packt Publishing Ltd, 2015. Citado na página 3.

[10] Dias, Ana Júlia Guimarães: *Análise de padrões de evasão em programas de pós-graduação da UFOP.*, 2023. Citado na página 10.