Análise comparativa de modelos de aprendizagem automática para a previsão da permanência de estudantes de graduação presenciais na UFMT

DANIEL VALENTINS DE LIMA[I]
ANDERSON CASTRO SOARES DE OLIVEIRA[II]
EILSON CASTRO SOARES DE OLIVEIRA[III]
http://dx.doi.org/10.22347/2175-2753v16i52.4358

Resumo
Este estudo visa comparar modelos de regressão logística, árvores de decisão e redes neurais na classificação da permanência dos estudantes nos cursos presenciais da UFMT. Foram analisados dados das matrículas de estudantes que iniciaram a graduação presencial nos anos de 2016, 2017 e 2018, nos cinco *campi* da universidade. Os resultados indicam desempenhos semelhantes entre os modelos, com a regressão logística apresentando maior acurácia e sensibilidade. Além disso, a regressão logística demonstrou maior flexibilidade na definição de um limiar para a classificação dos estudantes. Essas informações podem subsidiar políticas de ingresso e permanência na UFMT, visando a melhoria do ensino, permanência e conclusão dos estudantes.
Palavras-chave*:* Ensino superior; Evasão; Mineração de Dados; Modelagem estatística.

Submetido em: 04/08/2023
Aprovado em: 20/08/2024

[I] Universidade Federal do Mato Grosso (UFMT), Cuiabá (MT), Brasil; https://orcid.org/0000-0001-7583-2635; e-mail: dvalentins@outlook.com.
[II] Universidade Federal do Mato Grosso (UFMT), Cuiabá (MT), Brasil; https://orcid.org/0000-0001-6222-9300; e-mail: anderson.oliveira@ufmt.br.
[III] Instituto Federal de Educação, Ciência e Tecnologia de Mato Grosso (IFMT), Cuiabá (MT), Brasil; https://orcid.org/0000-0001-5853-2104; e-mail: eilson.oliveira@ifmt.edu.br.

Comparative analysis of machine learning models for evaluation of permanence in presential undergraduate students at UFMT

Abstract
This study aims to compare logistic regression, decision trees, and neural networks in classifying the permanence of students in UFMT's on-campus courses. Data from student enrollments who began their undergraduate studies in 2016, 2017, and 2018 across the university's five campuses were analyzed. The results indicate similar performances among the models, with logistic regression showing higher accuracy and sensitivity. Additionally, logistic regression demonstrated greater flexibility in defining a threshold for student classification. This information can support UFMT's admission and retention policies, aiming to improve teaching, retention, and student graduation.
Keywords: Higher education; University dropout; Data Mining; Statistical modeling.

Análisis comparativo de modelos de aprendizaje automático para la predicción de la permanencia de estudiantes de grados presenciales en la UFMT

Resumen
Este estudio tiene como objetivo comparar modelos de regresión logística, árboles de decisión y redes neuronales en la clasificación de la permanencia de los estudiantes en los cursos presenciales de la UFMT. Se analizaron datos de las matrículas de estudiantes que iniciaron sus estudios de pregrado en los años 2016, 2017 y 2018 en los cinco *campus* de la universidad. Los resultados indican desempeños similares entre los modelos, con la regresión logística mostrando mayor precisión y sensibilidad. Además, la regresión logística demostró mayor flexibilidad en la definición de un umbral para la clasificación de los estudiantes. Esta información puede respaldar las políticas de admisión y permanencia en la UFMT, buscando mejorar la enseñanza, la permanencia y la conclusión de los estudiantes.
Palabras clave: Educación superior; Deserción; Minería de Datos; Modelado estadístico.

Introdução

A Universidade Federal de Mato Grosso (UFMT) foi fundada em 10 de dezembro de 1970, a partir da fusão de duas instituições existentes no estado: a Faculdade de Direito, existente desde 1934, e o Instituto de Ciências e Letras de Cuiabá, criado em 1966, constituindo assim a primeira universidade pública de Mato Grosso sediada na capital. Suas atividades iniciaram em 1972 com doze cursos, entre eles Ciências Contábeis, Direito, Economia, Engenharia Civil, Física, Geografia, História Natural, Letras, Matemática, Pedagogia, Química e Serviço Social (Dorileo, 1977, 2005).

Ao longo dos anos, a UFMT passou por uma expansão significativa, aumentando o número de cursos e vagas no *Campus* Cuiabá e promovendo a interiorização com a oferta de cursos na modalidade parcelada, a distância e com turmas especiais. Adicionalmente, quatro *campi* foram estruturados fora da sede: Rondonópolis, Araguaia, Sinop e Várzea Grande. Desde 2018, o *Campus* de Rondonópolis foi transformado na Universidade Federal de Rondonópolis (UFR) (Universidade Federal de Mato Grosso, 2020).

No contexto das políticas públicas destinadas à expansão universitária, a UFMT experimentou um crescimento expressivo entre 2006 e 2018. Influenciada pelos programas federais Expandir e REUNI, a universidade ampliou o número de cursos presenciais de 66 para 106 e dobrou o número de vagas, de 3.048 para 6.036. Esse aumento significativo na oferta de vagas, embora represente um avanço no acesso ao ensino superior, contrasta com o modesto crescimento dos concluintes, que foi de apenas 16%, passando de 1.893 para 2.191 (Universidade Federal de Mato Grosso, 2018). Esse descompasso entre o número de vagas e concluintes destaca a necessidade de investigar mais profundamente a permanência e conclusão dos cursos na UFMT, refletindo um fenômeno comum a outras universidades federais, conforme observado no aumento nacional de 23,2% nas vagas do ensino superior entre 2010 e 2019 (Malange; Nogueira; Zardo, 2021).

De acordo com Malange, Nogueira e Zardo (2021), a expansão do ensino superior público possibilitou o ingresso de uma parcela significativa de estudantes de "escolas públicas e com vulnerabilidade social". É necessário, portanto, desenvolver ações que investiguem e promovam a permanência e conclusão de curso desses estudantes. A evasão e a retenção, problemas não recentes no ensino superior, têm causas multifatoriais, incluindo motivos financeiros, psicológicos, ambientais e de

integração acadêmica (percepção do desempenho acadêmico), entre outros (Coimbra; Silva; Costa, 2021; Nascimento; Dantas; Castro; Queiroz, 2024; Santos Junior; Real, 2019; Teixeira; Quito, 2021).

No intuito de compreender melhor esse fenômeno, vários estudos vêm sendo desenvolvidos utilizando a mineração de dados. A natureza desses estudos se divide entre a avaliação de desempenho de diferentes modelos de predição ou a classificação dos fatores de evasão e/ou permanência. Alguns estudos buscam, ainda, propor sistemas de alerta sobre o risco de evasão (Carrano; Albergaria; Infante; Rocha, 2019; Dutra; Souza; Fernandes, 2022; Jesus; Rodrigues; Costa Junior, 2021; Lanes; Alcântara, 2018; Teodoro; Kappel, 2020).

Este trabalho visa comparar modelos de regressão logística, árvores de decisão e redes neurais, aplicados especificamente aos registros da UFMT, na classificação da permanência dos estudantes em cursos presenciais. A pesquisa se destaca pela comparação direta de diferentes abordagens preditivas para identificar a mais eficaz. Inspirada em métodos de estudos anteriores, a abordagem integra técnicas em uma estrutura prática voltada para aprimorar as políticas de permanência na instituição. Esta estratégia enriquece as políticas de ingresso e retenção, oferecendo à comunidade acadêmica dados cruciais sobre as necessidades de públicos específicos e subsidiando decisões sobre programas destinados a melhorar o ensino e a conclusão dos cursos.

Assim, a expansão das universidades federais brasileiras, enquanto aumenta o acesso ao ensino superior, apresenta o desafio de não apenas atrair, mas também reter estudantes até a conclusão de seus cursos. A UFMT, como um estudo de caso, reflete esse fenômeno nacional, mostrando um crescimento substancial no número de vagas e matrículas, mas com taxas de conclusão que não acompanham na mesma proporção. Este cenário levanta questões críticas sobre os fatores que influenciam a permanência dos estudantes. Portanto, a questão de pesquisa que guia este trabalho é: "Como os modelos de regressão logística, árvores de decisão e redes neurais podem ser comparativamente avaliados em sua capacidade de prever a permanência dos estudantes na UFMT, e quais informações podem ser derivadas para informar e melhorar o combate a retenção?" A hipótese central deste estudo é que a aplicação desses modelos preditivos, se adequadamente calibrada e interpretada, pode fornecer compreensões valiosas sobre os padrões de

permanência, subsidiando decisões mais eficazes em políticas de ingresso e retenção na universidade.

Referencial teórico

Neste trabalho, o referencial teórico concentra-se nos métodos de classificação supervisionada, essenciais para analisar e prever fenômenos complexos, como a permanência estudantil. Inicialmente, aborda-se a classificação supervisionada, seguida pelo treinamento de classificadores e pela avaliação do desempenho dos modelos, detalhando como os modelos são ajustados e avaliados. Posteriormente, discutem-se as técnicas especificamente utilizadas, como regressão logística, árvores de decisão e redes neurais, destacando suas funcionalidades e aplicabilidade.

Classificação supervisionada

Classificação supervisionada consiste em um conjunto de métodos que tem como objetivo descobrir o relacionamento existente entre atributos de entrada $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ e atributos de saída Y. O relacionamento encontrado é representado em uma estrutura denominada modelo, e estes modelos geralmente descrevem e explicam fenômenos subjacentes no conjunto de dados e podem ser utilizados para prever valores futuros conhecendo-se alguns valores observados. A ideia dos classificadores supervisionados é a de mapear o espaço de entrada em classes predefinidas (Izbicki; Santos, 2020; Rokach; Maimon; Shmueli, 2023).

Na classificação supervisionada é comum dividir o conjunto de dados em dois grupos: treinamento e teste. O conjunto de treinamento é a parte dos dados utilizada para criar o classificador, e o conjunto de teste é utilizado para avaliar a performance do classificador. Essa divisão é feita por meio de amostragem aleatória, na qual é especificada a proporção de observações desejada para cada novo conjunto. É comum que 75% dos dados pertençam ao conjunto de treinamento, e os 25% restantes pertençam ao conjunto de teste (Alpaydin, 2020).

Treinamento de classificadores

O treinamento de um classificador consiste nos ajustes dos parâmetros internos do modelo ou algoritmo utilizado. Durante esta etapa, deseja-se obter os parâmetros que maximizem a performance do classificador, ou seja, sua capacidade de

predição. Assim, durante o treinamento é comum utilizar técnicas de validação cruzada para avaliar o desempenho do classificador e ajustar os seus parâmetros.

A validação cruzada *k-fold* separa os dados de treinamento em k grupos distintos, onde cada grupo é chamado de *fold*. Ao gerar cada par, mantém-se uma das K partes fora como conjunto de validação e se faz a combinação das demais $K-1$ partes como sendo o conjunto de treinamento (Alpaydin, 2020). Fazendo-se isso K vezes, e em cada vez deixando outra parte de fora, obtêm-se K pares, conforme equação 1.

$$\begin{array}{ll} Y_1 = X_1 & \Gamma_1 = X_2 \cup X_3 \cup \ldots \cup X_k \\ Y_2 = X_2 & \Gamma_1 = X_2 \cup X_3 \cup \ldots \cup X_k \\ \vdots & \vdots \\ Y_k = X_k & \Gamma_k = X_1 \cup X_2 \cup \ldots \cup X_{k-1} \end{array} \quad (1)$$

Os resultados de todos os julgamentos, um para cada observação do conjunto de dados, serão submetidos à média, e essa representará a estimativa final do erro. Há duas vantagens em adotar este procedimento. A primeira é que se garante o maior número possível de dados para o conjunto de treinamento, e a segunda é que se trata de um processo determinístico, não envolvendo nenhuma amostragem aleatória. As desvantagens estão no custo computacional, pois todo o procedimento deverá ser executado K vezes e também o fato de não poder ser estratificado (Izbicki; Santos, 2020; Rokach; Maimon; Shmueli, 2023).

Avaliação do desempenho dos modelos

A avaliação de um modelo de classificação pode ser realizada por meio de várias métricas, como a probabilidade de pertencer a uma classe de saída $Y = 1$ ou obtidas de uma tabela chamada de matriz de confusão. A probabilidade de um indivíduo pertencer a uma classe de saída $Y = 1$ é obtida a partir das características de entrada X indicadas pelo modelo (Alpaydin, 2020; James; Witten; Hastie; Tibshirani, 2013; Rokach; Maimon; Shmueli, 2023). Neste estudo, esta probabilidade será denominada escore de permanência, ou seja, a probabilidade de o estudante não evadir.

Conjuntamente ao escore de permanência, serão utilizadas métricas a partir da matriz de confusão. No caso da classificação binária, é uma matriz contendo quatro cenários: dois cenários são aqueles em que o modelo acertou a predição, e

os outros dois aqueles em que o modelo errou a predição (Tabela 1) (James; Witten; Hastie; Tibshirani, 2013; Izbicki; Santos, 2020).

Tabela 1 – Matriz de confusão

Predito	**Real**	
	Evento	**Não Evento**
Evento	VP – Verdadeiro Positivo	FP – Falso Positivo
Não Evento	FN – Falso Negativo	VN – Verdadeiro Negativo

Fonte: Izbicki e Santos (2020).

A partir da matriz de confusão (tabela 1), podem ser obtidas métricas de desempenho de classificadores. A primeira delas, a acurácia, representa a proporção de acertos do classificador, ou seja, o total de verdadeiros positivos e verdadeiros negativos em relação à amostra estudada, como apresentado na equação 2 (James; Witten; Hastie; Tibshirani, 2013; Izbicki; Santos, 2020).

$$Acurácia = \frac{VP + VN}{VP + VN + FP + FN} \quad (2)$$

Outros dois indicadores comumente utilizados são a sensibilidade e a especificidade. A sensibilidade, equação 3, indica a proporção de resultados rotulados como positivos entre todos que realmente são positivos. Já a especificidade, equação 4, mede a proporção de resultados corretamente classificados como negativos pelo classificador (James; Witten; Hastie; Tibshirani, 2013; Izbicki; Santos, 2020).

$$Sensibilidade = \frac{VP}{VP + FN} \quad (3)$$

$$Especifidade = \frac{VN}{VN + FP} \quad (4)$$

Os valores preditivos positivos e negativos apresentados nas equações 5 e 6 também podem ser utilizados para avaliar um classificador. O valor preditivo positivo (VPP) define a probabilidade de o evento de interesse em uma amostra ser classificado como positivo, enquanto o valor preditivo negativo (VPN) define a

probabilidade de o não evento de interesse em uma amostra ser classificado como negativo.

$$VPP = \frac{VP}{VP + FP} \quad (5)$$

$$VPN = \frac{VN}{VN + FN} \quad (6)$$

Modelos de classificação

Os modelos de classificação têm como objetivo descobrir o relacionamento existente entre atributos de entrada e atributos de saída. O relacionamento encontrado é representado em uma estrutura que geralmente descreve e explica fenômenos ocultos no conjunto de dados, podendo ser utilizado para prever valores futuros com base em alguns valores observados. Os algoritmos mais utilizados deste tipo de classificação são os de regressão logística, *naive* Bayes, máquina de vetores de suporte, redes neurais, árvores de decisão e florestas aleatórias (Alpaydin, 2020; Izbicki; Santos, 2020; James; Witten; Hastie; Tibshirani, 2013; Rokach; Maimon; Shmueli, 2023).

Regressão Logística

A regressão logística pode ser utilizada no contexto de classificação, quando a decisão a respeito da classe a que uma nova observação pertence é binária. Assim, a variável resposta $\boldsymbol{Y}$ assumirá o valor 1 caso a observação pertença a determinada classe e receberá 0 caso não pertença a essa classe. Portanto, o objetivo é obter a probabilidade $\boldsymbol{P(Y = 1|X)}$ (James; Witten; Hastie; Tibshirani, 2013; Mccullagh; Nelder, 2019).

Considere o vetor $\boldsymbol{X} = (X_1, X_2, \dots, X_n)$ que representa um conjunto de k atributos de entrada e e $\pi(x)$ como a proporção de ocorrência do evento de interesse, ou seja, a probabilidade de pertencer à classe 1. A relação entre essa probabilidade e o conjunto de atributos, definidos pelo vetor $\boldsymbol{X}$, pode ser definida pela função logit, dada pela expressão:

$$log\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_x \quad (7)$$

em que $\pi(x)$ é dado por:

$$P(Y = 1|X) = \pi(x) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_x}}{1 + e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_x}} \quad (8)$$

Os parâmetros $\beta_0, \beta_1, \beta_2, \ldots, \beta_n$ são geralmente estimados pelo método da máxima verossimilhança ou método de mínimos quadrados ponderado (Mccullagh; Nelder, 2019).

Árvore de decisão

As árvores de decisão são modelos hierárquicos compostos por particionamentos recursivos das características de entrada X_i. As árvores são compostas por nós decisórios internos e nós terminais. Cada nó interno contém um teste sobre uma variável X_i e os resultados formam os ramos da árvore. Os nós folhas, nas extremidades da árvore, representam os valores de classificação para a variável dependente Y (Breiman; Friedman; Olshen; Stone, 2017; Izbicki; Santos, 2020; James; Witten; Hastie; Tibshirani, 2013).

Existem várias formas de indução de uma árvore de decisão, sendo um dos mais utilizados o método CART. Este algoritmo tem como objetivo criar suas decisões através de uma divisão estritamente binária, focada em seus atributos principais e seus percursores. Ou seja, ao usar o primeiro nó (raiz da árvore), este só poderá estar ligado a dois nós, assim sucessivamente ao crescimento da árvore, até que encontre uma folha (Breiman; Friedman; Olshen; Stone, 2017; Izbicki; Santos, 2020; James; Witten; Hastie; Tibshirani, 2013).

No algoritmo CART, para obtenção de cada partição, existem várias possibilidades. Assim, cada partição é escolhida utilizando o índice de Gini. Este critério avalia todas as possibilidades de divisões e escolhe a de melhor desempenho, repetindo esse processo recursivamente até que a coleção ideal seja encontrada (Breiman; Friedman; Olshen; Stone, 2017).

Redes Neurais

Uma rede neural Multicamadas Perceptrons (MLP) é uma das arquiteturas de redes neurais artificiais mais utilizadas e conhecidas. Neste tipo de rede, as entradas X_i são apresentadas na primeira camada, chamada camada de entrada. Essa camada distribui as informações para as camadas escondidas da rede. A última camada é a camada de saída, onde a classificação é obtida. A camada de

entrada e a camada de saída podem ser separadas por uma ou mais camadas intermediárias. Além disso, os neurônios de uma camada estão conectados apenas aos neurônios da camada imediatamente posterior, não havendo realimentação nem conexões entre neurônios da mesma camada (Alpaydin, 2020; Sabry, 2023).

Nas redes MLP, as camadas intermediárias efetuam o processamento. Essas conexões guardam os pesos que serão multiplicados pelas entradas, garantindo o conhecimento da rede e gerando a classificação. Assim, o treinamento de uma rede MLP consiste em um problema de obtenção dos pesos sinápticos das camadas intermediárias, de forma que produzam boas saídas (Alpaydin, 2020; Sabry, 2023).

Uma outra componente importante da rede MLP é a função de ativação, que introduz um componente não linear nas redes neurais e define o intervalo de variação da saída ou variável resposta $\boldsymbol{Y}$. Existem diversas funções de ativação, dentre as quais se destacam: a função limiar ou em degrau, a função linear por partes, a função sigmoide logística e a função tangente sigmoide (Alpaydin, 2020; Sabry, 2023).

O algoritmo de treinamento mais utilizado em modelos MLP é o *backpropagation*, que consiste no ajuste dos pesos da rede baseado na minimização do erro médio quadrático pelo algoritmo do gradiente. Nesse algoritmo, para cada valor da saída $\boldsymbol{Y}$ gerado pela rede, é calculado um erro, que representa a diferença entre o valor previsto pela rede e o valor real desejado. Esses valores de erro são então retropropagados através da rede, do final (camada de saída) para o início (camada de entrada). Durante essa retropropagação, o erro é distribuído e os pesos das conexões entre os neurônios são ajustados de acordo com a magnitude do erro e a taxa de aprendizado. Este processo de ajuste é repetido múltiplas vezes, passando por todo o conjunto de treinamento, até que o erro total seja minimizado e a rede produza a resposta desejada com a precisão necessária (Alpaydin, 2020; Sabry, 2023).

Materiais e métodos

Os dados são provenientes do Sistema de Informações de Gestão Acadêmica (SIGA), e a extração dos mesmos foi realizada pela Gerência de Informações Institucionais da Pró-Reitoria de Planejamento (GEII/PROPLAN), órgão da Universidade Federal de Mato Grosso (UFMT). Embora os dados sejam de natureza

pública, não são de livre acesso, sendo necessário obter permissão para acessá-los e garantir sua confidencialidade.

O presente estudo se refere às matrículas de estudantes que iniciaram a graduação em cursos presenciais nos anos de 2016, 2017 e 2018 na UFMT. As informações contidas nos dados referem-se aos ingressantes dos 106 cursos de graduação presenciais dos cinco *campi*: Araguaia, Cuiabá, Rondonópolis, Sinop e Várzea Grande.

Os dados considerados constam de 14.303 cadastros de estudantes. A variável de interesse Y é a situação do estudante no final do segundo semestre letivo de 2018, sendo consideradas duas situações:

- 1 – se o estudante está cursando a graduação;
- 0 – se o estudante foi desvinculado da instituição ou foi transferido de curso.

Foram consideradas 14 variáveis de entrada X, conforme apresentado no quadro 1. Estas foram escolhidas por serem variáveis completas no banco de dados da instituição. Além disso, procurou-se captar três dimensões: acadêmicas, demográficas e sociais, de forma a identificar padrões e fatores críticos sobre a variável de interesse Y.

Os modelos de classificação para a situação dos estudantes de graduação da UFMT foram obtidos utilizando três diferentes técnicas: regressão logística, redes neurais e árvores de decisão. A escolha dessas técnicas se baseou na sua ampla utilização e eficácia demonstrada em estudos semelhantes sobre permanência estudantil.

A base de dados foi dividida em conjuntos de treinamento e teste, seguindo as etapas do processo KDD (*Knowledge Discovery in Databases*) (Alpaydin, 2020), que inclui a preparação e particionamento dos dados.

Quadro 1 – Lista e descrição das variáveis de entrada e seus respectivos níveis de mensuração

Variável	Níveis	Comentário
Campus	Araguaia, Cuiabá, Rondonópolis, Sinop e Várzea Grande	Em 2018, o *campus* de Rondonópolis tornou-se a Universidade Federal de Rondonópolis (UFR).

Continua

Conclusão

Variável	Níveis	Comentário
Área	Ciências Agrárias, Ciências Biológicas, Ciências da Saúde, Ciências Exatas e da Terra, Ciências Humanas, Ciências Sociais Aplicadas, Engenharias, Linguística, Letras e Artes, Outros - Ciências Sociais, Outros - Biomedicina	Área de classificação dos cursos (CNPq).
Tipo	1- Bacharelado; 0 - Licenciatura	Tipo de curso.
Integralização	1 - Crédito; 0 - Outros	Regime para integralização das componentes curriculares do curso. Outros compostos por seriado anual e seriado semestral.
Tempo de integralização	8, 9, 10, 12	Número mínimo de semestres para integralização do curso.
Periodização	1 - Anual; 0 - Semestral	Forma de periodização do curso.
Turno	1 - Único; 0 - Integral ou dois turnos	Turno de oferta do curso.
Tipo de Vaga	1 - Ampla concorrência; 0 - Cotas	Tipo de vaga concorrida no ingresso no curso.
Escola	1 - Privada; 0 - Pública	Rede de ensino da escola que o estudante cursou o ensino médio.
Sexo	1 - Masculino; 0 - Feminino	
Nascimento	1 - Mato Grosso; 0 - Outros	Estado de nascimento do estudante.
Estado civil	1 - Solteiro; 0 - Outros	Estado civil do estudante ao ingressar na universidade. Outros composto por casados, divorciados e viúvos.
Raça	1 - Branca; 0 - Outros	Raça declarada pelo estudante. Outros composto por Amarela, Indígena, Negra, Pardo, Mulato. Categorias utilizadas no SIGA.
Idade	Idade em anos completos	

Fonte: Universidade Federal de Mato Grosso (2024).

O conjunto de treinamento continha 75% dos dados e o conjunto de teste 25%, com a amostragem realizada por meio de amostragem aleatória simples. Desta forma, o conjunto de treinamento continha 10.727 observações e o conjunto de teste 3.576. Para lidar com o desbalanceamento dos dados, foi utilizado o método ROSE (*Random Over-Sampling Examples*) (Lunardon; Menardi; Torelli, 2014), que produz uma amostra de dados sintéticos balanceados a partir de uma abordagem

bootstrap suavizada. A seleção das variáveis foi realizada por meio da validação cruzada *k-fold* com 50 divisões, considerando-se as variáveis que melhor contribuíram para uma boa classificação dos estudantes.

Para mensurar a qualidade da classificação das três técnicas, foram utilizados os escores de permanência e as métricas da matriz de confusão: acurácia, sensibilidade, especificidade, valor preditivo positivo e valor preditivo negativo, tanto na base de treinamento quanto na de teste. A escolha dessas métricas se justifica pela necessidade de obter medidas não agregadas do desempenho do modelo, permitindo uma avaliação detalhada e específica de diferentes aspectos da performance. Essas métricas são essenciais para identificar não apenas a eficácia geral, mas também a capacidade de cada modelo de lidar com diferentes tipos de erros (falsos positivos e falsos negativos).

O ajuste dos modelos foi realizado utilizando o *software* R (R Core [...], 2020) por meio do pacote caret (Kuhn, 2020). Para o ajuste da regressão logística, foi utilizado o método de mínimos quadrados iterativamente ponderados (McCullagh; Nelder, 2019). A árvore de decisão foi construída utilizando o algoritmo CART, que aplica o índice Gini para escolher as divisões (Breiman; Friedman; Olshen; Stone, 2017). Já a rede MLP foi treinada utilizando o algoritmo *backpropagation* com cinco camadas ocultas, função de ativação sigmoide e 5000 épocas (Haykin, 1999).

Resultados e discussão

Perfil dos estudantes

Na Tabela 2 é apresentada a distribuição de frequências em relação às variáveis explicativas. Essas frequências permitem uma visão geral de alguns aspectos da oferta de cursos, do perfil dos estudantes que ingressaram na UFMT e de suas características demográficas no período de 2016 a 2018. Por isso, são descritas abaixo.

Analisando a Tabela 2, é possível dimensionar alguns aspectos da oferta de cursos na UFMT, como a distribuição entre os *campi*, as áreas e o perfil dos cursos. O *campus* Cuiabá apresentou o maior percentual de estudantes, com 51,90%, seguido de Rondonópolis, com 19,39%. O *campus* com menor percentual foi o de Várzea Grande, ocupando 5,17% dos dados, seguido por Sinop com 11,43% e Araguaia, com 12,12%. O *campus* Várzea Grande foi o último a ser implantado na UFMT; somente em 2014 teve os primeiros ingressantes nos cursos.

Foi observado que mais da metade das matrículas de estudantes estão concentradas em três áreas de cursos: Ciências Sociais Aplicadas, com 20,37%, Ciências Agrárias, com 18,45%, e Ciências Exatas e da Terra, com 14,10%. Um segundo grupo de áreas possui percentual superior a 10% das matrículas, sendo composto por Ciências da Saúde, com 12,99%, Engenharias, com 12,65%, e Ciências Humanas, com 11,49%. Um terceiro grupo de áreas soma conjuntamente um percentual de 9,94%, sendo composto por Linguística, Letras e Artes, Ciências Biológicas, Outros – Ciências Sociais e Outros – Biomedicina.

Quanto ao perfil dos cursos, observa-se que a maioria das matrículas está na modalidade de bacharelado (76,60%), com regime de integralização de crédito (72,33%), periodização semestral (86,67%) e com oferta integral ou em mais de dois turnos (55,88%).

Tabela 2 - Distribuição de frequência dos estudantes

Variável	Nível	Frequência	Percentual
Total		14303	100,00%
Campus			
	Araguaia	1733	12,12%
	Cuiabá	7423	51,90%
	Rondonópolis	2773	19,39%
	Sinop	1635	11,43%
	Várzea Grande	739	5,17%
Área			
	Ciências Agrárias	2639	18,45%
	Ciências Biológicas	474	3,31%
	Ciências da Saúde	1858	12,99%
	Ciências Exatas e da Terra	2017	14,10%
	Ciências Humanas	1644	11,49%
	Ciências Sociais Aplicadas	2913	20,37%
	Engenharias	1810	12,65%
	Lingüística, Letras e Artes	704	4,92%
	Outros – Biomedicina	111	0,78%

Continua

Variável	Nível	Frequência	Percentual
	Outros - Ciências Sociais	133	0,93%
Tipo			
	Bacharelado	10956	76,60%
	Licenciatura	3347	23,40%
Integralização			
	Crédito	10346	72,33%
	Outros	3957	27,67%
Periodização			
	Anual	12396	86,67%
	Semestral	1907	13,33%
Variável	Nível	Frequência	Percentual
Turno			
	Integral ou dois turnos	7992	55,88%
	Único	6311	44,12%
Tipo de Vaga			
	Ampla concorrência	7359	51,45%
	Cotas	6944	48,55%
Escola			
	Outros	10183	71,19%
	Privada	4120	28,81%
Sexo			
	Feminino	7377	51,58%
	Masculino	6926	48,42%
Nascimento			
	MT	9791	68,45%
	Outros	4512	31,55%
Estado civil			
	Outros	770	5,38%
	Solteiro	4512	94,62%
Raça			

Continua

Conclusão

Variável	Nível	Frequência	Percentual
	Branca	8921	62,37%
	Outros	5382	37,63%

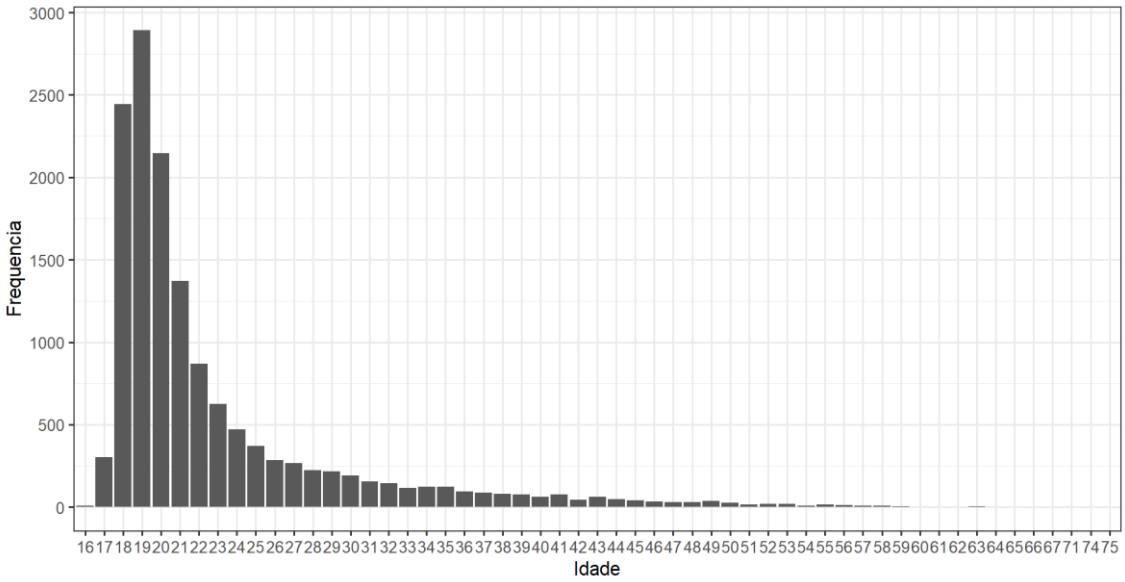
Fonte: Os autores (2022).

Observando o perfil de ingresso dos estudantes, é possível constatar que a maioria das matrículas são de vagas de ampla concorrência (51,45%), ou seja, de pessoas que ingressaram fora do sistema de cotas. Apesar de possuir um menor percentual, as matrículas dos estudantes que entraram pelo sistema de cotas representam uma parte significativa, com 48,55% das matrículas. A maioria dos estudantes veio da rede pública, com 71,19%, enquanto 28,81% vieram de ensino particular.

Apesar do sistema de cotas e da predominância da escola pública, quanto à questão étnico-racial, 62,37% dos estudantes são considerados brancos, o que destoa da população em geral de Mato Grosso, que possuía 30,8% das pessoas autodeclaradas brancas em 2018 (IBGE, 2018).

Nas características demográficas, observa-se que as mulheres representam um percentual levemente superior ao de homens, com 51,58% sendo do sexo feminino e 48,42% do masculino. A maioria dos estudantes provém de Mato Grosso, com 68,45%, e são solteiros (94,62%). Na Figura 1 é apresentada a distribuição etária dos estudantes, na qual se percebe uma grande concentração de estudantes na faixa etária de 18 a 21 anos. Estudantes de 18 anos representam 17,08% do conjunto de dados, enquanto os de 19 anos correspondem a 20,21%, de 20 anos contribuem com 14,99%, e estudantes de 21 anos compõem 9,57%. Também se observa que na faixa etária de 22 a 30 anos há 24,57% dos dados, enquanto na de 31 a 40 anos há 7,43% e, acima de 41 anos, 3,94%.

Figura 1 - Distribuição da faixa etária dos estudantes



Fonte: Os autores (2022).

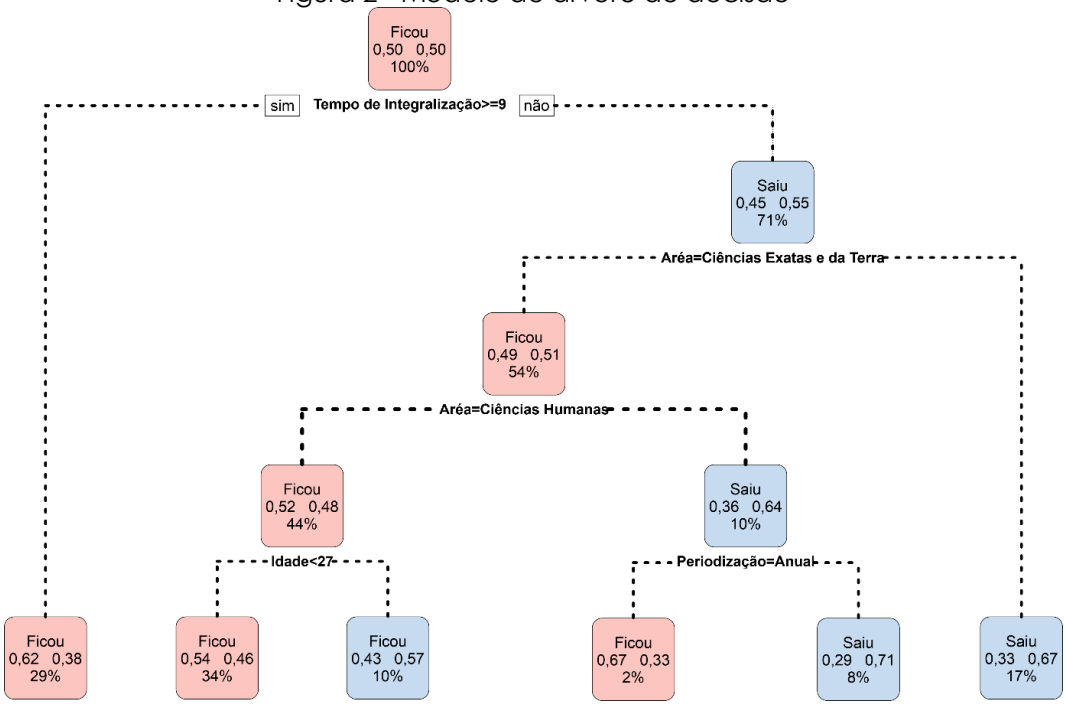
Modelos de classificação

Árvore de Decisão

Na Figura 2, apresenta-se a árvore de decisão gerada pelo modelo, que revelou duas principais divisões entre os estudantes com base na variável "tempo de integralização". O primeiro grupo, representando 29% dos dados, é composto por estudantes matriculados em cursos com duração superior a nove semestres, sendo classificados pelo modelo como "cursando" com uma precisão de 62%. Por outro lado, o segundo grupo, que compreende 71% dos dados, engloba estudantes matriculados em cursos com duração inferior a nove semestres, os quais foram classificados como "desistentes" ou "transferidos" com uma precisão de 55%.

No segundo nível de análise, a árvore de decisão utiliza a variável "área" para refinar a classificação dos estudantes pertencentes ao grupo 2. Especificamente, quando o curso pertence à área de Ciências Exatas e da Terra e tem uma duração média inferior a 4,8 anos, o modelo classificou esses estudantes como "desistentes" ou "transferidos", representando 17% do conjunto de dados, com uma precisão de 67%. Esses resultados destacam a relevância tanto do tempo de integralização quanto da área do curso na predição das trajetórias acadêmicas dos estudantes.

Figura 2 - Modelo de árvore de decisão



Fonte: Os autores (2022).

Nos últimos nós da árvore, são utilizadas as variáveis "periodização do curso" e "idade do aluno", condicionadas pelas variáveis já mencionadas e pela variável "área" (CNPQ), para cursos de Ciências Humanas. Se o estudante não for de um curso de Ciências Humanas e tiver idade menor que 27 anos, o modelo o classifica como "cursando", com uma precisão de 54%. Caso contrário, o modelo o classifica como "desistente" ou "transferido", com uma precisão de 57%.

Na Tabela 3, observa-se que a acurácia do modelo foi a mesma tanto no conjunto de treino quanto no de teste, com 60%. A métrica de sensibilidade diminuiu 1% no conjunto de teste, apresentando 59% no conjunto de treino e 60% no conjunto de teste. Houve um aumento de 2% na medida de especificidade, apresentando 65% nos dados de treino e 67% nos dados de teste.

Tabela 3 - Tabela de métricas do modelo de árvore de decisão

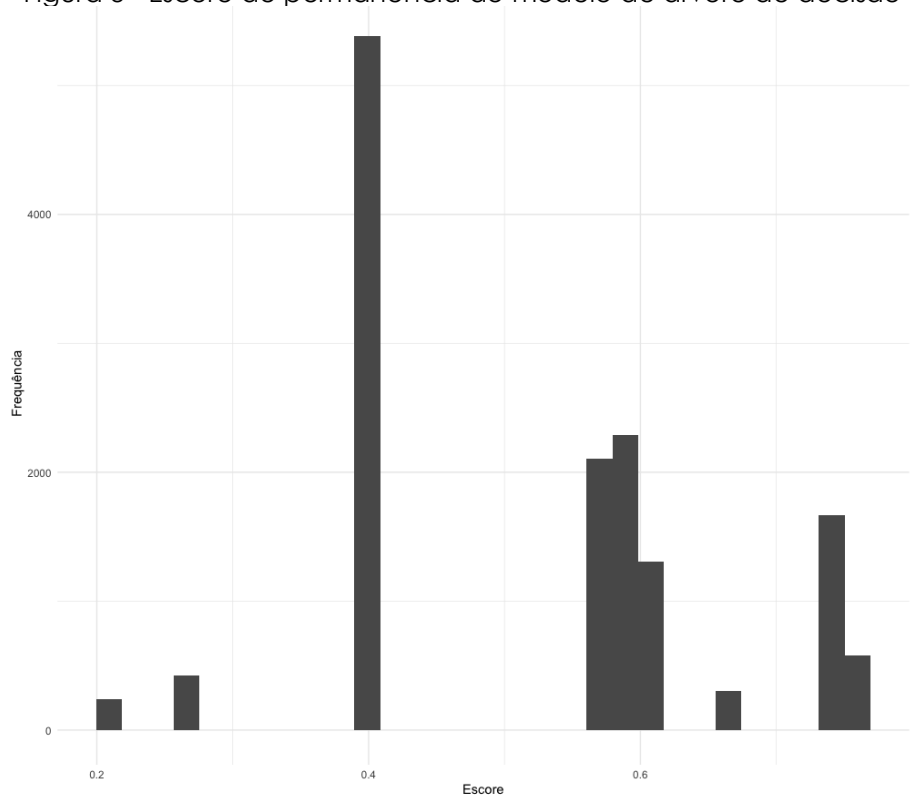
Métrica	Treino	Teste
Acurácia	0,60	0,60
Sensibilidade	0,60	0,59
Especificidade	0,65	0,67
VPP	0,95	0,96
VPN	0,11	0,11

Fonte: Os autores (2022).

Devido ao alto nível de desbalanceamento na variável resposta, o valor preditivo positivo (VPP) é alto, com 95% nos dados de treino e 96% nos dados de teste, enquanto o valor preditivo negativo (VPN) é baixo, com 11% em ambos os conjuntos.

Na figura 3, é apresentado o escore de permanência obtido pelo modelo, mostrando um claro espaçamento entre as probabilidades, com concentração de frequência em determinadas regiões. Há uma alta concentração na probabilidade de aproximadamente 40%, região que classifica o estudante como "desistente" ou "transferido". A segunda maior frequência ocorre na região de probabilidades de 56% a 61%. Por fim, observa-se algumas ponderações com alta probabilidade de pertencer à classe de estudantes.

Figura 3 - Escore de permanência do modelo de árvore de decisão



Fonte: Os autores (2022).

Regressão Logística

Na Tabela 4 é apresentada a estimativa dos parâmetros e a razão de chance para a regressão logística, observando-se que todas as variáveis selecionadas pelo modelo apresentam estimativa positiva, indicando que elas contribuem para a permanência dos alunos. Observando a razão de chances em relação à permanência no curso, ser estudante de um curso de turno único aumenta a chance

de permanência em 20%. Em relação ao perfil de ingresso, aqueles oriundos de escola privada têm 22% mais chance de permanecer, e os que entraram por cota têm 6% mais chance. Nas características demográficas, ser solteiro aumenta a chance em 57%, ser de Mato Grosso em 10%, do sexo masculino em 8%, e para cada aumento na idade há um acréscimo de chance de 2% que estão "cursando".

Tabela 4 - Estimativa dos parâmetros e razão de chance para regressão logística

Variável	Estimativa	Razão de chances
Intercepto	1,34	3,81
Turno (Único)	0,18	1,20
Escola (Privada)	0,20	1,22
Tipo de Vaga (Cotas)	0,06	1,06
Estado civil (Solteiro)	0,45	1,57
Estado de Nascimento (Mato Grosso)	0,10	1,10
Sexo (Masculino)	0,07	1,08
Idade	0,02	1,02

Fonte: Os autores (2022).

É apresentado na Tabela 5 o conjunto de métricas do modelo de regressão logística. A acurácia foi a mesma, tanto no conjunto de treino quanto no de teste, com 61%. A métrica de sensibilidade teve um decréscimo de desempenho para 60% nos dados de teste, em comparação com 61% nos dados de treino. Houve um aumento de 5% na medida de especificidade, com 60% no conjunto de treino e 65% no conjunto de teste. Os valores de VPP e VPN foram os mesmos do modelo anterior, com o VPP alcançando um valor muito alto e o VPN um valor muito baixo.

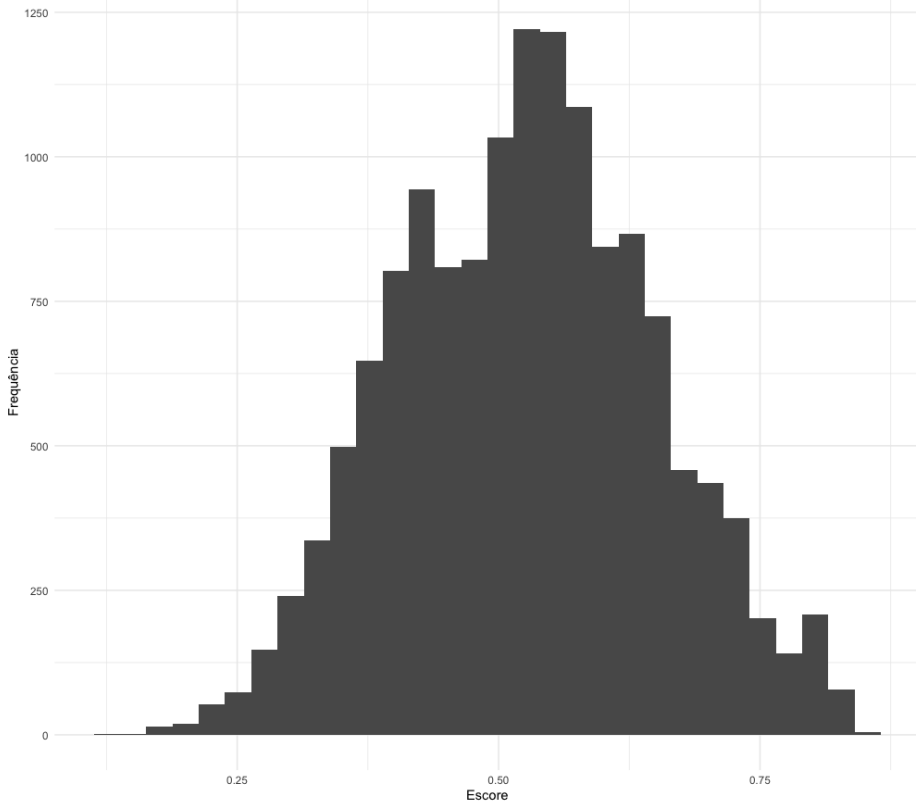
Tabela 5 - Métricas de performance do modelo de regressão logística

Métrica	Treino	Teste
Acurácia	0,61	0,61
Sensibilidade	0,61	0,60
Especificidade	0,60	0,65
VPP	0,95	0,96
VPN	0,11	0,11

Fonte: Os autores (2022).

O escore de permanência, apresentado na Figura 4 do modelo de regressão logística, aproxima-se de uma distribuição normal. O pico de frequência está na região de probabilidades entre 50% e 58%, com a cauda esquerda apresentando menor frequência que a cauda direita, esta última demonstrando um número considerável de observações com aproximadamente 80%.

Figura 4 - Escore de permanência do modelo de regressão logística

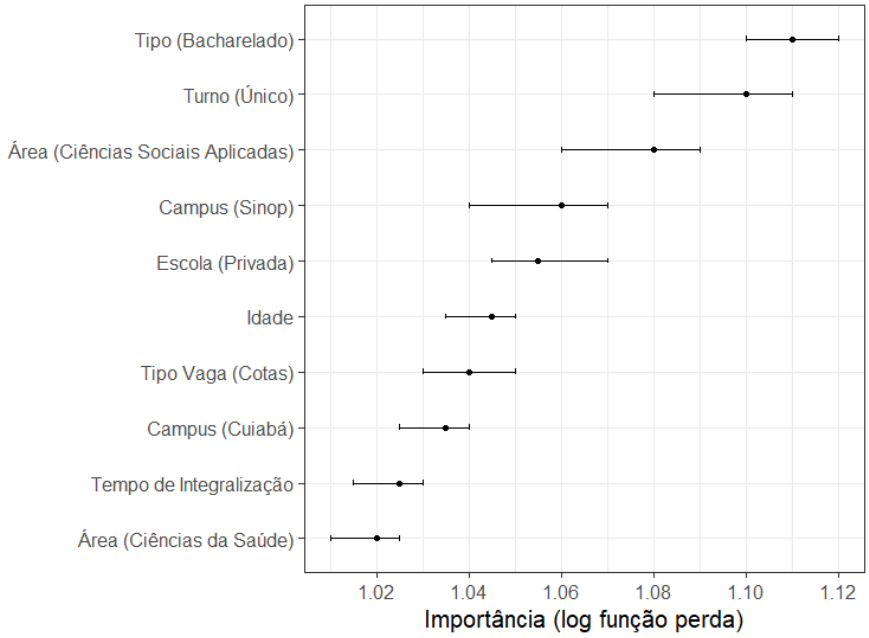


Fonte: Os autores (2022).

Redes Neurais

Na Figura 5 são apresentadas as dez variáveis mais importantes segundo o modelo de redes neurais MLP. As quatro variáveis mais importantes são referentes às características do curso: ser bacharelado, turno único, área de Ciências Sociais Aplicadas e ser do *campus* Sinop. Em seguida, estão as características do aluno: ser de escola privada, idade e tipo de vaga por cotas. As variáveis consideradas menos importantes foram as características do curso: *campus* Cuiabá, tempo de integralização do curso e área de Ciências da Saúde.

Figura 5 - Variáveis mais importantes do modelo de redes neurais



Fonte: Os autores (2022).

Na Tabela 6, nota-se que a acurácia do modelo de redes neurais foi inferior em relação aos dois modelos anteriores, com 58% nos dados de treino e 57% nos dados de teste. Na medida de sensibilidade, houve decréscimo de performance, com 57% nos dados de treino e 56% nos dados de teste. Embora a métrica de especificidade tenha performado nos dados de treino consideravelmente melhor que nos modelos anteriores, com 70%, nos dados de teste a performance foi de 67%. O VPP foi o mesmo em ambos os conjuntos, alcançando 96% de performance. O VPN performou 11% no conjunto de treino e 10% no conjunto de teste.

Tabela 6 - Tabela de métricas do modelo redes neurais

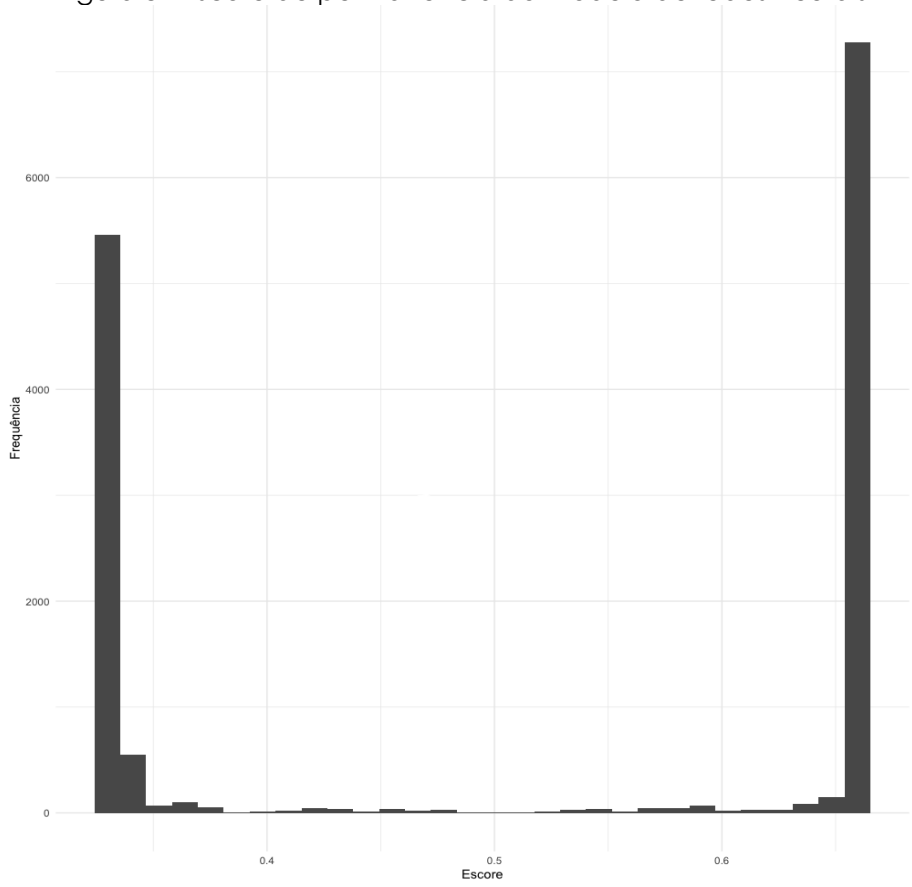
Métrica	Treino	Teste
Acurácia	0,58	0,57
Sensibilidade	0,57	0,56
Especificidade	0,70	0,67
VPP	0,96	0,96
VPN	0,11	0,10

Fonte: Os autores (2022).

Conforme a Figura 8, percebe-se uma forte divisão feita pelo modelo de redes neurais no que se refere ao escore de permanência dos estudantes. Há uma forte

concentração na região de probabilidade de aproximadamente 67% e uma alta concentração também na região de probabilidade de aproximadamente 25%.

Figura 8 - Escore de permanência do modelo de redes neurais



Fonte: Os autores (2022).

Comparação dos modelos

Em termos de acurácia, o modelo de regressão logística teve uma leve vantagem com 61% tanto nos dados de treino quanto nos de teste, enquanto o modelo de árvore de decisão apresentou 60% e o modelo de redes neurais teve 58% no treino e 57% no teste.

Na sensibilidade, que mede a capacidade do modelo de identificar corretamente os casos positivos, o modelo de regressão logística novamente teve um desempenho ligeiramente melhor, seguido pelo modelo de árvore de decisão e, por último, pelo modelo de redes neurais.

A especificidade, que indica a capacidade do modelo de identificar corretamente os casos negativos, foi onde o modelo de redes neurais se destacou, alcançando 70% nos dados de treino e 67% nos dados de teste. O modelo de árvore

de decisão apresentou 65% no treino e 67% no teste, enquanto o modelo de regressão logística teve 60% no treino e 65% no teste.

O VPP (Valor Preditivo Positivo) foi consistentemente alto em todos os modelos, com 95% para os modelos de árvore de decisão e regressão logística, e 96% para o modelo de redes neurais. Isso sugere que, quando um modelo prevê que um estudante permanecerá, essa previsão é muito confiável. Um alto VPP é particularmente útil em contextos onde é crucial minimizar os falsos positivos, garantindo que os recursos e esforços sejam direcionados a estudantes que realmente têm maior probabilidade de permanecer.

Por outro lado, o VPN (Valor Preditivo Negativo) foi muito baixo em todos os modelos, indicando que a capacidade dos modelos de identificar corretamente os casos negativos (estudantes que saem) é limitada.

Os escores de permanência também variaram entre os modelos. O modelo de regressão logística apresentou uma curva próxima de uma distribuição normal. O modelo de redes neurais dividiu as probabilidades em duas grandes regiões, enquanto o modelo de árvore de decisão concentrou as probabilidades em seis regiões, três delas com grande frequência.

Desta forma cada modelo tem suas vantagens e desvantagens. O modelo de regressão logística teve uma leve vantagem em termos de acurácia e sensibilidade, enquanto o modelo de redes neurais se destacou na especificidade. Todos os modelos apresentaram um alto VPP, indicando que as previsões de permanência são confiáveis. No entanto, o VPN baixo é uma limitação comum a todos os modelos, sugerindo dificuldades na identificação precisa dos estudantes que não permanecem. Essas observações sugerem que, embora nenhum modelo seja perfeito, cada um pode oferecer resultados valiosos dependendo do contexto e da prioridade das métricas de desempenho.

Comparação com outros estudos

Neste estudo, foram aplicados modelos de regressão logística, árvores de decisão e redes neurais para analisar a permanência de estudantes na UFMT, uma abordagem que se alinha com as práticas comuns na mineração de dados educacionais, mas com uma aplicação focada nas características e desafios específicos da instituição. Autores como Lanes e Alcântara (2018) e Carrano, Albergaria, Infante e Rocha (2019) mostraram o potencial de técnicas similares para

identificar riscos de evasão, combinando várias técnicas para uma análise robusta de indicadores de evasão.

Em estudos como o de Jesus, Rodrigues e Costa Junior (2021) e Dutra, Souza e Fernandes (2022), observam-se aplicações de modelos preditivos em contextos variados, onde estratégias específicas foram utilizadas para entender melhor as causas da evasão e retenção estudantil. Esses estudos, embora centrados em outras instituições e utilizando outras bases de dados, fornecem perspectivas valiosas sobre como diferentes variáveis e modelos podem ser eficazes em contextos educacionais diversos. Neste estudo, a regressão logística apresentou uma leve vantagem em termos de acurácia e sensibilidade, o que está em consonância com os resultados de Jesus, Rodrigues e Costa Junior (2021), que também encontraram que variáveis relacionadas ao desempenho acadêmico tinham maior peso na predição de permanência.

O trabalho de Teodoro e Kappel (2020) e Lanes e Alcântara (2018) explora o uso de técnicas avançadas de classificação automática para prever o desempenho dos alunos e potenciais evasões, oferecendo um paralelo interessante a esta pesquisa. Em particular, Teodoro e Kappel (2020) destacam a alta especificidade das redes neurais, corroborando os achados deste estudo de que as redes neurais se destacaram nessa métrica. No entanto, enquanto Lanes e Alcântara (2018) utilizaram validação cruzada *k-fold* com menos divisões, neste estudo optou-se por uma abordagem com 50 divisões, o que pode ter contribuído para a diferença nos resultados observados.

Dutra, Souza e Fernandes (2022) também utilizaram técnicas de aprendizado de máquina semelhantes às deste estudo. A acurácia dos modelos neste estudo foi semelhante, com a regressão logística apresentando uma acurácia ligeiramente superior. No entanto, eles destacaram a importância da variável socioeconômica, que não foi um foco principal neste estudo.

Este estudo se distingue pelo foco na comparação direta de diferentes modelos preditivos para identificar o mais adequado para os dados específicos da UFMT. Essa abordagem, embora inspirada nos métodos usados nos estudos mencionados, se dedica a integrar essas técnicas em uma estrutura diretamente aplicável para a melhoria das políticas de permanência na UFMT. A análise detalhada das métricas de performance, como acurácia, sensibilidade, especificidade, VPP e VPN, mostrou que, apesar de todos os modelos apresentarem

uma acurácia relativamente baixa, o alto VPP indica que as previsões de permanência são muito confiáveis, o que é crucial para a alocação eficiente de recursos na instituição.

Conclusão

A partir da comparação das três abordagens, quanto aos escores de permanência e as métricas da matriz de confusão, foi possível observar que o modelo de regressão logística é o mais capaz de prever a permanência dos estudantes na UFMT.

A vantagem do modelo de regressão logística não ocorre de modo linear em todas as medidas avaliadas. Os resultados das três abordagens são semelhantes nas medidas de performance, apesar de que cada modelo utilizado seleciona um grupo de variáveis diferentes como prioritárias. Os modelos de árvore de decisão e redes neurais performaram ligeiramente melhor na medida de especificidade.

O modelo de regressão logística chegou a melhores resultados nas medidas de acurácia e sensibilidade, bem como apresenta um comportamento mais distribuído dos escores de permanência, o que permite maior flexibilidade na definição de um novo limiar (*threshold*) para a classificação dos estudantes. Os modelos de redes neurais e de árvore de decisão dividiram as probabilidades em um pequeno número de regiões, dificultando o estabelecimento de um limiar mais adequado para prever as classes.

A partir desses resultados, pode-se afirmar que o modelo de regressão logística pode facilitar a identificação das variáveis que mais influenciam a permanência dos estudantes da UFMT. A compreensão dessas variáveis é central no desenvolvimento das políticas institucionais da Universidade, tendo o potencial de contribuir para identificar grupos de estudantes prioritários, bem como ser o insumo de ações e programas que objetivem a aumentar a permanência dos estudantes nos cursos, podendo fomentar as discussões sobre os programas de suporte acadêmico e assistência estudantil.

Considerando a centralidade dessas políticas, a aplicação de rotinas, com vista a obter estatísticas sobre a permanência, pode ser implementado como suporte a tomada de decisão dos conselhos universitários e na gestão administrativa. Neste aspecto, o modelo de regressão possui uma vantagem adicional, por ser menos exaustivo computacionalmente, o que facilita sua implementação na rotina

administrativa, e possuir *outputs* com maior facilidade de compreensão pela comunidade universitária, pois o debate sobre a permanência estudantil envolve os diversos setores da Universidade, como estudantes, professores e Comunidade Externa, devendo as informações serem comunicadas para todos.

Em resumo, apesar dos modelos de árvore de decisão e redes neurais terem mostrado alguns pontos fortes, a regressão logística provou ser a abordagem mais robusta, tendo o potencial de ser melhor interpretável pela comunidade acadêmica na análise de permanência dos estudantes.

A temática da permanência do estudante desafia as políticas públicas, sendo fundamental a continuidade das investigações no âmbito da UFMT, bem como o diálogo com investigação efetuadas em outras Universidades. Em futuras pesquisas, a depender da disponibilidade dos dados, poderão ser incluídas outras variáveis de interesse social, como aspectos socioeconômicos e psicossociais, que não foram foco principal deste estudo, mas que se mostraram relevantes em outros estudos comparáveis. A inclusão dessas variáveis pode oferecer uma visão ainda mais abrangente dos fatores que influenciam a permanência dos estudantes.

Além de ampliar o escopo das variáveis, poderá haver a aplicação de outras abordagens metodológicas, buscando investigar outras técnicas de aprendizado de máquina e métricas de performance, para avaliar a eficácia dos modelos de forma mais abrangente.

Direcionamento para trabalhos futuros

Os resultados deste estudo podem ser utilizados para direcionar diversos trabalhos futuros. Primeiramente, as conclusões obtidas sobre as variáveis que mais influenciam a permanência dos estudantes podem ajudar a universidade a desenvolver políticas mais direcionadas e eficazes para aumentar a retenção estudantil. Por exemplo, programas de suporte acadêmico e financeiro podem ser intensificados para grupos de estudantes identificados como de maior risco de evasão.

Além disso, futuras pesquisas podem explorar a inclusão de variáveis adicionais, como aspectos socioeconômicos e psicossociais, que não foram foco principal deste estudo, mas que se mostraram relevantes em outros estudos comparáveis. A inclusão dessas variáveis pode oferecer uma visão ainda mais abrangente dos fatores que influenciam a permanência dos estudantes.

Outro aspecto importante para trabalhos futuros é a aplicação de técnicas de aprendizado de máquina mais avançadas e a comparação com outras métricas de performance, para avaliar a eficácia dos modelos de forma mais abrangente. Essas técnicas podem incluir, por exemplo, Floresta Aleatória e AdaBoost, que podem oferecer melhorias na predição e na compreensão dos fatores de evasão.

Adicionalmente, a implementação prática dos modelos desenvolvidos pode ser explorada através da criação de sistemas de alerta precoce. Esses sistemas poderiam identificar, em tempo real, os estudantes com maior risco de evasão, permitindo intervenções preventivas personalizadas e mais eficazes.

Finalmente, a replicação deste estudo em outras instituições de ensino superior pode proporcionar um entendimento mais amplo e generalizável dos fatores de permanência estudantil, contribuindo para a melhoria das políticas educacionais em um nível mais amplo.

Referências

ALPAYDIN, E. *Introduction to machine learning*. 4. ed. Cambridge: MIT press, 2020.

BREIMAN, L.; FRIEDMAN, J.; OLSHEN, R. A.; STONE, C. J. *Classification and regression trees*. Nova York: CRC Press, 2017.

CARRANO, D.; ALBERGARIA, E. T. de; INFANTE, C.; ROCHA, L. Combinando técnicas de mineração de dados para melhorar a detecção de indicadores de evasão universitária. *In:* SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 30., 2019, Brasília, DF. *Anais* [...]. Brasília: [s. n.], 2019.

COIMBRA, C. L.; SILVA, L. B. e; COSTA, N. C. D. A evasão na educação superior: definições e trajetórias. *Educação e Pesquisa*, São Paulo, v. 47, 2021. DOI: https://doi.org/10.1590/S1678-4634202147228764. Disponível em: https://www.scielo.br/j/ep/a/WRKk9JVNBnJJsnNyNkFfJQj/#. Acesso em: 20 ago. 2023.

DORILEO, B. P. *Ensino superior em Mato Grosso:* até a implantação da UFMT. Campinas, SP: Komedi, 2005.

DORILEO, B. P. *Universidade, o fazejamento*. Cuiabá: Ed. UFMT, 1977.

DUTRA, J. F.; SOUZA, J. P. L. de; FERNANDES, D. Y. de S. Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do IFPB. *Revista Principia* - Divulgação Científica e Tecnológica do IFPB, João Pessoa, v. 59, n. 3, p. 1009-1027, 2022. DOI: http://dx.doi.org/10.18265/1517-0306a2021id5488. Disponível em: https://periodicos.ifpb.edu.br/index.php/principia/article/view/5488/1795. Acesso em: 20 ago. 2023.

IBGE. *Pesquisa nacional por amostra de domicílios contínua anual:* tabela 6408: população residente, por sexo e cor ou raça 2018. Rio de Janeiro: IBGE, 2018. Disponível em: https://sidra.ibge.gov.br/Tabela/6408. Acesso em: 8 set. 2022.

HAYKIN, S. *Neural networks:* a comprehensive foundation. 2. ed. Upper Saddle River, NJ: Prentice Hall, 1999.

IZBICKI, R.; SANTOS, T. M. dos. *Aprendizado de máquina:* uma abordagem estatística. São Carlos, SP: Rafael Izbicki, 2020.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. *An introduction to statistical learning:* with applications in R. Nova York: Springer, 2013.

JESUS, H. O.; RODRIGUES, L. C.; COSTA JUNIOR, A. de O. Predição de evasão escolar na licenciatura em computação. *Revista Brasileira de Informática na Educação*, Florianópolis, v. 29, p. 255-272, 2021. DOI: https://doi.org/10.5753/rbie.2021.29.0.255. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/2997. Acesso em: 20 ago. 2022.

KUHN, M. *Caret:* classification and regression training: R package version 6.0-90. [*S. l.*]: [s. n.], 2020.

LANES, M.; ALCÂNTARA, C. Predição de alunos com risco de evasão: estudo de caso usando mineração de dados. *In:* SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 29., 2018, Fortaleza. *Anais* [...]. Brasília: [s. n.], 2018.

LUNARDON, N.; MENARDI, G.; TORELLI, N. Rose: a package for binary imbalanced learning. *The R Journal*, [*S. l.*], v. 6, n. 1, p. 79-89, 2014.

MALANGE, F. C. V.; NOGUEIRA, P. S.; ZARDO, L. A. R. O acesso à educação superior pública no Brasil sob ótica dos dados nacionais. *REVELLI* - Revista de Educação, Linguagem e Literatura, Anápolis, v. 13, p. 1-23, 2021. DOI: https://doi.org/10.51913/revelli.v13i0.12176. Disponível em: https://www.revista.ueg.br/index.php/revelli/article/view/12176. Acesso em: 20 ago. 2023.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. 2. ed. Nova York: Chapman and Hall, 2019.

NASCIMENTO, F. F. do; DANTAS, L. C. de O.; CASTRO, A. F. de; QUEIROZ, P. G. G. Técnicas de mineração de dados e aprendizado de máquina aplicados à evasão estudantil: um mapeamento sistemático da literatura. *Revista Brasileira de Informática na Educação*, Florianópolis, v. 32, p. 270–294, 2024. DOI: https://doi.org/10.5753/rbie.2024.3296. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/3296. Acesso em: 20 ago. 2023.

R CORE team: R: a language and environment for statistical computing. Vienna, Austria: [s. n.], 2020.

ROKACH, L.; MAIMON, O.; SHMUELI, E. (ed.). *Machine learning for data science handbook*: data mining and knowledge discovery handbook. 3. ed. Berlim: Springer Nature, 2023.

SABRY, F. *Multilayer perceptron:* fundamentals and applications for decoding neural networks. [*S. l.*]: One Billion Knowledgeable, 2023.

SANTOS JUNIOR, J. da S.; REAL, G. C. M. Fator institucional para a evasão na educação superior: análise da produção acadêmica no Brasil. *Revista Internacional de Educação Superior*, Campinas, SP, v. 6, 2019. DOI: https://doi.org/10.20396/riesup.v6i0.8656028. Disponível em: https://periodicos.sbu.unicamp.br/ojs/index.php/riesup/article/view/8656028. Acesso em: 20 ago. 2023.

TEIXEIRA, M. D. de J.; QUITO, F. de M. Taxas longitudinais de diplomação, evasão e trancamento: método para análise da trajetória discente na educação superior. *Avaliação*: Revista da Avaliação da Educação Superior, Campinas, SP, v. 26, n. 2, p. 546-567, 2021.

TEODORO, L. de A.; KAPPEL, M. A. A. Aplicação de técnicas de aprendizado de máquina para predição de risco de evasão escolar em instituições públicas de ensino superior no Brasil. *Revista Brasileira de Informática na Educação*, Porto Alegre, v. 28, 2020. DOI: https://doi.org/10.5753/rbie.2020.28.0.838. Disponível em: http://milanesa.ime.usp.br/rbie/index.php/rbie/article/view/v28p838. Acesso em: 20 ago. 2023.

UNIVERSIDADE FEDERAL DE MATO GROSSO. Pró-Reitoria de Planejamento. *Anuário estatístico 2020:* ano base 2019. Cuiabá: UFMT, 2020.

UNIVERSIDADE FEDERAL DE MATO GROSSO. Pró-Reitoria de Planejamento. *Série histórica de cursos e vagas (Reuni)*. Cuiabá: UFMT, 2018.

UNIVERSIDADE FEDERAL DE MATO GROSSO. Pró-Reitoria de Planejamento. Sistema de Informações de Gestão Acadêmica. *Lista e descrição das variáveis de entrada e seus respectivos níveis de mensuração*. Documento interno. Cuiabá: UFMT, 2024.