South American Development Society Journal

Vol.: 10 | Nº.: 30 | **Ano:** 2024 | **ISSN:** 2446-5763 | **DOI:** 10.24325/issn.2446-5763.v10i30p182-195
Data de Submissão: 01/09/2024 | **Data de Publicação:** 22/12/2024

MODELOS PREDITIVOS DA EVASÃO DE ESTUDANTES: COMPARATIVO ENTRE REGRESSÃO LOGÍSTICA, REDES NEURAIS E ALGORITMOS GENÉTICOS

Diogo Martins Gonçalves de Morais
Faculdade Engenheiro Salvador Arena/Instituto de Tecnologia e Liderança
pro7113@cefsa.edu.br

Dallas Kelson Francisco de Souza
Universidade Estadual de Campinas
dallaskelson@gmail.com

Resumo

O presente estudo tem como objetivo realizar uma análise técnica e comparativa da acurácia de modelos preditivos de evasão de estudantes, explorando três abordagens distintas: algoritmo genético, redes neurais e regressão logística. Uma pesquisa aplicada e exploratória foi conduzida, envolvendo uma revisão da literatura para validar a relevância dos algoritmos em contextos educacionais e avaliar as variáveis influenciadoras da evasão no ensino superior. O resultado da revisão da literatura permitiu constatar uma significativa presença desses algoritmos em modelos preditivos de evasão e confirmou a pertinência das variáveis utilizadas. Além disso, o estudo detalhou como os três algoritmos foram operacionalizados e apresentou uma análise comparativa do desempenho dos modelos, utilizando dados de 447 estudantes de Administração e Engenharia de uma instituição privada, abrangendo 11 variáveis influenciadoras da evasão. Os resultados indicam uma acurácia de 90% para o modelo elaborado com regressão logística, 84,4% para o modelo elaborado com Redes Neurais, e 76% para o modelo elaborado com algoritmo genético, evidenciando uma maior eficiência do modelo logístico em prever evasão de estudantes na amostra da presente pesquisa.

Palavras-chave: Modelos preditivos; Machine Learning; Gestão Educacional

Abstract

This study aims to perform a technical and comparative analysis of the accuracy of predictive models for student dropout, exploring three distinct approaches: genetic algorithms, neural networks, and logistic regression. An applied and exploratory research was conducted, involving a literature review to validate the relevance of these algorithms in educational contexts and to assess the influencing variables of dropout in higher education. The literature review revealed a significant presence of these algorithms in predictive dropout models and confirmed the pertinence of the variables used. Moreover, the study detailed how the three algorithms were operationalized and presented a comparative performance analysis of the models using data from 447 students of Business Administration and Engineering from a private institution, covering 11 influencing variables of dropout. The results indicate an accuracy of 90% for the model developed with logistic regression, 84.4% for the model developed with neural networks, and 76% for the model developed with a genetic algorithm, highlighting a greater efficiency of the logistic model in predicting student dropout in the sample of this research.

Keywords: Predictive models; Machine Learning; Educational Management

Introdução

Os dados do Instituto Nacional de Pesquisas Educacionais Anísio Teixeira (INEP, 2024) mostram uma significativa expansão do ensino superior no Brasil. O número de Instituições de Ensino Superior (IES) cresceu de, aproximadamente, 2,3 mil para 2,5 mil, enquanto o número de estudantes matriculados aumentou de 6,7 milhões para 8,9 milhões no período de 2011 a 2021. Ao analisar as categorias administrativas das IES brasileiras que absorveram esses alunos, observa-se que 77% dos alunos matriculados no ensino superior estão em IES privadas, ao passo que 23% estão IES públicas.

Embora os números representem um cenário positivo, por se tratar de um aumento no acesso à educação superior, os dados também sugerem que os responsáveis pela gestão das IES enfrentam o desafio da evasão de estudantes na Educação Superior. Isso porque, de acordo com os dados do INEP (2024), a taxa de

desistência acumulada foi de 59%. Ou seja, entre 2012 e 2021, em média, a cada cem alunos matriculados, aproximadamente 59 desistiram do curso, uma proporção que ressalta a necessidade de estratégias eficazes para minimizar a evasão nas IES.

Dessa maneira, compreender os fatores que influenciam a evasão dos estudantes e buscar ferramentas para minimizá-la constitui um recurso estratégico para a gestão das instituições de ensino superior. Nesse contexto, a literatura científica apresenta diversos estudos que propõem o desenvolvimento de modelos preditivos de evasão de estudantes a partir dos dados de perfis desses indivíduos, obtidos na ocasião do vestibular e no decorrer do curso, tais como os de Morais, Souza e Cassoni (2020), Silva, Cabral e Pacheco (2020), Lopes Filho e Silveira (2021), Silva (2022) e Osório e Santacoloma (2023). Essas pesquisas demonstram a contemporaneidade e relevância do tema, visto que a predição desse tipo de evento pode auxiliar a tomada de decisão estratégica nas instituições, gerando resultados positivos ao longo do tempo.

Diante disso, o presente estudo tem como objetivo realizar uma análise técnica e comparativa da performance de três técnicas distintas para a previsão de evasão de estudantes. Assim, por meio de um Algoritmo Genético, Redes Neurais e Regressão Logística, busca-se identificar as abordagens de modelos preditivos de evasão de estudantes através de uma revisão da literatura, com a finalidade de validar a presença dos algoritmos que serão estudados nesse contexto educacional. Além disso, pretende-se identificar os fatores que influenciam a evasão dos estudantes com a mesma revisão da literatura, visando validar as variáveis utilizadas na construção dos modelos que serão comparados. Por fim, busca-se construir um modelo preditivo da evasão de estudantes com três abordagens distintas – o Algoritmo Genético, a Regressão Logística e as Redes Neurais – escrito na linguagem Python, a partir do mesmo conjunto de dados, a fim de comparar a performance dos três modelos.

O presente estudo foi realizado com dados de uma IES privada. Para essas instituições, altas taxas de evasão podem resultar em perdas financeiras, comprometendo a reputação da instituição e prejudicando, consequentemente, o planejamento a longo prazo. Além disso, a evasão do curso pode causar prejuízo financeiro para os alunos, afetando negativamente seus planos de carreira. Dessa forma, o presente estudo visa contribuir para a tomada de decisão estratégica das IES, bem como subsidiar políticas públicas que busquem minimizar as desistências dos estudantes ao longo dos anos.

Material e Métodos

O presente estudo pode ser caracterizado como uma pesquisa exploratória, pois possibilitará o aprofundamento da discussão sobre o fenômeno da evasão de estudantes, além da construção e análise de novos cenários hipotéticos. Conforme abordado por Gil (2017), a pesquisa exploratória busca proporcionar maior familiaridade com o problema, tornando-o mais explícito, permitindo a construção de hipóteses. Seu planejamento possui caráter flexível, uma vez que considera diversos aspectos da temática estudada.

Considerando que os cenários simulados foram construídos a partir da manipulação de dados, aplicação de algoritmos de Aprendizado de Máquina e análise de performance dos algoritmos, o estudo também pode ser caracterizado como uma pesquisa aplicada de natureza quantitativa.

Os dados utilizados para a construção do modelo preditivo da evasão de estudantes com o uso do Algoritmo Genético, Regressão Logística e Redes Neurais foram cedidos pelos pesquisadores e autores do artigo intitulado “Um Modelo Preditivo da Evasão de Estudantes no Ensino Superior”, escrito por Morais, Souza e Cassoni (2020), publicado em 02 de fevereiro de 2020, no periódico FTT Journal of Engineering and Business. Os autores utilizaram dados de 448 estudantes do ensino superior, distribuídos em quatro cursos de graduação: Administração, Engenharia de Alimentos, Engenharia de Controle e Automação e Engenharia de Computação, oriundos de uma faculdade privada de pequeno porte do município de São Bernardo do Campo/SP.

Com o uso do Algoritmo Genético e a partir de 11 variáveis explicativas – idade, instituição de ensino médio, quem incentivou a cursar graduação, familiar que concluiu o ensino superior, ano de conclusão do ensino médio, ocupação do pai, ocupação da mãe, nota final no vestibular, tipo de curso de ensino médio, situação e curso –, o modelo apresentado pelos autores reportou uma taxa de acerto de 76%.

O presente estudo propõe a elaboração do modelo preditivo com os mesmos dados de Morais, Souza e Cassoni (2020), diferenciando-se pelo uso de algoritmos de Aprendizado de Máquina adicionais e distintos, como a Regressão Logística e as Redes Neurais, a fim de realizar a comparação de performance e discutir os resultados no âmbito da ciência de dados.

Dessa maneira, as 11 variáveis explicativas usadas no Algoritmo Genético serão utilizadas como variáveis de entrada da Rede Neural proposta e como variáveis independentes no modelo de Regressão Logística proposto. De forma análoga, a variável predita no Algoritmo Genético que se refere à situação de “matriculado” ou “evadido” do curso, será representada como a probabilidade de um estudante evadir do curso na Regressão Logística e como a função de saída na Rede Neural proposta.

Resultados e Discussão

Nesta seção, são apresentados os resultados do presente estudo, relacionados ao atendimento do objetivo que se referem à revisão da literatura, de modo a identificar as abordagens dos modelos preditivos de evasão de estudantes usados no Brasil, assim como o tipo de algoritmo, as variáveis que influenciam a evasão e outras características dos modelos, o referencial teórico que fundamenta os algoritmos envolvidos no estudo, a elaboração dos modelos preditivos com o mesmo banco de dados utilizado por Morais, Souza e Cassoni (2020), porém com abordagens distintas, e, por fim, a comparação dos desempenhos dos modelos preditivos.

Resultado da pesquisa sobre as características dos modelos preditivos de evasão de estudantes do Ensino Superior

A revisão da literatura foi realizada por meio da plataforma de periódicos da Capes e do Google Acadêmico, com o uso da combinação das palavras-chave: modelos preditivos, análise preditiva, evasão e ensino superior. Além disso, em busca de evidências mais recentes, a pesquisa foi limitada ao período entre 2019 e 2023. A partir dessas características, identificou-se 43 artigos, dissertações e teses, dos quais apenas 11 estudos foram selecionados por envolverem diretamente a elaboração completa e aplicação de modelos preditivos de evasão de alunos no ensino superior no Brasil. Dos 11 artigos avaliados, apenas 5 apresentaram modelos preditivos juntamente com os indicadores de performance utilizados, permitindo uma comparação mais precisa sobre a acurácia dos modelos. Esses estudos são apresentados no Quadro 1.

Quadro 1 - Quadro Sinóptico sobre os algoritmos usados nos modelos preditivos

Título	Autores	Ano	Algoritmo	Variáveis	Acurácia

				influenciadoras da evasão	
Combinando Técnicas de Mineração de Dados para Melhorar a Detecção de Indicadores de Evasão Universitária	CARRANO, D.; TULER, E.; ROCHA, L.	2019	Árvores de decisão	Local de residência, idade, gênero, satisfação com o curso, perfil socioeconômico e desempenho acadêmico.	81%
Evasão ou permanência? Modelos preditivos para a gestão do Ensino Superior	SILVA, F. C.; CABRAL, T. L. O.; PACHECO, A. S. V.	2020	Regressão Logística Binária	Desempenho acadêmico, desempenho no vestibular, local de residência e idade	81,93%
Um Modelo Preditivo da Evasão de Estudantes no Ensino Superior	MORAIS, D.M.G.; SOUZA, A. A. M.; CASSONI, V.	2020	Algoritmo Genético	Desempenho acadêmico, desempenho no vestibular, perfil socioeconômico	76%
Predição de Evasão Escolar na Licenciatura em Computação	JESUS, H. O.; RODRIGUEZ, L.C.; COSTA JUNIOR, A. O.	2021	Rede Neural	Desempenho acadêmico, nível de frequência nas aulas, período matriculado no curso e carga horária do curso.	98%
Predictive Model to Identify College Students with High Dropout Rates	OSORIO, J. K. H.; SANTACOLOMA, G. D.	2023	Regressão Logística Binária	Desempenho acadêmico, desempenho na educação básica, perfil socioeconômico e psicossocial, e idade.	61,97%

Fonte: elaboração própria

Por meio do Quadro 1, é possível observar a presença de estudos que abordaram a problemática da evasão de estudantes no ensino superior por meio de modelos preditivos, em que os principais algoritmos usados foram Árvores de Decisão, Algoritmo Genético, Regressão Logística Binária e Rede Neural.

Verifica-se que os modelos apresentados possuem bom desempenho e aderência entre si, considerando as variáveis de influência da evasão utilizadas nos modelos. Dessa forma, é possível validar o uso do Algoritmo Genético, Regressão Logística e Rede Neural nesse contexto educacional.

Resultado da pesquisa sobre o referencial teórico dos algoritmos utilizados nos modelos preditos da evasão de estudantes

Algoritmo Genético nos modelos preditivos de evasão

O Algoritmo Genético é uma técnica de otimização baseada nos princípios da genética e da seleção natural. Inspirado no processo evolutivo biológico, este algoritmo simula a sobrevivência do mais apto para encontrar soluções ótimas para problemas complexos. Ele opera através de uma população de indivíduos, cada um representando uma solução potencial, e usa operadores como seleção, cruzamento (ou recombinação) e mutação para evoluir esses indivíduos ao longo de várias gerações (HOLLAND, 1992).

No presente estudo, define-se “modelo preditivo da evasão de estudantes” como uma função matemática capaz de prever se um estudante será classificado como evadido ou matriculado (variável dependente “y”) a partir de um conjunto “n” de variáveis influenciadoras (variável independente “x”). O Algoritmo Genético inicia com um conjunto de funções matemáticas que são geradas de forma aleatória, que competem entre si na busca da melhor eficácia de previsão. Neste contexto, cada função concorrente é considerada um indivíduo “j” de uma população, como apresentado na eq. (1).

$$y_j = \alpha + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n, \quad (1)$$

Onde α representa o intercepto, e β_1, β_2, ... β_n são os coeficientes das variáveis independentes x_1, x_2, ... x_n. A variável dependente y_i representa a condição do estudante diante da instituição de ensino: o estudante será considerado “matriculado” quando $y_j > 0$, e “evadido” quando $y_j < 0$. As variáveis independentes x_i representam os fatores que influenciam a evasão, como o desempenho acadêmico, desempenho no vestibular, tipo de escola de ensino médio, grau de instrução dos pais, fatores socioeconômicos, entre outros. Dessa forma, os coeficientes α, β_1, β_2, ..., β_n são parâmetros a serem estimados pelo Algoritmo Genético, inicialmente gerados de forma aleatória.

O Algoritmo Genético inicia o processo de busca da melhor função com uma população inicial criada a partir da geração aleatória dos coeficientes α, β_1, β_2, ..., β_n.

Essa população inicial é testada com um banco de dados de 448 estudantes, para os quais os valores das variáveis x_i e a respectivas situações y_j (matriculado ou evadido) já são conhecidos.

Durante o teste, cada indivíduo (função matemática) na população recebe um percentual de acertos e erros, conhecido como “fitness”, calculado conforme a eq. (2).

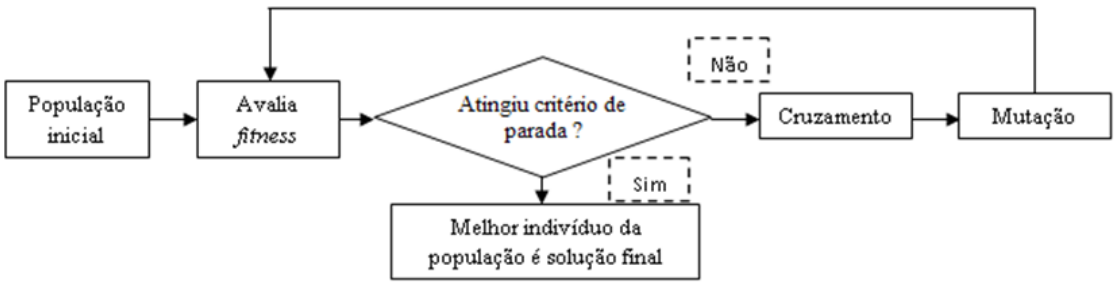
$$Fitness = \left(\frac{n° \, de \, matriculados \, identificados}{n° \, de \, matriculados \, do \, banco \, de \, dados}\right) x \left(\frac{n° \, de \, evadidos \, identificados}{n° \, de \, evadidos \, do \, banco \, de \, dados}\right) \quad (2)$$

O “fitness” é calculado como o produto de duas razões: a primeira razão é o número de estudantes matriculados identificados corretamente pela função matemática e o número total de matriculados no banco de dados, e a segunda razão é dada pelo número de evadidos identificados corretamente pela função matemática e o número total de evadidos do banco de dados.

Depois de calcular o “fitness” de toda a população inicial, ocorre o cruzamento de dois indivíduos, escolhidos por meio de um sorteio ponderado por probabilidades diferentes. Essa ponderação é realizada de acordo com o Fitness calculado, ou seja, os indivíduos com maior “fitness” têm maiores probabilidades de serem sorteados.

Após a seleção dos dois indivíduos, realiza-se o cruzamento dessas duas funções, gerando um conjunto de novas funções, chamadas de filhos e mutantes. A Figura 1 ilustra o funcionamento do Algoritmo Genético nesse contexto.

Figura 1 – Funcionamento do Algoritmo Genético



Fonte: elaboração própria.

Por fim, o teste é realizado novamente com o uso do banco de dados em todos esses novos indivíduos e aqueles com os maiores valores de “fitness” prevalecem em uma nova população, que passa pelo mesmo processo até que um indivíduo atinja um valor de “fitness” que seja estacionário ou adequado para a interrupção da busca.

Regressão Logística nos modelos preditivos de evasão

A Regressão Logística é um modelo estatístico-matemático que relaciona uma variável dependente y a uma variável independente x a partir de uma relação não linear. Esse modelo é apropriada em casos em que o fenômeno se apresenta de forma qualitativa, sendo necessário transformar variável qualitativa em variável quantitativa. Para isso, utiliza-se a variável y em um formato binário onde, 0 representa um evento e 1 outro evento (FÁVERO e BELFIORE, 2020).

Nesse caso, ao final, obtém-se um modelo matemático que permite calcular a probabilidade de ocorrência de um fenômeno a partir das variáveis independentes, como expresso pela eq. (3).

$$p_i = \frac{1}{1+e^{-(\alpha+\beta_1 x_{1i}+\beta_2 x_{2i}+\cdots+\beta_k x_{ki})}} \quad (3)$$

Onde p_i é a probabilidade do evento Y ocorrer; α representa uma constante; β_j (j = 1, 2, ..., *k*) são os parâmetros logísticos estimados para cada variável independente; x_j (*j*=1, 2, ..., k) são as variáveis independentes; e i representa uma observação da amostra (HAIR, 2019; FÁVERO, 2020).

No presente estudo, a variável y representa a probabilidade de um estudante evadir do curso, enquanto as variáveis independentes x_j representam fatores que influenciam a evasão, como o desempenho acadêmico, desempenho no vestibular, tipo de escola de ensino médio, grau de instrução dos pais, fatores socioeconômicos, entre outros.

Redes Neurais nos modelos preditivos de evasão

Uma Rede Neural Artificial é um modelo computacional inspirado no funcionamento do cérebro humano, composta por unidades básicas chamadas neurônios artificiais, organizados em camadas e interconectados por sinapses. Cada neurônio recebe sinais de entrada, processa-os e transmite sinais de saída (SKANSI, 2018).

As saídas das Redes Neurais são funções dadas pela soma ponderada das entradas e um viés, onde cada neurônio ativa-se ao exceder um limiar definido por sua função de ativação, que pode ser linear, sigmoide ou tangente hiperbólica. A função sigmoide, em particular, é amplamente usada por sua habilidade em manejar não linearidades e complexidades dos dados, tornando-a uma escolha eficaz para estas redes.

O aprendizado em uma rede neural ocorre por meio do ajuste dos pesos sinápticos, geralmente através de um processo conhecido como treinamento, no qual a rede é exposta a um grande conjunto de dados e, gradativamente, melhora sua performance na tarefa desejada.

Dessa forma, a saída de cada neurônio de uma rede neural pode ser calculada pela eq. (4), onde "x_i" representa cada variável de entrada no neurônio e "w_i" representa seu respectivo peso.

$$y(x) = \sum_{i=1}^{n} x_i \,.\, w_i + viés \tag{4}$$

O processamento das informações desde a camada de entrada, passando pelas camadas seguintes e chegando à camada de saída, é conhecido como "forward propagation" (ou propagação direta); seu valor é calculado pela eq. (4). Uma vez que a informação alcance a camada de saída, deve-se calcular a taxa de acerto do modelo, que é necessária para corrigir os valores dos pesos ""w_i" utilizados no algoritmo.

Os novos valores para os pesos são calculados por meio de uma função de otimização, denominada gradiente descendente. Esse processo é conhecido como "backpropagation" e utiliza o valor da derivada parcial da função de ativação de cada neurônio para identificar a inclinação de cada um dos pesos. Assim, o "backpropagation" continua mudando o valor dos pesos até encontrar uma redução que seja proporcional à taxa de aprendizado.

No presente estudo, pretende-se obter a função de saída y(x) (vide eq. 4) que tenha passado por esse processamento, que representa a probabilidade de um estudante evadir do curso, enquanto as variáveis independentes x_i representam fatores de influência da evasão considerados na presente pesquisa.

O desempenho dos modelos preditivos de evasão de estudantes

Até a seção anterior, esta pesquisa buscou apresentar as abordagens e características dos modelos preditivos da evasão dos estudantes do ensino superior na literatura, e apresentou o funcionamento dos três algoritmos frequentemente usados na elaboração desses modelos. Nas próximas seções, o presente estudo apresentará o resultado do objetivo específico deste Trabalho de Conclusão de Curso, que converge para o objetivo geral deste estudo: realizar uma análise técnica e comparativa do desempenho de um modelo preditivo da evasão de estudantes, com base nas técnicas de Algoritmo Genético, Redes Neurais e Regressão Logística.

Análise comparativa entre o Algoritmo Genético e a Regressão Logística

Para obter o modelo de Regressão Logística a partir do conjunto de dados de evasão de estudantes, inicialmente foram considerados 448 estudantes, mas esse número foi reduzido para 447 após a exclusão de um aluno devido a dados faltantes. Além disso, realizou-se a transformação de algumas variáveis categóricas em variáveis dummy, condição necessária para o uso da Regressão Logística. A variável dependente (situação do estudante) foi codificada numericamente, sendo transformada em duas categorias, com os valores 0 e 1 para representar os estados de evasão e matrícula, respectivamente.

Por fim, os dados foram divididos em dois conjuntos: um conjunto de treinamento com 357 dados (80% da amostra) para ajustar o modelo e um conjunto de teste com 90 dados (20% da amostra) para avaliar seu desempenho.

A equação logística usada no modelo preditivo, com todos os coeficientes estimados, é apresentada na Equação 5.

$$\log\left(\frac{p}{1-p}\right) = -1{,}3486 + 0{,}2054\times\text{Idade} - 0{,}1084\times\text{Tipo de Ensino Médio} - 0{,}1963\times\text{Tipo de Escola} - 0{,}0874\times\text{Incentivo} - 0{,}2082\times\text{Familia} - 0{,}0379\times\text{Ano de Ingresso} - 0{,}0632\times\text{Profissão do Pai} - 0{,}0820\times\text{Profissão da Mãe} - 0{,}0019\times\text{Vestibular} + 0{,}5388\times\text{Tipo de Vaga} + 0{,}2562\times\text{Curso} \quad (5)$$

A acurácia do modelo apresentado na Equação 5 foi de 90%. Essa medida resulta da razão entre o número de previsões corretas pelo modelo e o número total de previsões. A avaliação foi baseada na amostra denominada "amostra teste", composta por 20% dos dados. Dessa forma, a acurácia de 90% é considerada acima do satisfatório. Em comparação, Morais, Souza e Cassoni (2020) encontraram uma acurácia de 76% utilizando o Algoritmo Genético.

Análise comparativa entre o Algoritmo Genético, a Regressão Logística e a Rede Neural Artificial

Para obter a Rede Neural Artificial a partir do conjunto de dados de evasão de estudantes, as variáveis preditoras foram normalizadas, e todas as variáveis categóricas foram convertidas para formatos numéricos. Utilizou-se, por conveniência, uma Rede Neural com duas camadas ocultas, cada uma contendo 10 neurônios. Ressalta-se que essa escolha foi arbitrária, visando encontrar um equilíbrio entre a capacidade de modelagem e a eficiência computacional. Além disso, a função utilizada para fazer as previsões é baseada na função sigmoide aplicada à saída da última camada da rede.

A saída dessa função está no intervalo (0, 1), representando a probabilidade de a instância pertencer à classe positiva (neste caso, "Evadido"). Se essa probabilidade for maior que 0,5, a previsão é "Evadido"; se for menor, a previsão é "Matriculado". Dessa forma, a acurácia do modelo no conjunto de teste foi de, aproximadamente, 84,4%. Isso indica um desempenho inferior ao do modelo de Regressão Logística para este conjunto de dados específico, porém ainda superior a acurácia de 76% da modelagem realizada com o mesmo banco de dados, com o uso do Algoritmo Genético.

Dessa maneira, pode-se concluir que a Regressão Logística apresenta o melhor desempenho para este tipo de aplicação. Além disso, a modelagem com o uso de Regressão Logística traz outras vantagens, como a sua simplicidade e interpretabilidade, permitindo que cientistas de dados e gestores educacionais compreendam facilmente quais variáveis influenciam mais significativamente a evasão de estudantes.

Considerações Finais

O presente estudo realizou uma comparação técnica entre três metodologias de modelagem preditiva para identificar evasão de estudantes: Regressão Logística, Redes Neurais e Algoritmos Genéticos, sendo esta última abordagem usada como referência para as demais, por se tratar de um estudo publicado. Conclui-se, neste contexto, que a Regressão Logística demonstrou superioridade, alcançando uma acurácia de 90%, seguida pelas Redes Neurais, com 84,4%, e pelo Algoritmo Genético, com 76%.

Para além do desempenho medido pela acurácia, a comparação das três abordagens em um mesmo contexto e com o mesmo banco de dados sugere que a simplicidade e interpretabilidade da Regressão Logística tornam essa técnica como recomendável para o uso de cientistas de dados e gestores educacionais. Constatou-se também que os modelos que envolvem aprendizado de máquina podem ser aplicados em contextos educacionais, com aplicações factíveis e dados acessíveis aos gestores, como coordenadores de cursos, e podem ser usados como ferramenta estratégica para mitigar o problema da evasão de estudantes.

Não obstante, reconhece-se que não é possível fazer nenhum tipo de generalização sobre a superioridade da Regressão Logística sobre as demais escolhas algorítmicas, visto que este estudo envolveu um conjunto de dados específicos de estudantes do Ensino Superior, de cursos específicos de uma Faculdade privada da região metropolitana de São Paulo. Recomenda-se, para estudos futuros, a comparação dos desempenhos destes modelos em outros contextos.

Por fim, recomenda-se que a escolha do modelo mais adequado para prever a evasão estudantil transcenda a análise de acurácia, englobando fatores como implementação, interpretabilidade dos resultados e disponibilidade de dados.

Referências

FÁVERO, L.P.; BELFIORE, P. **Data science for business and decision making**. Academic Press: Cambridge, 2019.

FÁVERO, L. P.; BELFIORE, P. **Manual de análise de dados**. LTC: Rio de Janeiro, 2020.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 6. ed. Atlas: São Paulo, 2017.

HAIR, J. F., BLACK, W. C., BABIN, B. J., ANDERSON, R. E. **Multivariate Data Analysis**. Cengage: London, 2019.

HOLLAND, John H. **Adaptation in Natural & Artificial Systems**: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence. Cambridge: MIT Press, 1992.

Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira [INEP]. 2024. **Censo da Educação Superior**. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior > Acesso em: 8 de agosto de 2024.

LOPES FILHO, J. A. B.; SILVEIRA, I. F. Detecção precoce de estudantes em risco de evasão usando dados administrativos e aprendizagem de máquina. **Revista Ibérica de Sistemas e Tecnologias de Informação**, v. 40, n. 1, 2021.

MORAIS, D. M. G. DE; SOUZA, A. A. M. DE; CASSONI, V. Um modelo preditivo da evasão de estudantes no ensino superior. **FTT Journal of Engineering and Business**, v. 1, n. 5, fev. 2020.

OSORIO, J. K. H.; SANTACOLOMA, G. D. Predictive Model to Identify College Students with High Dropout Rates. **Revista Electrónica de Investigación Educativa**, v.25, n.13, 2023.

SKANSI, S. **Introduction to Deep Learning**: From Logical Calculus to Artificial Intelligence. New York: Springer Publishing Company Incorporated, 2018.

SILVA, F. C.; CABRAL, T. L. O.; PACHECO, A. S. V. Evasão ou permanência? Modelos preditivos para a gestão do Ensino Superior**. Arquivos Analíticos de Políticas Educativas**, v. 28, n. 149, out. 2020.

SILVA, J. J. **Uma comparação de técnicas de Aprendizado de Máquina para a predição de evasão de estudantes no ensino público superior**. 2022. 77 f. Dissertação de Mestrado - Universidade de São Paulo, São Paulo.

SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 30, 2019. Brasília. **Anais** [...]. Brasília: Congresso Brasileiro de Informática na Educação, 2019. 10 p. Tema: A Computação na perspectiva da diversidade, inclusão e inovação na educação para o século XXI.