MODELAGEM E PREVISÃO DE EVASÃO DE ESTUDANTES EM CURSO DE ENGENHARIA DE COMPUTAÇÃO UTILIZANDO APRENDIZADO DE MÁQUINA

DOI: 10.37702/2175-957X.COBENGE.2024.5122

Autores: *GABRIELLY BARCELOS CARIMAN, ROBERTA LIMA GOMES*

Resumo: *As altas taxas de evasÃ£o em cursos de ensino superior focados em tecnologia contribuem significativamente para a escassez de profissionais de TI. Para abordar essa questÃ£o, o Departamento de Engenharia de ComputaÃ§Ã£o da universidade conduziu uma pesquisa visando um diagnÃ³stico precoce e preciso do desempenho dos alunos. Isso permite que intervenÃ§Ãµes sejam feitas a tempo de aumentar as taxas de conclusÃ£o. Modelos de aprendizado de mÃ¡quina foram empregados para prever o desempenho dos alunos e a probabilidade de evasÃ£o. Modelos de regressÃ£o logÃstica e Ãrvore de decisÃ£o foram treinados e avaliados, apresentando resultados promissores com mais de 80% de acurÃ¡cia na previsÃ£o de evasÃ£o.*

Palavras-chave: *PrevisÃ£o de evasÃ£o, EvasÃ£o no ensino superior, Modelos de aprendizado de mÃ¡quina.*

MODELAGEM E PREVISÃO DE EVASÃO DE ESTUDANTES EM CURSO DE ENGENHARIA DE COMPUTAÇÃO UTILIZANDO APRENDIZADO DE MÁQUINA

1 INTRODUÇÃO

Com a evolução da engenharia em quase todos os setores da sociedade, a demanda por profissionais de Tecnologia da Informação e Comunicação (TIC) vem crescendo a cada ano (Brasscom, 2021). A pandemia da COVID-19 evidenciou ainda mais essa carência, em decorrência de uma massiva digitalização de processos e serviços. Apesar da necessidade de atrair mais profissionais, observa-se que os índices de evasão nos cursos de nível superior em TIC estão entre os mais altos do país (Garcia; Gomes, 2022). Também, são alarmantes os índices de retenção, evidenciando as dificuldades enfrentadas pelos estudantes em se adaptarem às demandas dos cursos.

No intuito de superar esses obstáculos, as instituições de ensino superior estão adotando medidas para apoiar os estudantes, conforme orientado pelo Plano de Desenvolvimento Institucional Universidade 2021-2030. Exemplificando, os programas PaEPE oferecem bolsas a estudantes de graduação para atividades de monitoria e apoio administrativo. O Colegiado de Engenharia de Computação do Centro Tecnológico realizou uma ação conjunta com bolsistas PaEPE em 2022. O objetivo foi diagnosticar o desempenho futuro dos estudantes com base em seus percursos formativos, permitindo intervenções precoces para aumentar a taxa de conclusão dos cursos. O estudo analisou os históricos curriculares de todos os estudantes, inclusive egressos e evadidos, e utilizou aprendizado de máquina (regressão logística e árvore de decisão) para prever a evasão dos alunos, comparando a acurácia dos modelos.

Por fim, este artigo segue a estrutura a seguir. Inicialmente, na Seção 2, são delineadas as definições e métricas utilizadas para analisar a evasão e retenção nos cursos de Engenharia, contextualizando o cenário educacional brasileiro. Em seguida, na Seção 3, são apresentados os detalhes dos dados utilizados para a modelagem e previsão. Posteriormente, na Seção 4, é descrita a metodologia adotada. Logo após, na Seção 5, são expostos os resultados da aplicação de modelos de aprendizado de máquina na previsão de evasão, ressaltando as particularidades encontradas nos cursos de Engenharia de Computação. Finalmente, na Seção 6, são apresentadas as considerações finais e são indicadas possíveis direções futuras.

2 EVASÃO E RETENÇÃO NOS CURSOS DE ENGENHARIA

Várias definições para evasão podem ser encontradas na literatura (Hoed, 2017), representando um fenômeno complexo relacionado a diferentes fatores sociais, técnicos e econômicos. Uma das formas de se mensurá-la é por meio da Taxa de Desistência Acumulada (TDA) representando a “saída antecipada, antes da conclusão do ano, série ou ciclo, por desistência [..]” (MEC, 2017).

Segundo o Censo da Educação Superior de 2021 (INEP, 2022), a média nacional para a TDA entre 2012 e 2021 dos cursos de graduação presenciais foi de 59%. De acordo com Hoed (2017), os cursos nas áreas de Matemática e Computação estão entre os que mais contribuem para essa média expressiva. De fato, os dados apresentados pela plataforma disponibilizada pelo INEP indicam uma TDA de 70% para o mesmo

período, se considerados apenas os cursos da área de "Computação e Tecnologia da Informação e Comunicação".

Uma problemática que pode estar diretamente relacionada à evasão são os índices de retenção, que implicam em um tempo adicional levado pelo estudante para completar o curso superior, podendo esse atraso acarretar na evasão do estudante (Campello; Lins, 2008). Geralmente, a retenção está associada ao baixo rendimento acadêmico. De acordo com Bernardo *et al.* (2017), o desempenho acadêmico é a variável que mais influência (dentre um conjunto de 46 estudadas) na decisão do estudante abandonar ou não um curso. Alguns trabalhos ainda apontam que reprovações são ainda mais acentuadas nos primeiros anos de curso, em disciplinas do ciclo básico (De Almeida, Godoy, 2016; Nunes *et al.*, 2020).

Tendo isso em vista, a previsão do comportamento dos estudantes é fundamental para aprimorar o currículo e oferecer intervenções acadêmicas adequadas. Identificar as causas da evasão acadêmica e adotar medidas preventivas são desafios complexos, mas mecanismos automatizados podem ser promissores nesse combate. Portanto, na mesma linha do presente trabalho, alguns autores se apoiam na definição de modelos matemáticos/estatísticos, analisando, dentre outras informações, dados de retenção para permitir a identificação de perfis de estudante que apresentem maior potencial de evasão (De Almeida; Godoy, 2016; Hoed, 2017; Saccaro; França; Jacinto, 2019).

Particularmente, a Mineração de Dados (MD) e o Aprendizado de Máquina (AM) têm sido amplamente utilizados para analisar e extrair informações de conjuntos de dados, permitindo a previsão do desempenho dos estudantes e o risco de abandono escolar (Rastrollo-Guerrero, Gómez-Pulido e Durán-Domínguez, 2020; Alturki e Alturki, 2021) (Opazo *et al.*, 2021). Mas, alguns autores defendem que modelos específicos para cada universidade (Opazo *et al.*, 2021), ou para cada curso (Silva *et al.*, 2012) são mais eficazes do que um modelo geral, pois consideraram características individuais. Além disso, também é importante considerar informações mais abrangentes dos estudantes, como dados socioeconômicos, para aprimorar a precisão das previsões de evasão acadêmica (Marques *et al.*, 2020; Moreira Da Silva *et al.*, 2022; Rolim; Silva, 2021).

Desta forma, no presente trabalho foi conduzido um estudo para desenvolver modelos de previsão de evasão no curso de Engenharia de Computação da Universidade, utilizando técnicas de aprendizado de máquina, tais como Regressão Logística e Árvore de Decisão. Foram considerados dados como reprovações dos alunos, método de ingresso e informações socioeconômicas, representadas pela condição de cotista ou não.

3 DADOS UTILIZADOS NA MODELAGEM E PREVISÃO

Para a elaboração dos modelos utilizados nesta pesquisa, os dados foram obtidos através do sistema acadêmico da Universidade entre os anos de 1990 a 2023. Posteriormente, os dados foram exportados em formato de tabela estruturada (em arquivos CSV), totalizando 66.898 linhas.

As principais colunas das tabelas incluem informações como o número de matrícula do aluno (anonimizado pela coordenação de curso), o código do curso, o código da disciplina e a situação do item, indicando se o aluno foi aprovado na disciplina ou não. Outros dados importantes incluem o semestre, o Coeficiente de Rendimento Acadêmico (CRA), o ano e a forma de ingresso, além de informações sobre cotas.

Com o objetivo de prever a evasão, os dados foram agrupados de forma a facilitar o tratamento das informações. Foi criada uma nova coluna categórica denominada "grupo forma evasão", representando diferentes cenários, como "Formado", "Evadido", "Sem

Evasão" (estudantes ativos) e "Outros". Da mesma forma, os dados sobre a forma de ingresso foram agrupados na coluna "forma de ingresso", representando valores como "Vestibular", "Sisu" e "Outro". Além disso, a situação na disciplina foi agrupada em "Aprovado" e "Reprovado", com uma coluna adicional para indicar reprovações por frequência. Com esse novo conjunto de dados, a base totalizou 1.378 linhas.

Por fim, foram realizadas transformações adicionais, como a agregação dos dados por Matrícula de Estudante, o cálculo do tempo total de graduação para cada aluno e a análise do número de disciplinas não aprovadas por área de conhecimento e por ano de graduação, a fim de calcular a taxa de reprovação por área e por ano.

Na Tabela 1, encontra-se o resumo da descrição, apresentando as principais informações das colunas mais relevantes.

Tabela 1 - Resumo da descrição das principais colunas.

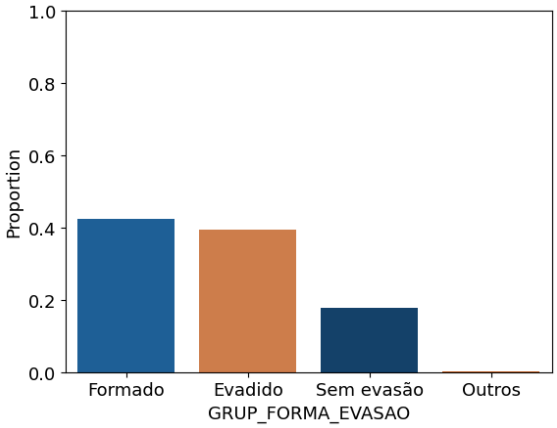
Variável	Descrição
Grupo Forma Evasão	Contém informações sobre os grupos de formas de evasão: ● Formado: valor 0 ● Evadido: valor 1 ● Sem Evasão: valor 2 ● Outros: valor 3
Forma de Ingresso	Coluna com as categorias de forma de ingresso: ● Vestibular ● Sisu ● Outro
Grupo Situação	Coluna com a informação do resultado de um aluno em uma disciplina: ● Aprovado: valor 0 ● Reprovado: valor 1
Reprovado por Frequência	● Recebe o valor 1 se reprovou por frequência em pelo menos uma matéria ● Recebe o valor 0 se nunca reprovou por frequência
Cotista	Coluna que indica se o estudante entrou como cotista ou não: ● Cotista: valor 0 ● Não Cotista: valor 1
Reprovada por Cursada	Cálculo do número de disciplinas reprovadas dividido pelo número de disciplinas feitas. ● Esse cálculo é aplicado no primeiro, segundo, terceiro, quarto e quinto ano de graduação ● Também é feito para as disciplinas de matemática, informática, elétrica e física ● Os resultados variam entre: ○ Não reprovou em nenhuma disciplina que fez: valor 0 ○ Reprovou em todas as disciplinas selecionadas: valor 1

Fonte: Produção dos próprios autores.

3.1 Distribuição dos dados

Inicialmente, ao analisar os dados sobre os grupos de formas de evasão, nota-se uma distribuição quase equitativa entre os estudantes que não estão mais ativos no curso e os ativos. Os formados representam 42,52%, os evadidos 39,55%, os sem evasão 17,71% e os outros 0,22%. Essa distribuição pode ser observada na Figura 1.

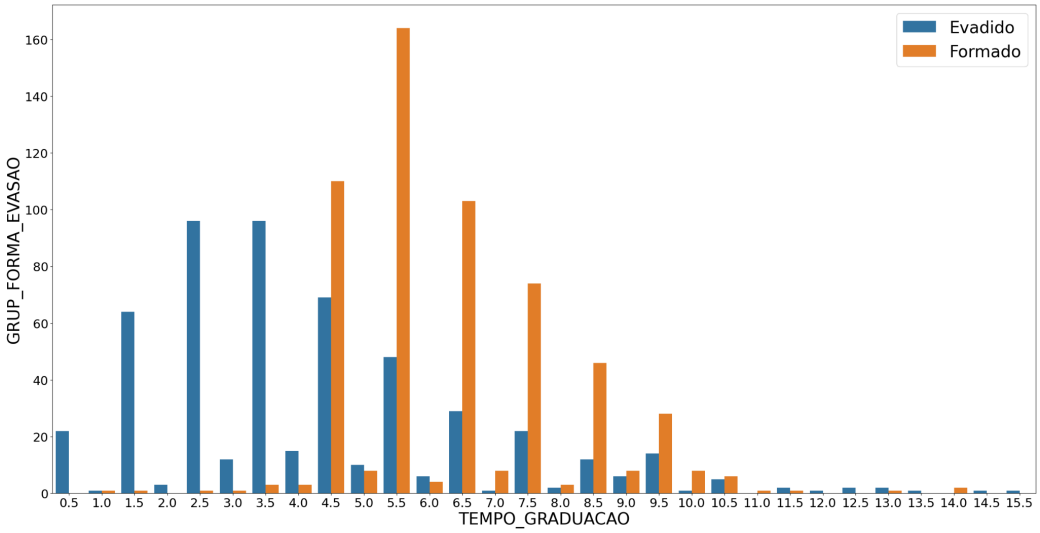
Figura 1 – Distribuição das formas de evasão.



Fonte: Produção dos próprios autores.

Ao examinar os dados referentes ao tempo de graduação, observamos uma distribuição mais assimétrica à direita. Portanto, a média é maior que a mediana, e a cauda direita é mais longa e esticada, como evidenciado na Figura 2. Isso se traduz no fato de que a mediana do tempo de graduação dos alunos que se formaram é de 5 anos e meio, enquanto dos que evadiram é de 3 anos e meio. Esses dados, também, corroboram a constatação de que a maioria dos estudantes que abandonam o curso o fazem em menos de 5 anos. Alguns estudantes que se formaram em menos tempo do que a duração mínima estabelecida para o curso podem ter concluído disciplinas anteriormente em outro curso ou faculdade, e depois aproveitado essas matérias. Assim, eles se formam em menos tempo do que o mínimo de 5 anos para Engenharia de Computação.

Figura 2 – Tempo de graduação em anos.



Fonte: Produção dos próprios autores.

4 METODOLOGIA EXPERIMENTAL

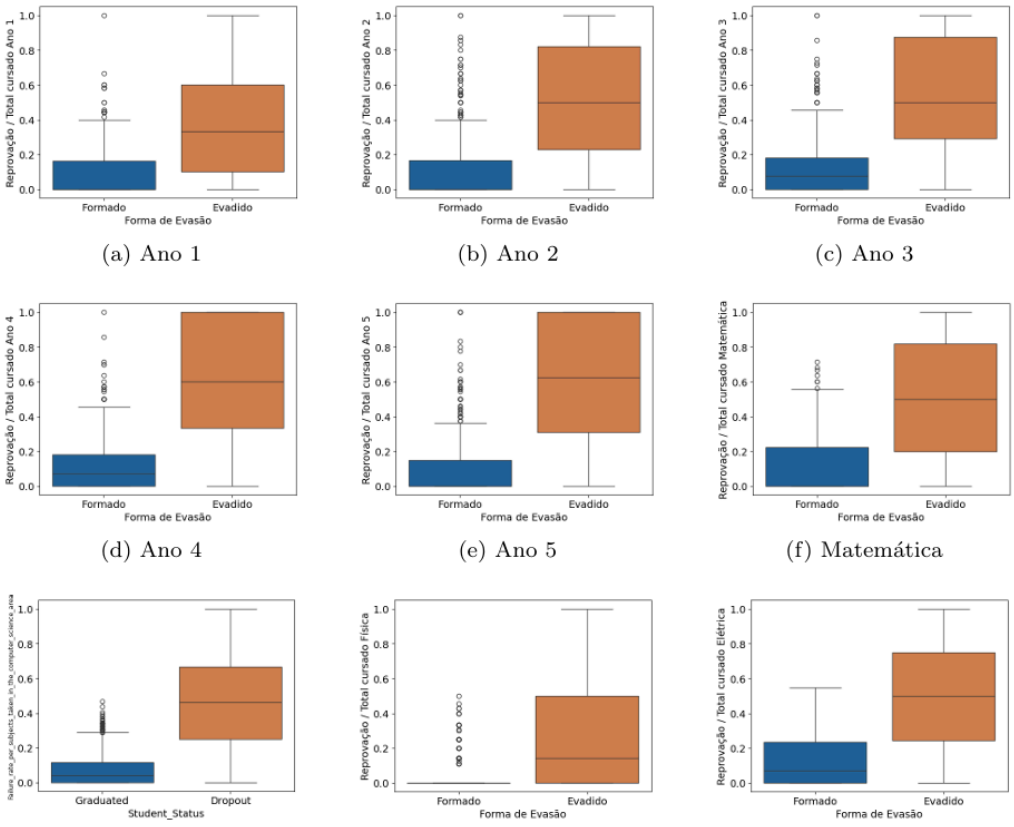
Nesta seção, será abordada a metodologia empregada na predição de evasão. Isso envolve uma análise de dados para selecionar as variáveis a serem utilizadas nos modelos de classificação, os modelos empregados para a classificação e o método de experimentação e avaliação de desempenho dos modelos.

4.1 Análise exploratória dos dados

Para selecionar as melhores variáveis nos modelos de aprendizado de máquina, foi conduzida uma análise bivariada utilizando gráficos de box plot (Morettin; Bussab, 2017) dos dados de reprovações, segmentados pelo número de disciplinas cursadas.

Ao examinar os box plots na Figura 3, que representam as reprovações por disciplinas cursadas, observa-se uma notável variação nos dados entre os graduados e os desistentes. Em geral, os graduados exibem uma mediana próxima de zero, sugerindo que a maioria das disciplinas cursadas por eles resultou em aprovação, apesar da presença de diversos valores atípicos superiores. Por outro lado, os desistentes mostram uma dispersão de dados mais ampla e uma mediana substancialmente maior, indicando que enfrentaram um número maior de reprovações em comparação com os graduados.

Figura 3 – Box plots dos dados de reprovação por disciplinas cursadas analisando a forma de evasão.



(a) Ano 1 (b) Ano 2 (c) Ano 3

(d) Ano 4 (e) Ano 5 (f) Matemática

(g) Informática (h) Física (i) Elétrica

Fonte: Produção dos próprios autores.

Considerando as disparidades notáveis de comportamento entre os grupos de graduados e desistentes, evidenciadas pelas variáveis de reprovação por disciplinas nos anos 1, 2, 3, 4 e 5, bem como nas disciplinas de Matemática, Informática, Física e Elétrica, essas variáveis serão utilizadas nos modelos de previsão de evasão. Além disso, serão incluídas as variáveis categóricas que indicam se o aluno reprovou por frequência, sua forma de ingresso e se é cotista ou não. Isso é motivado pela hipótese inicial de que essas três variáveis também desempenham um papel significativo na diferenciação de comportamentos entre alunos desistentes e formados.

4.2 Técnicas de aprendizado de máquina utilizadas

A área de Aprendizado de Máquina (machine learning) é fundamental na Inteligência Artificial, dedicada ao desenvolvimento de técnicas computacionais que permitem que sistemas adquiram conhecimento de forma automatizada, sendo capazes de fazer escolhas (Monard; Baranauskas, 2003).

O foco deste trabalho é a previsão da evasão de alunos, um problema de classificação em que os alunos são categorizados como evadidos ou não. O estudo parte do pressuposto de que métricas selecionadas oferecerão insights relevantes para o modelo, utilizando classificadores supervisionados básicos para obtenção de respostas. Inicialmente, são escolhidos dois classificadores para comparação: Regressão Logística (Logistic Regression) e Árvore de Decisão (Decision Tree).

A Regressão Logística é uma técnica estatística que constrói um modelo a partir de um conjunto de observações. Seu objetivo é prever os valores de uma variável categórica com base em variáveis independentes, contínuas ou binárias (Gonzalez, 2018). Modelando a probabilidade de um evento ocorrer de acordo com as variáveis independentes, ela utiliza a função logística, também conhecida como função sigmoid, para calcular a probabilidade de uma observação pertencer a uma categoria específica, variando de 0 a 1.

As árvores de decisão simplificam problemas complexos ao dividir os dados em conjuntos menores e identificar padrões não lineares entre as variáveis. Elas usam critérios de decisão para subdividir os dados criando, nós internos, representando escolhas e nós-folha indicando uma condição de parada. Esse processo continua até alcançar uma condição de término (Da Fonseca, 1994), resultando em uma estrutura binária. Uma das métricas utilizadas para classificação é a entropia, que mede a desordem em um sistema de dados. Durante o treinamento, é essencial evitar o overfitting (Monard; Baranauskas, 2003), onde a árvore se adapta demais aos dados de treinamento e compromete sua generalização, configurando hiperparâmetros, como a altura máxima da árvore, para limitar o número de nós-folha e garantir uma generalização adequada.

4.3 Método de Estimação de Desempenho

Inicialmente, a base é composta por um total de 1.378 alunos. Para prever quais destes podem evadir ou não, excluem-se os estudantes ativos (e outros que não podem ser considerados no modelo de previsão de evasão). Assim, a base passa a contabilizar 1.130 estudantes, incluindo apenas os que já concluíram o curso ou os que evadiram.

Considerando variáveis como "reprovação por cursadas", surge um problema com valores nulos para alunos que não estão mais no curso. Para lidar com isso, preenche-se os valores nulos com a mediana para evitar distorções. No entanto, em colunas com uma alta proporção de valores ausentes, preencher com a mediana pode não ser eficaz devido à falta de informações suficientes para representar com precisão os dados. Portanto, a coluna "REPROVADA_POR_CURSADA_ANO5", que contém o número de disciplinas reprovadas dividido pelo número total de disciplinas cursadas durante o quinto ano de graduação do aluno, será excluída devido à grande quantidade de dados faltantes (35,58%). Para avaliar o desempenho do modelo, divide-se os dados em 70% para treinamento e 30% para teste. Um modelo de regressão logística é desenvolvido e treinado com 70% dos dados, sendo posteriormente avaliado nos 30% restantes usando a matriz de confusão e a acurácia, devido ao equilíbrio razoável entre evasão e graduação na base de dados.

Conforme afirmado por Monard e Baranauskas (2003), a matriz de confusão é uma ferramenta essencial para avaliar o desempenho de um modelo de classificação. Como demonstrado no Quadro 1, ela compara classificações corretas e previsões para categorias em um conjunto de exemplos, organizando os resultados em duas dimensões: categorias reais e previstas. Acertos estão na diagonal principal, e as células fora dela representam erros de classificação. Em um classificador perfeito, todos os elementos da diagonal secundária seriam zero, indicando a ausência de erros.

Quadro 1 - Exemplo de uma matriz de confusão

		Valor Previsto	
		Negativo	**Positivo**
Valor Real	**Negativo**	Verdadeiro Negativo (TN)	Falso Positivo (FP)
	Positivo	Falso Negativo (FN)	Verdadeiro Positivo (TP)

Fonte: Produção dos próprios autores.

Por meio da matriz de confusão, é possível extrair informações valiosas, como a acurácia, que pode ser definida como a porcentagem de predições corretas, expressa na Equação (1) (Luque *et al.*, 2019):

$$\frac{TP + TN}{TP + FN + TN + FP} \quad (1)$$

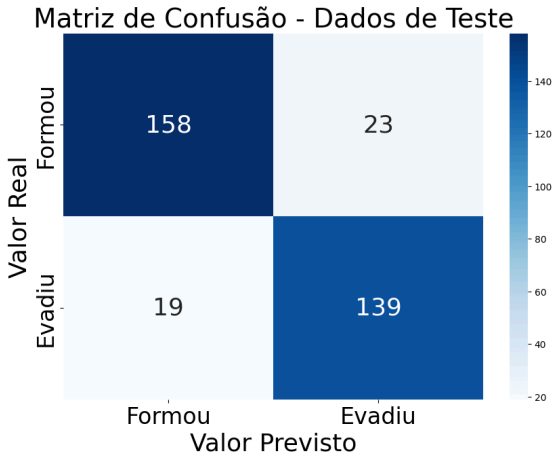
O propósito desta estimativa é reduzir a discrepância global. A acurácia nos dados de teste é a medida mais amplamente utilizada para avaliar e contrastar modelos de classificação (Castro; Braga, 2011). Assim, optou-se por empregar essa medida na comparação dos modelos. O modelo de árvore de decisão também foi avaliado e treinado nos mesmos conjuntos de dados (divididos em 70% para treinamento e 30% para teste), conforme realizado nos modelos de regressão logística. Antes de aplicar esse modelo nos testes, foi crucial otimizá-lo. O hiperparâmetro escolhido foi o critério de entropia, utilizado para avaliar o grau de desordem, e o tamanho máximo das árvores foi definido como 1, 2 ou 3.

Para determinar o modelo final, foram treinados três tipos de árvores distintas, com alturas máximas de 1, 2 e 3, respectivamente. Os resultados foram comparados utilizando a acurácia como métrica de avaliação. A árvore que alcançou a maior acurácia foi selecionada como o modelo final, sendo então treinada e posteriormente utilizada nos testes para prever a evasão.

5 RESULTADOS

A aplicação do modelo de regressão logística nos dados resultou na matriz de confusão apresentada na Figura 4. É possível observar que dezenove casos de evasão foram erroneamente classificados como não evasão pelo modelo, e houve vinte e três casos em que pessoas que realmente não evadiram foram erroneamente classificadas como evadidas. Esse modelo alcançou uma acurácia de 87,61%. A Tabela 2 relaciona as variáveis independentes com seus respectivos coeficientes.

Figura 4 – Matriz de confusão da regressão logística.



Fonte: Produção dos próprios autores.

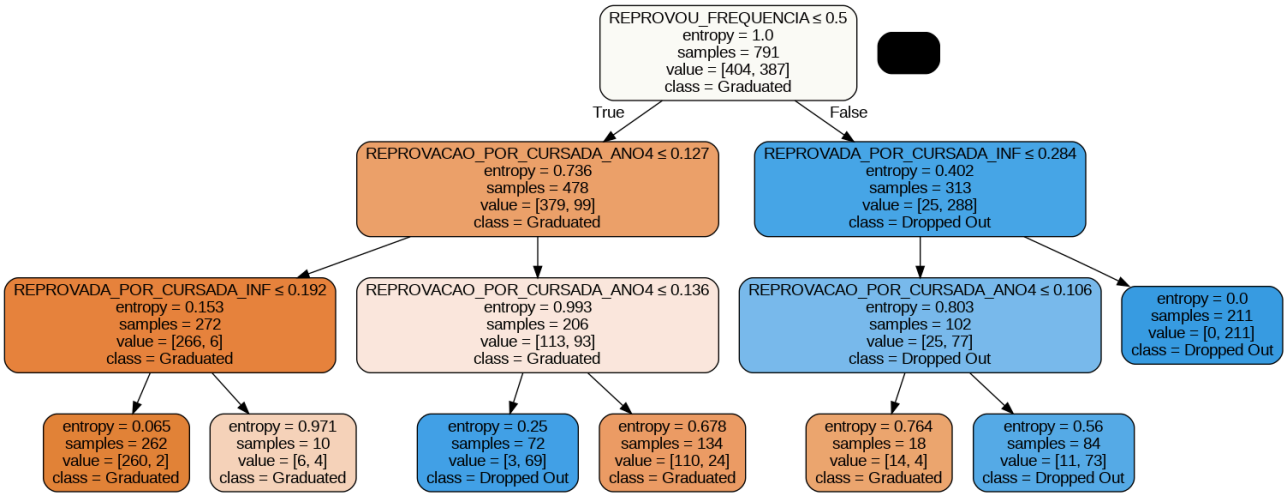
Tabela 2 - Coeficientes das variáveis utilizadas na regressão logística.

Variável	Coeficiente
Disciplinas reprovadas dividido por cursadas de Informática	1,3096
Reprovado por frequência	1,1084
Disciplinas reprovadas dividido por cursadas de Elétrica	0,6160
Disciplinas reprovadas dividido por cursadas de Matemática	0,5551
Forma de Ingresso	0,5195
Disciplinas reprovadas dividido por cursadas de Física	0,3662
Disciplinas reprovadas dividido por cursadas no segundo ano de graduação	0,3140
Disciplinas reprovadas dividido por cursadas no terceiro ano de graduação	0,1594
Disciplinas reprovadas dividido por cursadas no quarto ano de graduação	0,1260
Disciplinas reprovadas dividido por cursadas no primeiro ano de graduação	0,0737
Cotista	0,0050

Fonte: Produção dos próprios autores.

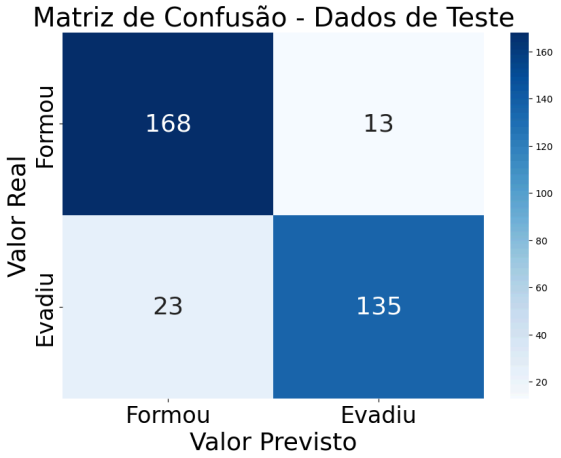
O modelo de árvore de decisão aplicado é mostrado na Figura 5. A aplicação deste modelo resultou na matriz de confusão apresentada na Figura 6, onde podemos observar que vinte e três casos de evasão foram erroneamente categorizados como não evasão, enquanto treze casos de não evasão foram incorretamente classificados como evadidos. Este modelo alcançou uma taxa de acurácia de 89,38%. Na Tabela 3, é destacada a importância de cada variável de entrada no processo de decisão do modelo. Dado que o modelo de árvore de decisão seleciona apenas as variáveis que melhor conseguem separar a base entre evadidos e não evadidos, algumas variáveis sequer foram incluídas.

Figura 5 – Árvore de decisão.



Fonte: Produção dos próprios autores.

Figura 6 – Matriz de confusão da árvore de decisão.



Fonte: Produção dos próprios autores.

Tabela 3 - Coeficientes das variáveis utilizadas na árvore de decisão.

Variável	Coeficiente
Reprovado por frequência	0,5269
Disciplinas reprovadas dividido por cursadas no quarto ano de graduação	0,3742
Disciplinas reprovadas dividido por cursadas de Informática	0,0987
Disciplinas reprovadas dividido por cursadas de Matemática	0,0000
Forma de Ingresso	0,0000
Disciplinas reprovadas dividido por cursadas de Física	0,0000
Disciplinas reprovadas dividido por cursadas no segundo ano de graduação	0,0000
Disciplinas reprovadas dividido por cursadas no terceiro ano de graduação	0,0000
Disciplinas reprovadas dividido por cursadas de Elétrica	0,0000
Disciplinas reprovadas dividido por cursadas no primeiro ano de graduação	0,0000
Cotista	0,0000

Fonte: Produção dos próprios autores.

Na Tabela 4, encontramos um resumo de todas as acurácias dos modelos para cada um dos cursos.

Tabela 4 - Acurácias dos modelos.

Regressão logística	Árvore de decisão
87,61%	**89,38%**

Fonte: Produção dos próprios autores.

Dentre os resultados dos modelos, destacam-se as variáveis de reprovação por disciplinas em informática e a variável que verifica se o aluno reprovou por frequência em pelo menos uma disciplina, as quais têm alta importância em todos os modelos. Isso indica que estudantes com alto índice de reprovação em disciplinas de informática e os que desistem de uma disciplina podem ter maior propensão a abandonar o curso.

Por outro lado, a variável que indica se o aluno é cotista demonstra baixa importância em todos os modelos, sugerindo que ser cotista não tem um impacto significativo na evasão do estudante. Essa constatação está em linha com pesquisas que mostram que, apesar das diferenças de desempenho no ENEM, os estudantes cotistas conseguem superar a defasagem de formação em pouco tempo (De Godoi; Dos Santos, 2021).

Entre os modelos analisados, a árvore de decisão se destaca como o mais eficaz, alcançando uma acurácia de 89,38%.

6 CONSIDERAÇÕES FINAIS

A pesquisa proporcionou uma compreensão aprofundada dos fatores que impactam a evasão nos cursos de Engenharia de Computação. A análise exploratória dos dados revelou padrões de comportamento entre estudantes evadidos e graduados, destacando a importância das variáveis de reprovação em disciplinas de informática e reprovação por frequência em ao menos uma matéria.

Os modelos de regressão logística e árvore de decisão foram eficazes na previsão de evasão, com taxas de acurácia variando entre 87,61% e 89,38%, ressaltando a relevância das variáveis selecionadas nos modelos.

No entanto, a evasão é um fenômeno complexo influenciado por diversos fatores individuais, socioeconômicos e institucionais, exigindo análises mais abrangentes que considerem outras variáveis relevantes. Estratégias de intervenção precoce, como programas de tutoria e atividades extracurriculares, podem contribuir para a redução da evasão.

Para pesquisas futuras, sugere-se explorar abordagens avançadas de aprendizado de máquina, considerar mais dados socioeconômicos dos alunos e realizar estudos mais personalizados para melhorar a precisão das previsões de evasão. Além disso, a avaliação do impacto das intervenções propostas com base nas previsões de evasão é um tema importante para investigações subsequentes.

REFERÊNCIAS

ALTURKI, Sarah; ALTURKI, Nazik. Using educational data mining to predict students' academic performance for applying early interventions. **Journal of Information Technology Education: JITE. Innovations in Practice: IIP**, v. 20, p. 121-137, 2021.

BERNARDO, Ana *et al.* Freshmen program withdrawal: Types and recommendations. **Frontiers in psychology**, v. 8, p. 274165, 2017.

BRASSCOM. **Estudo da Brasscom aponta demanda de 797 mil profissionais de tecnologia até 2025**. 2021. Disponível em: https://brasscom.org.br/estudo-da-brasscom-aponta-demanda-de-797-mil-profissionais-de-tecnologia-ate-2025/. Acesso em: 10 maio 2024.

CAMPELLO, Antonio de Vasconcellos Carneiro; LINS, Luciano Nadler. Metodologia de análise e tratamento da evasão e retenção em cursos de graduação de instituições federais de ensino superior. **Encontro Nacional de Engenharia de Produção**, v. 28, n. 2008, p. 1-13, 2008.

CASTRO, Cristiano Leite de; BRAGA, Antônio Pádua. Supervised learning with imbalanced data sets: an overview. **Sba: Controle & Automação Sociedade Brasileira de Automatica**, v. 22, p. 441-466, 2011.

DA FONSECA, J. M. M. R. **Indução de Árvores de decisão**. 1994. Tese de Doutorado. Dissertação de Mestrado, Universidade Nova de Lisboa, Lisboa.

DE ALMEIDA, Eustáquio; GODOY, Elenilton Vieira. A evasão nos cursos de engenharia: uma análise a partir do COBENGE. 2016.

DE GODOI, Marciano Seabra; DOS SANTOS, Maria Angélica. Dez anos da lei federal das cotas universitárias: avaliação de seus efeitos e propostas para sua renovação e aperfeiçoamento. **Revista de Informação Legislativa**, v. 58, n. 229, p. 11-35, 2021.

GARCIA, Léo Manoel Lopes da Silva; GOMES, Raquel Salcedo. Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica. **Revista Educar Mais**, Pelotas, RS, v. 6, p. 937-957, 2022.

GONZALEZ, Leandro de Azevedo. Regressão logística e suas aplicações. 2018.

HOED, Raphael Magalhães. Análise da evasão em cursos superiores: o caso da evasão em cursos superiores da área de Computação. 2017.

INEP, Diretoria de Estatísticas Educacionais. **Censo da Educação Superior 2021: Divulgação dos resultados**. 2022. Disponível em: https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2021/apresentacao_censo_da_educacao_superior_2021.pdf. Acesso em: 10 maio 2024.

LUQUE, Amalia *et al.* The impact of class imbalance in classification performance metrics based on the binary confusion matrix. **Pattern Recognition**, v. 91, p. 216-231, 2019.

MARQUES, Leonardo Torres *et al.* Evasão acadêmica e suas causas em cursos de bacharelado em ciência da computação: Um estudo de caso na ufersa. In: **Anais do XXXI simpósio brasileiro de informática na educação**. SBC, 2020. p. 1042-1051.

MEC, Ministério da Educação. **Metodologia de Cálculo dos Indicadores de Fluxo da Educação Superior**. Brasília, 2017. Disponível em: https://download.inep.gov.br/informacoes_estatisticas/indicadores_educacionais/2017/metodologia_indicadores_trajetoria_curso.pdf. Acesso em: 16 maio 2024.

MONARD, M.; BARANAUSKAS, J. Conceitos sobre aprendizado de máquinas em Sistemas Inteligentes: Fundamentos e Aplicações. 2003.

MOREIRA DA SILVA, Diogo E. *et al.* Forecasting Students Dropout: A UTAD University Study. **Future Internet**, v. 14, n. 3, p. 76, 2022.

MORETTIN, P. A.; BUSSAB, W. O. Estatística básica: Saraiva Educação SA. 2017.

NUNES, Francinaldo Carlos *et al.* Estudo exploratório sobre a evasão no curso de Computação da UFCG: um olhar sobre a disciplina cálculo I. 2020.

OPAZO, Diego *et al.* Analysis of first-year university student dropout through machine learning models: A comparison between universities. **Mathematics**, v. 9, n. 20, p. 2599, 2021.

RASTROLLO-GUERRERO, Juan L.; GÓMEZ-PULIDO, Juan A.; DURÁN-DOMÍNGUEZ, Arturo. Analyzing and predicting students' performance by means of machine learning: A review. **Applied sciences**, v. 10, n. 3, p. 1042, 2020.

ROLIM, Joana Pacheco; SILVA, Rômulo César. Mineração de Dados Educacionais para identificação de Perfil de Retenção em um curso de Ciência da Computação. In: **Anais da IX Escola Regional de Informática de Goiás**. SBC, 2021. p. 195-204.

SACCARO, Alice; FRANÇA, Marco Túlio Aniceto; JACINTO, Paulo de Andrade. Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. **Estudos Econômicos (São Paulo)**, v. 49, p. 337-373, 2019.

SILVA, Francisca Islandia Cardoso da *et al.* Evasão escolar no curso de educação física da Universidade Federal do Piauí. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 17, p. 391-404, 2012.

MODELING AND PREDICTION OF STUDENT DROPOUT IN COMPUTER ENGINEERING COURSES USING MACHINE LEARNING

Abstract: *High dropout rates in tech-focused higher education programs contribute significantly to the IT professional shortage. To address this, the university's Department of Computer Engineering conducted a survey aiming for early and precise performance diagnosis among students. This allows for timely interventions to enhance completion rates. Machine learning models were employed to predict student performance and dropout likelihood. Logistic regression and decision tree models were trained and evaluated, yielding promising results with over 80% accuracy in predicting dropout.*

Keywords: *Prediction of dropout, Higher education dropout, Machine learning models.*