



Análise da Evasão em Cursos de Computação e Tecnologia no Brasil com o Apoio de Modelos de Regressão e *Explainable AI*

Raul da Silva Martins, Universidade Federal do Vale do São Francisco,
raul.silvamartins@discente.univasf.edu.br, <https://orcid.org/0009-0000-8780-7226>.

Débora da Conceição Araújo, Universidade Federal do Vale do São Francisco,
debora.caraujo@univasf.edu.br, <https://orcid.org/0000-0002-1226-6012>.

Iago Bruno Ferreira e Souza, Instituto Federal do Sertão Pernambucano,
iagoferreirabs@gmail.com, <https://orcid.org/0009-0001-3277-3281>.

Felipe Guilherme Oliveira Melo, Universidade Federal do Vale do São Francisco,
felipe.guilherme@univasf.edu.br, <https://orcid.org/0000-0002-5339-5718>.

Resumo: A evasão em cursos de computação e tecnologia representa um desafio significativo para as Instituições de Ensino Superior (IES) no Brasil. Considerando esse contexto, este trabalho teve por objetivo estimar a evasão e identificar as características mais influentes para tal abandono. Para as análises, utilizou-se a base de dados do Censo da Educação Superior em modelos de regressão avaliados quanto à sua capacidade preditiva. O modelo *Random Forest* apresentou o melhor desempenho e, a partir dele, aplicaram-se técnicas de explicabilidade a fim de compreender a contribuição de cada variável da base nas previsões. Os resultados evidenciaram padrões distintos entre IES públicas e privadas, com destaque para fatores como faixa etária e ocupação de vagas remanescentes. Esses achados apontam para a importância de políticas institucionais direcionadas a públicos específicos, contribuindo para estratégias mais eficazes de permanência e conclusão dos cursos. Espera-se que as discussões expostas nesta pesquisa contribuam para o aprimoramento de políticas educacionais eficazes para a redução da evasão nos cursos de computação e tecnologia no Brasil.

Palavras-chave: Previsão de evasão, Cursos de Computação, Inteligência Artificial Explicável.

Dropout Analysis in Computing and Technology Programs in Brazil Using Regression Models and Explainable AI

Abstract: Student dropout in computer science and technology programs poses a significant challenge for Higher Education Institutions (HEIs) in Brazil. This study aims to estimate dropout rates and identify the most influential factors contributing to student attrition in these fields. Using data from the Brazilian Higher Education Census, several regression models were tested for predictive performance. The Random Forest model demonstrated the best results and was subsequently interpreted through explainable AI techniques to assess the contribution of each variable to the predictions. The findings revealed distinct patterns between public and private HEIs, with key factors including admission type, course shift (day or night), and enrollment in remaining vacancies. These insights are expected to support the development of more effective educational policies aimed at reducing dropout rates.

Keywords: Dropout Prediction, Computer Science Programs, Explainable Artificial Intelligence.

1. Introdução

A evasão escolar representa um desafio crítico enfrentado pelas instituições dos mais diversos níveis de ensino. Este fenômeno gera impactos significativos nos âmbitos social,



acadêmico e econômico, comprometendo o desenvolvimento educacional e profissional dos indivíduos (Lopes; Lima e Fagundes, 2023; Lima e Fagundes, 2023). No Brasil, os índices de evasão são particularmente preocupantes, sendo observadas taxas elevadas tanto no ensino técnico quanto no superior, com variações conforme o perfil do estudante, a modalidade de ensino (Sousa *et al.*, 2023) ou a área do curso.

Tem-se, nesse sentido, os cursos da área de computação frequentemente identificados entre os mais altos índices de evasão (Jesus; Rodriguez e Junior, 2021; Alvim; Bittencourt e Duran, 2024). No entanto, são escassos os estudos que investigam tal abandono no contexto brasileiro (Alvim; Bittencourt e Duran, 2024). Em Oliveira e Medeiros (2024) é apresentado um modelo de previsão de evasão em nível de Universidade e seus variados cursos. Os modelos construídos em Oliveira e Medeiros (2024) basearam-se, principalmente, em informações oriundas de um questionário aplicado aos estudantes, porém os resultados são gerais, ou seja, as características individuais de cada curso não foram analisadas. A evasão em cursos de computação, especificamente, é investigada em Silva *et al.* (2021) em que um questionário é aplicado com o intuito de identificar os perfis dos estudantes atuais e desistentes, além de suas expectativas e dificuldades encontradas no curso. Modelos com o intuito de prever a evasão a partir dos dados não foram considerados no estudo. Em Jesus, Rodriguez e Junior (2021), a tarefa de previsão de evasão é considerada no curso de Licenciatura em Computação de uma Universidade. Atributos como a frequência dos estudantes, as disciplinas que estavam cursando no momento do questionário e suas respectivas notas são utilizados como entradas para o modelo. A rede neural *Multilayer Perceptron* (MLP) apresentou o melhor resultado no estudo (Jesus; Rodriguez e Junior, 2021) e, como trabalhos futuros, é possível apontar a utilização de métodos explicáveis que permitam aos autores entender as características que impactam na evasão identificada.

Diante da complexidade do problema, diversas estratégias são exploradas com a finalidade de compreender o fenômeno de evasão escolar nos cursos de computação e tecnologia. Entre essas estratégias, é possível observar o uso de técnicas de *Machine Learning* (ML) que se destacam como abordagens promissoras para problemas de predição, permitindo a identificação de padrões a partir de grandes volumes de dados (Nascimento *et al.*, 2024). Modelos preditivos baseados em ML oferecem a possibilidade de lidar com a estimação da evasão com base em diferentes fatores (Oliveira e Medeiros, 2024). Para isso, técnicas supervisionadas e não supervisionadas, incluindo métodos baseados em árvores, Máquinas de Vetor de Suporte e Redes Neurais Artificiais, têm sido aplicadas com sucesso (Costa; Júnior e Santana, 2023). Além de prever as taxas de evasão, é de fundamental importância, no contexto educacional, que os modelos também sejam capazes de identificar os fatores que acarretam tal fenômeno, permitindo intervenções mais assertivas e informadas. Destaca-se, portanto, a Inteligência Artificial Explicável (*eXplainable Artificial Intelligence* – XAI) que tem ganhado destaque por sua capacidade de tornar os modelos de previsão mais transparentes e compreensíveis (Mustofa *et al.*, 2025; Khosravi *et al.*, 2022).

À luz dessas informações, esta pesquisa fez uso dos microdados do Censo da Educação Superior (CES), divulgados pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (Inep) (Inep, 2025), com o objetivo de investigar a evasão nos cursos de computação e tecnologia no Brasil por meio da aplicação de modelos de regressão e de XAI. Ao final, os principais padrões identificados foram analisados com a finalidade de contribuir para a formulação de estratégias que auxiliem na mitigação da evasão e na melhoria das políticas educacionais voltadas para a permanência estudantil.



2. Materiais e Métodos

Esta seção descreve a base de dados utilizada para o estudo, as etapas de pré-processamento realizadas, bem como os modelos e técnicas de aprendizado de máquina construídos e seus métodos de avaliação.

2.1. Base de Dados

A base de dados utilizada para treinamento e teste dos modelos faz parte do Censo da Educação Superior de 2023*, que abrange os dados mais recentes divulgados pelo Inep até a realização da presente pesquisa. O CES coleta informações sobre as instituições de ensino superior, públicas e privadas, assim como sobre seus alunos e docentes, com o objetivo de fornecer estatísticas confiáveis, além de fundamentar políticas públicas e auxiliar a tomada de decisões dos gestores das instituições.

Os microdados do CES estão divididos em duas partes principais, a saber: 1) Cadastro de Instituições de Ensino Superior (IES); e 2) Cadastro de Cursos. A primeira base possui dados gerais sobre todas as IES do país, enquanto a segunda apresenta dados detalhados de cada curso que participou do censo. Essa pesquisa considerou a granularidade de cursos, por ter como objetivo investigar a evasão em cursos da área de computação e tecnologia. A base de Cadastro de Cursos possui mais de 600 mil cursos, cada um com 200 variáveis associadas.

Além dos microdados do CES, foi utilizada neste trabalho a base de dados de Indicadores de Fluxo da Educação Superior†. Os indicadores de fluxo permitem a avaliação dos cursos através de uma trajetória temporal com três dimensões: permanência, conclusão e desistência no curso de ingresso. A Taxa de Desistência Anual (TADA), presente nos indicadores de fluxo, foi utilizada como variável resposta para os modelos de previsão, pois esta representa o percentual de alunos desvinculados ou transferidos de um determinado curso.

2.2. Pré-processamento de dados

A etapa de pré-processamento envolveu a realização de mudanças na base de dados para torná-la adequada para o treinamento dos modelos de regressão (Kuhn *et al.*, 2013). Nesta etapa, após a remoção dos valores nulos, todas as variáveis quantitativas dos microdados foram selecionadas, com exceção da QT_SIT_DESVINCULADO e QT_SIT_TRANSFERIDO, tendo em vista que ambas fazem parte da fórmula adotada pelo Inep para a variável de saída (TADA). A respeito das variáveis categóricas, foram selecionadas TP_GRAU_ACADEMICO e IN_GRATUITO, optou-se por remover as demais variáveis categóricas, pois essas envolvem características individuais, como o nome da instituição de ensino (NO_IES). Essa medida visa evitar que os modelos aprendam padrões excessivamente específicos de determinadas IES, comprometendo a capacidade de generalização dos preditores. Ao final, 172 variáveis de entrada e uma variável resposta restaram na base de dados pré-processada‡.

Após o pré-processamento, a coluna CO_CINE_AREA_GERAL foi utilizada para filtrar cursos classificados como Computação e Tecnologias da Informação e Comunicação (TIC), correspondentes ao código 6 do *dataset*. Ao final, a base de dados é salva em disco no formato CSV, contendo 1442 cursos das áreas de computação e tecnologia. Para a realização dos experimentos, os dados foram divididos em dois

*Microdados do Inep: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/microdados/>

†Indicadores de Fluxo da Educação Superior: <https://www.gov.br/inep/pt-br/aceso-a-informacao/dados-abertos/indicadores-educacionais/indicadores-de-fluxo-da-educacao-superior>

‡Variáveis selecionadas: <https://drive.proton.me/urls/BCWAZH7PWW#RPgDAw1ew97l>



subconjuntos, a saber: 1) IES privadas (910 cursos); e 2) IES públicas (532 cursos). Para isso, foi utilizada a coluna TP_REDE que indica se o curso faz parte da rede pública ou privada de ensino. A divisão da base, considerando o tipo de instituição, foi realizada após testes piloto que indicaram que os modelos de regressão se comportavam melhor recebendo as entradas de IES públicas e privadas separadamente. A base de dados final, após todo o processamento assim como o código-fonte utilizado na pesquisa estão disponíveis no repositório do Github [§].

2.3. Protocolo dos Experimentos

O problema de previsão de evasão é abordado na literatura tanto como uma tarefa de classificação quanto de regressão (Nascimento *et al.*, 2024). A escolha dos modelos depende da análise da base de dados e, principalmente, da variável-alvo. No presente estudo, deseja-se estimar o valor de uma variável contínua (TADA), portanto, um problema de regressão. Os principais modelos de regressão apresentados na literatura para a tarefa de evasão (Nascimento *et al.*, 2024) foram selecionados para análise neste trabalho, a saber: *Random Forest* (RF), *Decision Tree* (DT), *Multilayer Perceptron* (MLP) e *Support Vector Machine* (SVM).

Para o treinamento e teste dos modelos supracitados, foi utilizada a biblioteca *Scikit-learn*,[¶] ferramenta que provê implementações otimizadas dos modelos e permite que os pesquisadores possam customizar seus hiperparâmetros (HP) de forma adequada ao problema. A etapa de definição dos HP é fundamental para que o modelo de ML apresente um bom resultado, pois os hiperparâmetros especificam a estrutura e o comportamento do preditor durante o treinamento (Yu e Zhu, 2020). Para otimização e definição dos HP, foi utilizado o algoritmo *Grid Search* (GS), considerado uma técnica clássica de otimização. O GS realiza uma busca exaustiva entre o conjunto de possíveis valores para cada combinação de HPs (Tran *et al.*, 2020). Nesse sentido, os valores considerados no GS para esta pesquisa encontram-se na tabela disponível no link [‡].

Após finalizar a execução do GS, é selecionado o conjunto de HP que obteve o melhor resultado para cada modelo. Para a análise dos modelos de regressão, foi utilizado o algoritmo *K-folds Cross Validation* (CV), que divide a base de dados em *k* blocos de tamanhos iguais e cada bloco é selecionado uma vez para ser o conjunto de teste, enquanto os demais são utilizados para treinamento. O processo se repete até que todos os *folds* tenham sido testados. A quantidade de *folds* comumente adotada varia entre 5 e 10, pois oferece estimativas de erro que não sofrem um viés excessivo (subestimação do erro) nem de variância muito alta (resultados instáveis), tornando tais valores escolhas seguras para a maioria dos casos (Nti *et al.*, 2021). Devido ao tamanho da base de dados deste estudo, consideramos o valor de *k*=5 para preservar um mínimo de 20% de exemplos em cada *fold*, de modo a não deixar o conjunto de teste pouco representativo. Após testes preliminares com a base de dados completa, o GS e o CV foram executados considerando as duas divisões da base de dados (IES públicas e IES privadas).

Para avaliar a assertividade das previsões realizadas, foram consideradas as métricas *Mean Absolute Error* (MAE), *Mean Squared Error* (MSE) e *Root Mean Squared Error* (RMSE), comumente utilizadas na literatura ao abordar modelos de regressão (Nascimento *et al.*, 2024). Tais estimadores também são conhecidas como métricas de erros, pois são baseadas no erro (viés) que um modelo apresenta ao comparar a sua predição com o valor real. A métrica MAE define a magnitude média do erro apresentado

[§]Repositório com código-fonte e base de dados: (<https://github.com/raulpy271/analise-evasao-escolar>)

[¶]Scikit-learn: (<https://scikit-learn.org/stable/>)

[‡]Hiperparâmetros: (<https://drive.proton.me/urls/4HAEHENAX4#CWeyUWxgEteM>)



pelo modelo, desconsiderando a sua direção. A MSE, por sua vez, é a média do erro quadrado apresentado e ressalta a presença de *outliers* nos dados. Por fim, a métrica RMSE consiste na extração da raiz quadrada do MSE, o que faz com que o valor da métrica possua a mesma escala da variável resposta.

A definição matemática das métricas encontra-se nas fórmulas agrupadas na Equação 1, em que a variável p_i representa o valor predito pelo modelo para a entrada i , r_i representa o valor da resposta correta para a mesma entrada e N representa a quantidade de entradas em que o modelo foi testado.

$$MAE = \frac{1}{N} \sum_{i=1}^N |p_i - r_i| \quad MSE = \frac{1}{N} \sum_{i=1}^N (p_i - r_i)^2 \quad RMSE = \sqrt{MSE} \quad (1)$$

Após os testes com os modelos de regressão, o melhor preditor para cada divisão da base de dados foi selecionado para a etapa de análise explicável. Nessa etapa, foi empregada a técnica *SHapley Additive exPlanations* (SHAP), proposta por Lundberg e Lee (2017). O SHAP se baseia na teoria dos valores de Shapley, oriunda da Teoria dos Jogos, e permite quantificar a contribuição de cada variável de entrada para a previsão realizada pelo modelo, oferecendo uma interpretação local e global dos resultados. Essa abordagem foi escolhida por sua robustez e ampla adoção em estudos que requerem maior transparência e interpretabilidade dos modelos.

3. Resultados e Discussões

Nesta seção são apresentados os resultados do treinamento e teste dos modelos considerando as instituições públicas e privadas, o que envolve avaliar a performance dos preditores, o tempo de treinamento e a importância das variáveis de entrada para as previsões. A estabilidade dos modelos será também avaliada através de ferramenta visual, com gráficos do tipo *boxplot* e as principais características extraídas por meio de XAI serão discutidas.

3.1. Análise de Evasão em Instituições Privadas

A Tabela 1 apresenta valores médios do tempo de treinamento e do desempenho dos modelos aplicados à base de dados composta por cursos de IES da rede privada. A média aritmética e o desvio padrão do tempo foram calculados a partir dos 5 valores obtidos ao realizar o método CV. A análise do tempo de treinamento dos modelos indica que todos os modelos executaram rapidamente, sem exceder mais que um segundo. Apesar disso, foi observado um tempo maior para os algoritmos RF e MLP, essa diferença pode ser prejudicial em uma base maior.

Tabela 1. Comparação da performance dos modelos para base de dados com cursos da rede privada.

Modelo	Tempo (s)	MAE	MSE	RMSE
DT	0.00 / 0.000	0.062 / 0.007	0.009 / 0.004	0.091 / 0.021
RF	0.65 / 0.009	0.061 / 0.007	0.008 / 0.004	0.090 / 0.021
MLP	0.44 / 0.100	0.062 / 0.006	0.009 / 0.004	0.091 / 0.020
SVM	0.02 / 0.003	0.078 / 0.007	0.010 / 0.003	0.098 / 0.015

Ao analisar as métricas de avaliação, é possível identificar que os algoritmos apresentaram uma performance semelhante. Para verificar se essa similaridade é estatisticamente significativa, poderiam ter sido realizados testes estatísticos, como o

t-Student ou testes não paramétricos, que avaliam se há diferenças significativas entre os resultados obtidos. No entanto, devido ao tamanho reduzido dos *folds* ($k=5$), foram realizadas apenas as análises visuais que exploram a distribuição dos resultados e identificam possíveis discrepâncias, conforme apresentado na Figura 1.

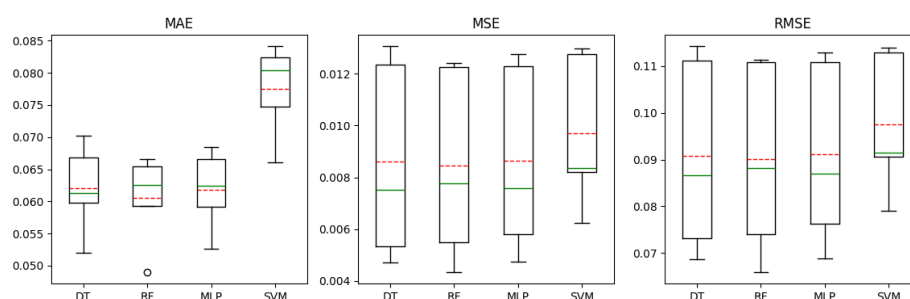


Figura 1. *Boxplot* com performance dos modelos para base de dados com cursos da rede privada.

A análise dos limites inferior e superior de cada caixa no *boxplot* indica que SVM e RF apresentaram, em geral, menor variação. Essa tendência foi mais evidente na métrica MAE, no caso do RF, e nas métricas MSE e RMSE, no caso do SVM. No entanto, o *Random Forest* foi o modelo que apresentou o melhor desempenho em todas as métricas, com erro médio de 6,1% na MAE, 0,8% na MSE e 9% na RMSE. Por esse motivo, o RF foi selecionado para a etapa de XAI.

Para entender o impacto das variáveis em um problema há métricas que definem a importância de cada variável (*feature importance*) de entrada no resultado de um preditor (Scornet, 2023). A escolha da métrica para estimar a importância de uma variável depende de quais modelos de ML o estudo deseja utilizar, pois algumas métricas assumem o uso de métodos específicos. Dado que o RF apresentou performances favoráveis, foi escolhida a métrica de *feature importance* utilizada no RF, *Mean Decrease Impurity* (MDI), de modo a garantir uma interpretação nativa da forma como o modelo tomou decisões ao longo do treinamento (Scornet, 2023).

A MDI é representada por um valor numérico atribuído a cada variável de entrada. Quando a MDI de uma variável tem valor 0, tem-se um indicativo de que não há importância no resultado do modelo, ou seja, quanto mais alto é o valor do MDI, maior é a importância da variável em questão, o cálculo da métrica pode ser encontrado em Louppe (2015). Assim, foi computado o MDI das 172 variáveis a partir do modelo RF treinado com a base de dados com as IES privadas. Dessas, 86 variáveis exerceram alguma importância na resposta do modelo, porém a soma do MDI das 15 variáveis mais importantes é de aproximadamente 0.67, ou seja, essas 15 variáveis exerceram um papel crucial na resposta do modelo, frente às demais 71, tendo em vista que os resultados do MDI foram normalizados no intervalo real (0,1).

A métrica MDI é útil para identificar as variáveis mais importantes no resultado do modelo, porém a mesma não fornece informações de direção, ou seja, não se sabe se a variável contribuiu para o aumento ou para a diminuição da evasão. Para se obter essa informação sobre a direção, foi utilizada a técnica SHAP, descrita na seção 2.3. O resultado da ferramenta é expresso de forma visual na Figura 2, nela é apresentada a métrica de valor SHAP para as 15 variáveis mais importantes. Para cada variável, é possível visualizar os seus diferentes valores e como eles impactam no resultado do modelo. Para melhor interpretação da Figura 2 vale ressaltar que cada linha representa uma variável do conjunto de dados, enquanto o eixo X mostra o impacto de cada variável



na predição do modelo. Sendo assim, é possível entender que valores SHAP positivos aumentam a chance de evasão, enquanto valores negativos diminuem. Sobre as cores, referem-se ao valor real da variável, sendo o vermelho relacionado a valores altos e o azul relacionado a valores baixos.

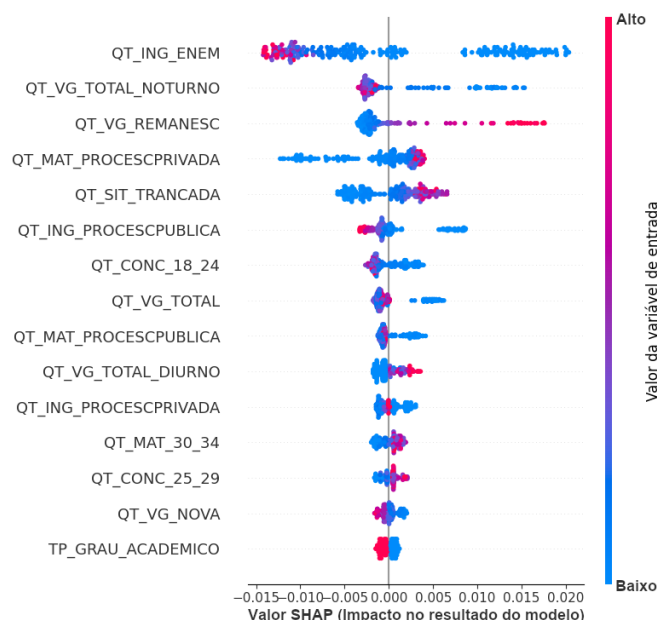


Figura 2. Valor SHAP das 15 variáveis mais importantes na base de cursos da rede privada treinado com modelo RF.

De acordo com o exposto na Figura 2, observa-se que os estudantes que ingressam via Exame Nacional do Ensino Médio (ENEM) tendem a apresentar menor evasão nas instituições privadas, em comparação àqueles que entram por processos seletivos próprios. Esse resultado pode indicar que o ingresso via ENEM está mais associado à escolha consciente do curso ou da instituição, tendo em vista que o ENEM democratizou o acesso ao ensino superior, permitindo que estudantes de diferentes regiões e contextos socioeconômicos concorram a vagas em todo o país (Barbosa e Pires, 2022). Enquanto os estudantes que optam por processos seletivos próprios, possivelmente, priorizam a proximidade geográfica ou a conveniência e não necessariamente o alinhamento com a área escolhida. Assim, a maior evasão observada nesse grupo pode estar relacionada a uma menor identificação inicial com o curso. Outro achado relevante diz respeito à relação entre o período noturno e o perfil etário dos estudantes. Cursos noturnos, possivelmente frequentados por alunos que trabalham durante o dia (Vasconcelos *et al.*, 2020), apresentam maior risco de evasão. Isso pode estar ligado à predominância de estudantes com mais de 25 anos nesse turno, que frequentemente acumulam responsabilidades profissionais e familiares. O modelo identificou que a faixa etária de 18 a 24 anos está mais associada à conclusão do curso, enquanto faixas mais avançadas indicam maior propensão à evasão. Esses resultados destacam a necessidade de políticas de apoio específicas para estudantes com idade superior a 25 anos, sobretudo nos cursos noturnos.

Por fim, um fator relevante está associado à variável que indica trancamento de matrícula, que aponta que estudantes que interrompem temporariamente seus estudos tendem a não retornar. De forma semelhante, o aumento no número de vagas remanescentes apresenta correlação com maiores índices de evasão, configurando-se como um indicador de evasão estrutural, frequentemente associado à baixa atratividade



do curso e a fragilidades nos mecanismos institucionais de permanência. Em cursos de computação e tecnologia, a baixa ocupação pode estar relacionada à percepção de que o curso não oferece condições adequadas de ensino, como laboratórios bem equipados ou acesso a tecnologias atualizadas.

3.2. Análise de Evasão em Instituições Públicas

Nesta seção é detalhada a previsão de evasão e análise dos resultados relativos à base de dados que considera apenas as IES públicas. A Tabela 2 apresenta a performance dos modelos. Sobre o tempo de treinamento, mais uma vez é possível observar o treinamento rápido para todos os modelos, além de que o tempo absoluto do RF e MLP foi reduzido em relação à base com instituições privadas, fato que poderia ser previsto ao comparar o tamanho das bases.

Tabela 2. Comparação da performance dos modelos para base de dados com cursos da rede pública.

Modelo	Tempo (s)	MAE	MSE	RMSE
DT	0.01 / 0.002	0.068 / 0.015	0.010 / 0.006	0.098 / 0.031
RF	0.46 / 0.007	0.068 / 0.015	0.010 / 0.006	0.097 / 0.030
MLP	0.34 / 0.042	0.071 / 0.014	0.011 / 0.006	0.100 / 0.030
SVM	0.01 / 0.001	0.072 / 0.015	0.010 / 0.005	0.097 / 0.028

A métrica MAE mostrou uma semelhança ainda maior na performance dos modelos, eles emitiram um erro absoluto entre 6.8 a 7.2 pontos percentuais, sendo o MLP e SVM com o maior erro absoluto e DT e RF com os menores erros. Quanto às métricas MSE e RMSE, assumem valores quase idênticos aos valores das mesmas métricas para os modelos treinados com a base de IES privadas. As médias e desvios padrão alcançados podem ser analisados de forma visual no gráfico 3, que apresenta um *boxplot* com a performance dos modelos.

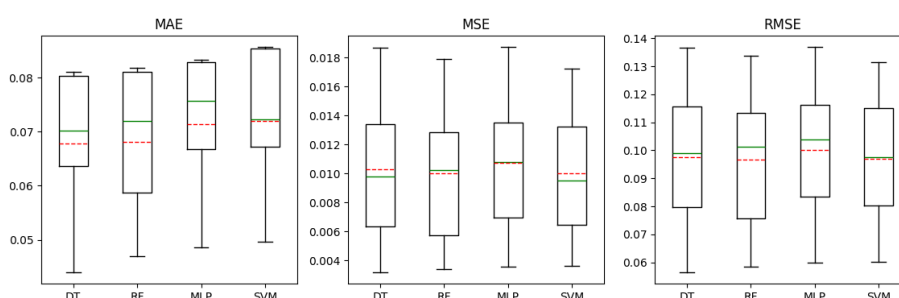


Figura 3. *Boxplot* com performance dos modelos para base de dados com cursos da rede pública.

No *boxplot* é possível observar que as barras têm escala maior em relação aos resultados com IES privadas, o que se justifica pelo fato de a base de IES públicas ser a de menor tamanho. Pode ser vista uma leve diferença entre as estatísticas média e mediana representadas pelas linhas de cores vermelha e verde, respectivamente. Ao analisarmos o desempenho médio dos modelos, identificou-se, mais uma vez, que o RF está presente entre os melhores modelos nas 3 métricas, portanto, foi escolhido para a análise explicável.

Para analisar a importância de cada variável de entrada, novamente foi adotada a métrica MDI, pois o modelo RF manteve-se com uma performance adequada. Para os cursos de computação das IES públicas, tem-se um total de 97 variáveis atingindo um valor de MDI maior que 0, ou seja, influenciando a previsão do modelo. A importância



das variáveis aparece mais dispersa para as IES públicas, e a soma do MDI das 15 variáveis mais importantes atingiu um total de 0.50. Ao compararmos as 15 variáveis mais importantes para as IES públicas e privadas, apenas 5 constam nas duas listas de características, o que indica que questões distintas influenciam a evasão em diferentes categorias administrativas, além de corroborar com os testes iniciais que mostraram desempenho melhor dos preditores ao separarmos as bases. A Figura 4 mostra o resultado da técnica SHAP aplicada ao modelo RF treinado com a base de instituições públicas.

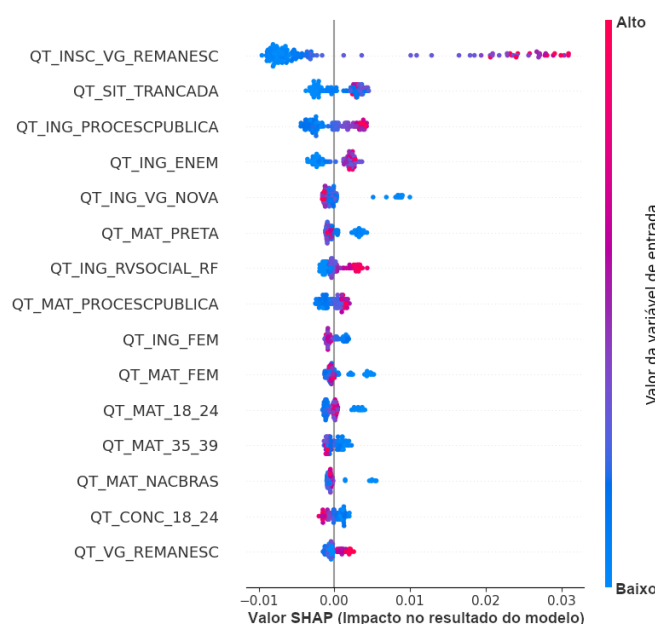


Figura 4. Valor SHAP das 15 variáveis mais importantes na base de cursos da rede pública treinado com modelo RF.

A análise da Figura 4 revela que, nas instituições públicas, o preenchimento de vagas remanescentes está fortemente associado à evasão, o que pode indicar que o ingresso tardio do estudante é o principal fator de impacto para a evasão em cursos de computação e tecnologia em IES públicas. Esse fato pode estar relacionado à natureza das disciplinas introdutórias, que frequentemente envolvem conteúdos da área de exatas, como Cálculo, Programação e Lógica, exigindo uma base sólida em matemática. Estudantes que ingressam por meio de vagas remanescentes podem enfrentar maior dificuldade para acompanhar o conteúdo já em andamento, o que compromete o aprendizado e pode aumentar a probabilidade de evasão. Portanto, acelerar o processo de ingresso de alunos remanejados pode ser um ponto importante para mitigar a evasão nas IES públicas. Agora, como um fator importante para a permanência, tem-se as políticas de ação afirmativa, como a reserva de vagas para grupos sociais específicos, que demonstraram relação com a redução da evasão, o que sugere a efetividade desses mecanismos tanto na seleção quanto na permanência dos estudantes no ensino superior. Assim como nas instituições privadas, a faixa etária também se mostra relevante: estudantes mais jovens (18 a 24 anos) apresentam maior probabilidade de conclusão, reforçando a importância de estratégias específicas de apoio para estudantes a partir dos 25 anos de idade, que podem enfrentar maiores desafios externos à vida acadêmica.

3.3. Implicações práticas e propostas de intervenção

Os achados desta pesquisa podem ser operacionalizados por gestores educacionais por meio da formulação de estratégias de retenção específicas para os perfis de risco



identificados.

Nesse contexto, se mostram importantes políticas institucionais voltadas ao acompanhamento dos estudantes de duas características principais: 1) maiores de 24 anos e 2) ingressantes da lista de remanejamento. Ações como agilidade na chamada de estudantes da lista de remanejamento, tutoria acadêmica intensiva nos primeiros semestres e programas de reintegração estudantil podem ser de fundamental importância para a permanência desses grupos.

Além disso, os resultados reforçam a importância de políticas públicas como o ENEM e as ações afirmativas, que mostraram relação com maior permanência estudantil. Isso indica que programas como o Sistema de Seleção Unificada (SISU), o Programa Universidade para Todos (Prouni) e as cotas sociais devem ser não apenas mantidos, mas ampliados e monitorados com foco também na permanência, e não apenas no acesso.

4. Considerações finais

Este estudo teve como objetivo prever a evasão em cursos de computação e tecnologia no Brasil por meio da aplicação de modelos de aprendizado de máquina, bem como identificar as características mais relevantes para o fenômeno do abandono estudantil, por meio de técnicas de *eXplainable AI*. A solução proposta baseou-se na utilização de modelos de regressão treinados com dados do Censo da Educação Superior do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira.

Dentre os principais achados, destaca-se que estudantes que ingressam por meio de processos seletivos públicos demonstram maior propensão à permanência, mesmo em instituições privadas, enquanto o ingresso a partir de vagas remanescentes aparece fortemente associado à evasão nas IES públicas. Entende-se que o ingresso tardio dificulta a integração do estudante, uma vez que ele pode ter perdido aulas iniciais importantes em disciplinas com altas taxas de retenção, como Cálculo e Programação. A análise por faixa etária indicou que estudantes entre 18 e 24 anos, de IES públicas e privadas, apresentam menor risco de evasão, o que evidencia a necessidade de políticas de apoio voltadas a estudantes acima dessa faixa etária, especialmente em cursos noturnos. Além disso, a presença de políticas afirmativas revelou-se positiva na permanência de alunos nas instituições públicas.

Como trabalhos futuros, propõe-se a incorporação de bases de dados que incluam informações individuais dos estudantes, o que possibilitará uma análise mais aprofundada das motivações associadas ao abandono. Além disso, a aplicação da metodologia em cursos de outras áreas do conhecimento poderá enriquecer o entendimento comparativo sobre os fatores que favorecem ou dificultam a permanência nos diferentes contextos acadêmicos. Em relação aos métodos de inteligência artificial, outros modelos podem ser considerados, como estratégias de aprendizado profundo; também pode-se analisar outras técnicas explicáveis, como o *permutation importance* (Louppe, 2015) e o *Local Interpretable Model Agnostic Explanation* (LIME) (Salih *et al.*, 2024).

Por fim, conclui-se que a presente pesquisa contribui para o entendimento do fenômeno da evasão nos cursos de computação e tecnologia no Brasil ao evidenciar quais características dos cursos possuem maior impacto tanto na desistência quanto na permanência estudantil, fornecendo subsídios para gestores educacionais de IES públicas e privadas desenvolverem políticas de retenção mais eficazes.



Referências

- Alvim, Í. V.; Bittencourt, R. A.; Duran, R. S. Evasão nos cursos de graduação em computação no Brasil. In: SBC. **Anais do IV Simpósio Brasileiro de Educação em Computação**. [S.l.], 2024. p. 1–11.
- Barbosa, M. L. d. O.; Pires, A. An exploratory analysis of admission and inclusion in Brazilian higher education: from inherited to relative merit. **Pro-Posições**, SciELO Brasil, v. 33, p. e20210070, 2022.
- Costa, C. T. da; Júnior, M. L. L.; Santana, A. M. União de dados por clusterização para construção de modelos de predição de evasão. **Revista Novas Tecnologias na Educação - RENOTE**, CINTED-UFRGS, v. 21, n. 2, 2023. Disponível em: <https://orcid.org/0009-0003-5144-0727>.
- Inep, I. N. de Estudos e P. E. A. T. **Censo da Educação Superior**. 2025. Disponível em: <https://www.gov.br/inep/pt-br/areas-de-atuacao/pesquisas-estatisticas-e-indicadores/censo-da-educacao-superior>.
- Jesus, H. O. D.; Rodriguez, L. C.; Junior, A. d. O. C. Predição de evasão escolar na licenciatura em computação. **Revista Brasileira de Informática na Educação**, v. 29, p. 255–272, 2021.
- Khosravi, H. *et al.* Explainable artificial intelligence in education. **Computers and education: artificial intelligence**, Elsevier, v. 3, p. 100074, 2022.
- Kuhn, M.; Johnson, K.; Kuhn, M.; Johnson, K. Data pre-processing. **Applied predictive modeling**, Springer, p. 27–59, 2013.
- Lima, J. A. S.; Fagundes, R. A. A. Análise comparativa de algoritmos de machine learning para prever a evasão escolar: Uma revisão sistemática da literatura. **Revista Novas Tecnologias na Educação - RENOTE**, CINTED-UFRGS, v. 21, n. 2, 2023. Disponível em: <https://orcid.org/0000-0001-5010-0024>.
- Lopes, C. E. S.; Lima, J. A. da S.; Fagundes, R. A. A. Aplicação de algoritmos de classificação para prever a evasão escolar. **Revista Novas Tecnologias na Educação - RENOTE**, CINTED-UFRGS, v. 21, n. 2, 2023. Disponível em: <https://orcid.org/0000-0001-7969-5754>.
- Louppe, G. **Understanding Random Forests: From Theory to Practice**. 2015. Disponível em: <https://arxiv.org/abs/1407.7502>.
- Lundberg, S. M.; Lee, S.-I. A unified approach to interpreting model predictions. In: Guyon, I. *et al.* (Ed.). **Advances in Neural Information Processing Systems 30**. Curran Associates, Inc., 2017. p. 4765–4774. Disponível em: <http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf>.
- Mustofa, S.; Emon, Y. R.; Mamun, S. B.; Akhy, S. A.; Ahad, M. T. A novel ai-driven model for student dropout risk analysis with explainable ai insights. **Computers and Education: Artificial Intelligence**, Elsevier, v. 8, p. 100352, 2025.
- 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 - RBIE**, Sociedade Brasileira de Computação, v. 32, p. 270–294, 2024. Disponível em: <https://sol.sbc.org.br/journals/index.php/rbie>.
- Nti, I. K.; Nyarko-Boateng, O.; Aning, J. *et al.* Performance of machine learning algorithms with different k values in k-fold cross-validation. **International Journal of Information Technology and Computer Science**, MECS Publisher, v. 13, n. 6, p. 61–71, 2021.
- Oliveira, R. dos S.; Medeiros, F. P. A. de. Modelo de predição de evasão escolar com base em dados de autoavaliação de cursos de graduação. **Revista Brasileira de Informática**



- na Educação - **RBIE**, Sociedade Brasileira de Computação, v. 32, p. 01–21, 2024. Disponível em: <https://doi.org/10.5753/rbie.2024.3542>.
- Salih, A. M. *et al.* A perspective on explainable artificial intelligence methods: Shap and lime. **Advanced Intelligent Systems**, Wiley, v. 7, n. 1, jun. 2024. ISSN 2640-4567. Disponível em: <http://dx.doi.org/10.1002/aisy.202400304>.
- Scornet, E. Trees, forests, and impurity-based variable importance in regression. In: Institut Henri Poincaré. **Annales de l'Institut Henri Poincaré (B) Probabilités et statistiques**. [S.l.], 2023. v. 59, n. 1, p. 21–52.
- Silva, R. A. d. S.; AF, B. B.; Fátima, P. F. Maria de; Santos, I. de S.; Andrade, R. M. Evasão em computação na ufc sob a perspectiva dos alunos. In: SBC. **Workshop sobre Educação em Computação (WEI)**. [S.l.], 2021. p. 338–347.
- Sousa, M. K. V. de; Albuquerque, D. W.; Filho, E. D.; Perkusich, M. B. Um survey investigativo sobre aspectos motivadores do ingresso e evasão dos alunos dos cursos técnicos da Área de informática. **Revista Brasileira de Informática na Educação - RBIE**, Sociedade Brasileira de Computação, v. 31, p. 467–487, 2023. Disponível em: <http://br-ie.org/pub/index.php/rbie>.
- Tran, N.; Schneider, J.-G.; Weber, I.; Qin, A. Hyper-parameter optimization in classification: To-do or not-to-do. **Pattern Recognition**, v. 103, p. 107245, 2020. ISSN 0031-3203. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0031320320300510>.
- Vasconcelos, R. M.; Amaral, L.; Ciampi, M. M.; Brito, C. R. Motivation to choose the night course in engineering. In: IEEE. **2020 IEEE World Conference on Engineering Education (EDUNINE)**. [S.l.], 2020. p. 1–4.
- Yu, T.; Zhu, H. **Hyper-Parameter Optimization: A Review of Algorithms and Applications**. 2020. Disponível em: <https://arxiv.org/abs/2003.05689>.