Revista Eletrônica de Iniciação Científica em Computação, 2026,24:1
ISSN: 3085-8461 • doi: 10.5753/reic.2026.7215

Artigo de Pesquisa/Research Paper

Predição da Evasão de Estudantes Ingressantes no IFCE *Campus* Tianguá Utilizando Técnicas de Aprendizagem de Máquina Supervisionada

Predicting Dropout of Incoming Students at IFCE Campus Tianguá Using Supervised Machine Learning Techniques

Everton Almeida [Instituto Federal de Educação, Ciência e Tecnologia do Ceará | *everton.almeida.veraso8@aluno.ifce.edu.br*]
Raquel Silveira [Instituto Federal de Educação, Ciência e Tecnologia do Ceará | *raquel_silveira@ifce.edu.br*]
Adonias Caetano de Oliveira [Instituto Federal de Educação, Ciência e Tecnologia do Ceará |*adonias.oliveira@ifce.edu.br*]

Bacharelado em Ciência da Computação, Instituto Federal de Educação, Ciência e Tecnologia do Ceará Campus Tianguá, AV. Tabelião Luiz Nogueira de Lima, 62324-075, Brasil.

Resumo. A evasão no ensino superior público brasileiro compromete a permanência acadêmica e a gestão institucional, gerando implicações sociais e econômicas. Esse contexto demanda estratégias multidisciplinares para promover a permanência dos discentes e otimizar a gestão acadêmica. Neste estudo, busca-se aplicar métodos de aprendizado de máquina na análise de dados educacionais para predizer os estudantes ingressantes do Instituto Federal do Ceará (IFCE), *campus* Tianguá, em potencial risco de evasão. Utiliza-se registros acadêmicos e sociodemográficos de estudantes, abrangendo o período de 2010.2 a 2025.2. Esses dados foram submetidos a processos de preparação e refinamento. Técnicas de classificação supervisionada foram testadas em diferentes configurações, considerando o desbalanceamento dos dados e ajustes de hiperparâmetros para aprimorar o desempenho dos algoritmos. Os resultados revelam a viabilidade de modelos preditivos em identificar padrões de evasão escolar, auxiliando na implementação de medidas preventivas, o que contribui para políticas mais eficazes de evasão no ambiente educacional.

Abstract. Student dropout in Brazilian public higher education undermines academic retention and institutional management, generating social and economic implications. This context requires multidisciplinary strategies to promote student persistence and optimize academic management. This study applies machine learning methods to educational data analysis to predict incoming students at the Federal Institute of Ceará (IFCE), Tianguá campus, at potential risk of dropout. Academic and sociodemographic records from students were used, covering the period from 2010.2 to 2025.2. These data were subjected to preparation and refinement processes. Supervised classification techniques were tested in different configurations, considering data imbalance and hyperparameter adjustments to improve algorithm performance. The results reveal the viability of predictive models in identifying school dropout patterns, aiding the implementation of preventive measures, which contributes to more effective policies against dropout in the educational environment.

Palavras-chave: evasão, aprendizado de máquina, análise preditiva, permanência acadêmica, educação superior

Keywords: dropout, machine learning, predictive analysis, academic retention, higher education

Recebido/Received: 17 December 2025 • **Aceito/Accepted:** 03 June 2026 • **Publicado/Published:** 10 July 2026

1 Introdução

A permanência dos estudantes no Ensino Superior é um desafio recorrente enfrentado por instituições públicas em todo o país. A evasão nas instituições de ensino superior do Brasil apresenta índices alarmantes, com taxas acima de 20% no ano de 2023, conforme o Censo da Educação Superior do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP) [INEP, 2024].

No âmbito do Instituto Federal do Ceará (IFCE), campus Tianguá, a diversidade de perfis estudantis, aliada a particularidades regionais, intensifica o desafio da evasão. Um estudo prévio mapeou os índices acadêmicos do IFCE campus Tianguá por meio de um dashboard interativo em Power BI, revelando que, entre 2010 e 2024, 47,6% dos ingressantes não concluíram seus cursos, com aumento notável da evasão durante a pandemia [Linhares *et al*., 2025].

A evasão estudantil persiste como problema estrutural no ensino superior brasileiro, afetando instituições públicas e privadas. Apesar de políticas nacionais de ampliação do acesso, como o Portal Único de Acesso ao Ensino Superior (PROUNI) e o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI), que expandem oportunidades de ingresso, elas não garantem permanência e conclusão dos cursos, gerando perdas para estudantes, instituições e sociedade, além de aprofundar desigualdades socioeconômicas e comprometer o desenvolvimento nacional [SEMESP, 2025].

A evasão no ensino superior público brasileiro representa um desafio multifacetado, com repercussões sociais, econômicas e institucionais [Arantes *et al*., 2021]. Fatores acadêmicos, pessoais e socioeconômicos interagem para moldar as trajetórias discentes, demandando políticas de permanência adaptadas ao contexto [Amorim *et al*., 2025]. Análises quantitativas revelam associações entre variáveis como idade ao ingresso, renda familiar, suporte financeiro e tipo de escola de origem com o risco de abandono [Saccaro *et al*., 2024].

O abandono precoce compromete a mobilidade social, particularmente para estudantes de origens vulneráveis, e resulta em desperdício de recursos públicos alocados em matrículas, infraestrutura e suporte, sem o correspondente retorno em formados. Ademais, elevadas taxas de evasão impactam o planejamento institucional e a manutenção da qualidade acadêmica [Arantes *et al*., 2021]. Identificar de forma sistemática os fatores mais relevantes para a evasão em contextos específicos é fundamental para o planejamento de políticas institucionais de permanência e êxito [Silva and Sampaio, 2022].

Nesse cenário, este trabalho tem como objetivo desenvolver um modelo preditivo para identificar a evasão entre estudantes ingressantes no IFCE campus Tianguá, visando identificar precocemente riscos de evasão, sendo um instrumento para subsidiar intervenções institucionais de permanência e êxito dos estudantes. Para isso, foram aplicadas etapas de preparação e processamento de dados em uma base de dados extraída do sistema QAcadêmico[1]. Após essa etapa, foram avaliados cinco algoritmos de classificação, permitindo comparar o desempenho desses algoritmos na predição da evasão escolar.

Este trabalho está organizado da seguinte maneira: a seção 2 apresenta os trabalhos relacionados com este estudo; a seção 3 aborda o referencial teórico; a seção 4 detalha a metodologia utilizada neste trabalho; a seção 5 expõe os resultados e discussões; e, por fim, a seção 6 apresenta conclusões e sugestões para pesquisas futuras.

2 Fundamentação Teórica

Esta seção apresenta a base teórica que sustenta o estudo, de modo a explicitar definições que embasam este trabalho. Está organizada em subseções sobre evasão escolar, análise de dados educacionais, aprendizagem de máquina, técnicas de balanceamento de dados, amostragem de dados e métricas de avaliação de modelos.

2.1 Evasão escolar

A evasão no ensino superior é usualmente definida como a saída do estudante do curso antes da sua conclusão, incluindo abandono voluntário, desligamento e transferências que resultam em não conclusão. As explicações teóricas clássicas enfocam interações entre fatores individuais, acadêmicos e institucionais, que relacionam integração acadêmica e social à permanência estudantil [INEP, 2024; Baggi and Lopes, 2011].

Pesquisas mais recentes reavaliam e expandem esse arcabouço, incorporando determinantes socioeconômicos, trajetórias educacionais (por exemplo, intervalo entre ensino médio e ingresso), condições de financiamento e aspectos regionais que afetam a vulnerabilidade dos ingressantes. Estudos empíricos realizados no contexto brasileiro e internacional mostram correlações consistentes entre baixa renda, atraso na transição educacional, origem em escola pública e maior risco de evasão [Saccaro *et al*., 2024].

No âmbito do IFCE *campus* Tianguá, a evasão reflete particularidades regionais e perfis diversificados, com taxas elevadas durante períodos como a pandemia, reforçando a necessidade de intervenções precoces baseadas em dados [Linhares *et al*., 2025].

2.2 Aprendizagem de máquina

A Aprendizagem de Máquina (do inglês, *Machine Learning*, ML) é um subcampo da Inteligência Artificial cujo objetivo é desenvolver algoritmos capazes de aprender padrões a partir de dados e realizar predições ou decisões sem programação explícita [Géron, 2019].

A Mineração de Dados Educacionais (EDM) aplica técnicas de ML para extrair padrões de dados educacionais, visando melhorar processos educacionais, como a predição de evasão. Essa abordagem integra análise exploratória, préprocessamento e modelagem para identificar fatores de risco em trajetórias estudantis [Han *et al*., 2012].

No contexto brasileiro, mapeamentos sistemáticos revelam o uso predominante de dados demográficos e acadêmicos em estudos de EDM para evasão, com algoritmos supervisionados destacando-se para subsidiar políticas de retenção. Lacunas incluem a escassez de referências consolidadas em instituições federais, enfatizando a relevância de análises locais como esta [Bastos *et al*., 2025].

Este estudo foca na aprendizagem supervisionada, onde modelos são treinados com dados rotulados para mapear as saídas a partir das entradas, como classificação de êxito ou evasão [Géron, 2019]. Os algoritmos utilizados incluem:

Árvore de Decisão (do inglês Decision Tree, DT): estrutura de decisões hierárquicas, dividindo o espaço de dados em regiões por meio de regras sucessivas sobre os atributos. A interpretabilidade é uma das principais vantagens, permitindo visualizar regras explícitas [Breiman, 2001; James *et al*., 2023].

Random Forest (RF): um método baseado em múltiplas árvores de decisão treinadas de forma independente com amostragem bootstrap (bagging). Cada árvore recebe subconjuntos aleatórios de instâncias e atributos, o que reduz variância e melhora a generalização [Breiman, 2001; Géron, 2019].

K-Nearest Neighbors (KNN): classifica por proximidade a k vizinhos, capturando padrões locais, mas sensível a escala e dimensionalidade [Hastie *et al*., 2009; Murphy, 2022].

Multi-Layer Perceptron (MLP): um tipo de rede neural artificial composta por camadas densas formadas por neurônios artificiais. Cada neurônio aplica uma transformação linear seguida de uma função de ativação. O treinamento utiliza retropropagação do erro [Goodfellow *et al*., 2016; Géron, 2019].

Extreme Gradient Boosting (XGBoost): é uma implementação otimizada de *gradient boosting*, no qual árvores fracas são treinadas sequencialmente, cada uma corrigindo os erros da anterior. Apresenta regularização, *subsampling*, paralelização e controle de complexidade da árvore, o que resulta em bom desempenho e eficiência em bases estruturadas [Chen and Guestrin, 2016].

[1] Um dos sistemas de gestão acadêmica utilizado pelo IFCE. O acesso ao sistema pode ser realizado por meio do *link*: `https://qacademico.ifce.edu.br`.

2.3 Balanceamento de dados

O desbalanceamento de classes ocorre quando uma das categorias de um problema de classificação possui número significativamente menor de instâncias do que as demais, o que leva modelos supervisionados a favorecerem a classe majoritária e apresentarem baixo desempenho na identificação da classe minoritária [He and Garcia, 2009].

Para mitigar esse viés, foram adotadas duas técnicas amplamente utilizadas de balanceamento de dados: *Random Oversampler* e *Synthetic Minority Oversampling Technique* (SMOTE), ambas voltadas ao *oversampling*, ou seja, ao aumento da representatividade da classe minoritária. [Géron, 2019].

Random Oversampler: realiza *oversampling* por meio da replicação aleatória de exemplos da classe minoritária até atingir uma distribuição mais equilibrada [Lemaître *et al*., 2016].

SMOTE: proposto por Chawla *et al*. [2002], introduz uma estratégia de *oversampling* sintético ao gerar novas instâncias artificiais da classe minoritária.

2.4 Amostragem dos dados

A avaliação adequada de modelos supervisionados depende de estratégias que permitam estimar sua capacidade de generalização, isto é, o desempenho em dados não utilizados durante o ajuste dos parâmetros. Entre os procedimentos mais consolidados na literatura estão o método *hold-out* e a validação cruzada, ambos amplamente empregados em estudos empíricos de classificação [Géron, 2019; James *et al*., 2023].

Hold-out: divide, aleatoriamente, dados em treino e teste. O conjunto de treinamento é usado para ajustar o modelo, enquanto o conjunto de teste é destinado exclusivamente à avaliação final [Murphy, 2022].

Validação cruzada k-fold: particiona os dados em *k folds* mantendo-se a distribuição das classes em cada um deles. Em cada iteração, um *fold* é usado como validação e os demais como treinamento, repetindo-se o processo até que todos os subconjuntos tenham desempenhado o papel de conjunto de validação. As métricas são então agregadas, produzindo estimativas mais estáveis do desempenho do modelo [Hastie *et al*., 2009].

2.5 Métricas de avaliação dos modelos

A avaliação de modelos supervisionados é fundamental para medir a capacidade de generalização [James *et al*., 2023; Murphy, 2022]. As principais métricas utilizadas nessa categoria de modelos são:

Acurácia: representa a proporção total de predições corretas em relação ao conjunto de dados. Embora forneça uma visão geral da performance, é limitada em cenários desbalanceados, nos quais pode superestimar o desempenho ao favorecer a classe majoritária.

Precisão: indica a fração de predições positivas que são realmente corretas, sendo útil para avaliar a ocorrência de falsos positivos.

Recall: mede a fração de positivos reais identificados corretamente, refletindo a capacidade do modelo de capturar casos relevantes da classe de interesse.

F1-score: média harmônica entre precisão e *recall*, equilibrando ambos os indicadores. Por ser menos sensível ao desbalanceamento, é frequentemente adotado como métrica principal em avaliações de classificadores binários [Géron, 2019].

F1 macro: média aritmética dos *F1-scores* de cada classe, atribuindo pesos iguais a todas elas. Dessa forma, essa métrica enfatiza o equilíbrio do desempenho entre classes, independentemente de sua frequência no conjunto de dados [Murphy, 2022].

F1 weighted: média ponderada dos *F1-scores* por classe, considerando a proporção de amostras de cada classe, refletindo o desempenho global do modelo de forma mais alinhada à distribuição real dos dados [Murphy, 2022].

Matriz de confusão: sintetiza os acertos e erros do algoritmo ao distinguir entre verdadeiros positivos, falsos positivos, verdadeiros negativos e falsos negativos. Essa representação permite identificar padrões de erro e compreender de forma direta como o modelo se comporta em cada classe [Géron, 2019; Murphy, 2022].

Area Under the Curve (AUC): a curva *Receiver Operating Characteristic* (ROC) e a AUC podem ser utilizadas para examinar a capacidade discriminativa do modelo em diferentes limiares de decisão, indicando sua habilidade geral de separação entre as classes [Géron, 2019].

O uso combinado dessas métricas fornece uma avaliação abrangente e confiável, permitindo comparar algoritmos de forma consistente e fundamentada [Murphy, 2022].

3 Trabalhos Relacionados

O emprego de técnicas de aprendizado de máquina para predizer a evasão estudantil tem ganhado destaque em estudos brasileiros recentes. Esses estudos exploram dados administrativos de instituições federais para revelar padrões de abandono e subsidiar políticas de permanência e êxito dos estudantes. A abordagem demonstra o potencial de algoritmos supervisionados em contextos educacionais, com ênfase em mineração e pré-processamento de dados.

Primão [2022] desenvolveu um modelo preditivo de evasão no Instituto Federal de Santa Catarina, integrando dados acadêmicos e socioeconômicos de estudantes de graduação antes e durante a pandemia de Covid-19. Testou algoritmos como DT, MLP e *XGBoost*, avaliados por acurácia, precisão, *recall* e *F1-score*. O *XGBoost* apresentou o melhor desempenho, com *F1-score* de 97,53% antes da pandemia e 90,32% durante.

Araújo [2025] analisou a predição de evasão no curso de Ciência da Computação da Universidade Federal do Ceará *Campus* Quixadá, utilizando dados de ingressantes de 2013 a 2019. Empregaram algoritmos como RF, Regressão Logística, DT, *Gradient Boosting* e SVM, com pré-processamento via Python, validação cruzada *k-fold* e ajuste de hiperparâmetros via *grid search*. O RF obteve acurácia de 85% e *recall* de 96%, destacando correlações entre notas iniciais, frequência e risco de abandono.

Malerba [2024] desenvolveu modelos preditivos na Uni-

versidade Federal de Itajubá, usando uma base pública do Kaggle com 35 atributos demográficos, socioeconômicos e de desempenho. Aplicou algoritmos de classificação supervisionada, incluindo Regressão Logística, *Gradient Boosting*, *XGBoost*, RF, DT e SVM, com pré-processamento e validação cruzada *k-fold*. No modelo para recuperação de evadidos, obteve *F1-scores* acima de 90% para formandos e 89% para evadidos. No modelo interpretável para desligamento, RF e DT destacaram-se com 91% para formandos e 84% para evadidos, identificando desempenho inicial como preditor chave.

Esses trabalhos convergem com este estudo no uso de algoritmos de aprendizado de máquina supervisionada e dados acadêmicos e socioeconômicos para predizer evasão. No entanto, diferenciam-se pelo foco em contextos institucionais específicos, enquanto esta pesquisa concentra-se no *campus* Tianguá do IFCE, abrangendo ensino superior e técnico, com dados de 2010.2 a 2025.2 e ênfase nos ingressantes. A avaliação sistemática de técnicas de balanceamento como SMOTE e *Random Oversampler* oferece uma contribuição complementar.

4 Metodologia

A metodologia adotada neste estudo foi estruturada para permitir a construção, preparação e avaliação de modelos de ML. O processo metodológico compreende cinco etapas principais: obtenção e caracterização dos dados, Análise exploratória dos dados, pré-processamento e engenharia de atributos, modelagem supervisionada e avaliação.

Essas etapas foram implementadas no ambiente *Google Colaboratory* utilizando a linguagem Python e as seguintes bibliotecas: *pandas, numpy, matplotlib, seaborn, scikit-learn*, e *XGBoost*.

Na Figura 1, é exibido o fluxograma do processo metodológico adotado neste estudo.

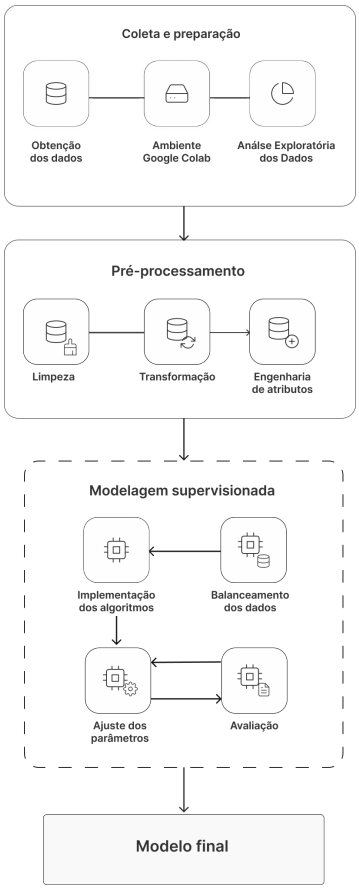

Figura 1. Fluxograma metodológico do estudo.

4.1 Obtenção e caracterização dos dados

Os dados foram extraídos do Sistema QAcadêmico, por meio de uma *view* da base de dados desse sistema, disponibilizada para os autores deste trabalho para fins de pesquisa. Para este trabalho, os dados foram filtrados, contendo exclusivamente registros dos estudantes do IFCE *campus* Tianguá, compreendendo 6.992 estudantes ingressantes no período 2010.2 a 2025.2, abrangendo tanto cursos de nível superior, quanto técnico.

Os dados disponibilizados na *view* já se encontram anonimizados, sem inclusão de informações pessoais sensíveis em conformidade com a Lei Geral de Proteção de Dados (LGPD) e princípios éticos de pesquisa.

Em termos de caracterização inicial, o *dataset* possui 174 variáveis, distribuídas em categorias sociodemográficas (tais como, sexo, cor/raça e renda familiar), acadêmicas (tais como, forma de ingresso, cota e modalidade do curso) e temporais (tais como, período letivo de ingresso e data de nascimento).

4.2 Análise exploratória dos dados

Realizou-se a análise exploratória e descritiva das 174 variáveis originais, de modo a compreender as variáveis disponíveis para a predição da evasão escolar dos estudantes. Complementarmente, analisou-se a qualidade dos dados, identificando o percentual de valores ausentes em cada uma das variáveis, assim como a distribuição dos dados e a correlação existente entre as variáveis.

Esta etapa subsidiou a aplicação de etapas de pré-processamento e a seleção de atributos que realmente estão disponíveis no momento do ingresso do estudante e que são relevantes para a referida tarefa.

4.3 Pré-processamento e engenharia de atributos

A etapa de pré-processamento e engenharia de atributos visa preparar os dados brutos para a aplicação de algoritmos de aprendizado de máquina, possibilitando que os dados estejam limpos, consistentes e otimizados para maximizar a performance preditiva dos algoritmos (GÉRON, 2019).

As principais etapas incluíram a definição da variável alvo (aquela que se pretende aprender), limpeza e seleção de variáveis, criação de novos atributos e transformação dos dados, conforme descrito a seguir.

4.3.1 Definição da variável alvo

Considerando a tarefa de predizer se um estudante ingressante tem potencial risco de evasão ou conclusão do curso com êxito, utilizou-se a variável "Situação de Matrícula" como variável alvo.

Essa variável registra o status acadêmico de forma granular do estudante. Com o objetivo de considerar apenas os estudantes que concluíram o ciclo de ensino, seja com êxito ou sem êxito, foram removidos todos os registros com situação "Matriculado", ou seja, sem desfecho definido. Após esse filtro, permaneceram 6.207 instâncias.

Em seguida, os valores da variável alvo "Situação de Matrícula" foram agrupados em duas categorias, de acordo com a classificação adotada pelo Plano Estratégico para Permanência e Êxito dos Estudantes do IFCE [IFCE, 2017]. Essa categorização é utilizada institucionalmente para o monitoramento dos indicadores de permanência e êxito estudantil e compreende três classes principais: "Matriculado", "Êxito" e "Sem Êxito". Ressalta-se que, conforme mencionado anteriormente, a categoria "Matriculado" foi removida desta análise por representar estudantes com trajetória acadêmica ainda em andamento.

- Êxito: Categoria composta pelos estudantes que concluíram o curso com sucesso. Esta categoria agrupa as situações de matrícula concluído, formado, concludente, aguardando colação de grau e ENADE.
- Sem Êxito: Categoria composta pelos estudantes que não concluíram o curso, ou seja, situações de evasão. Esta categoria agrupando as situações de matrícula abandono, cancelamento (voluntário ou compulsório), trancamento, transferência, intercâmbio, vínculo Institucional e não concluído.

A Tabela 1 apresenta a distribuição das instâncias do conjunto de dados, conforme os valores das classes utilizadas para a tarefa de predição da evasão escolar estudantil.

Tabela 1. Distribuição das instâncias do conjunto de dados, conforme as classes.

Classe	Quantidade	Porcentagem
Sem Êxito	3.891	63,7%
Êxito	2.316	37,3%
Total	**6.207**	**100%**

4.3.2 *Limpeza, padronização e seleção de variáveis*

A limpeza dos dados busca eliminar ruídos e inconsistências que podem afetar a qualidade das predições, enquanto a seleção das variáveis reduz o conjunto para atributos relevantes ao foco da predição da evasão escolar dos estudantes ingressantes, evitando o processamento de informações irrelevantes ou associadas ao pós-ingresso do estudante. Portanto, foram aplicados os seguintes procedimentos:

- Remoção de variáveis com mais de 50% de valores ausentes
- Eliminação de variáveis duplicadas ou equivalentes por inspeção manual
- Remoção de atributos associados ao pós-ingresso do estudante: Variáveis como coeficiente de rendimento ou situação semestral, disponíveis apenas após o ingresso do estudante, foram excluídas para garantir que os modelos reflitam predições baseadas unicamente em dados iniciais do estudante
- Padronização de formatos e correções de inconsistências textuais: Categorias com grafias variadas (por exemplo, "M"e "Masculino"em sexo) foram uniformizadas para evitar erros na codificação posterior e assegurar consistência nas análises categóricas.

4.3.3 *Criação de novas variáveis*

A engenharia de atributos motiva-se pela necessidade de extrair informações derivadas que capturem padrões implícitos nos dados originais, enriquecendo o poder preditivo dos modelos sem introduzir viés [Géron, 2019]. Foram criadas duas novas variáveis, estruturadas a partir de dados existentes no *dataset*:

- Idade de ingresso: Representa a idade do estudante ao ingressar no curso. Calculada como a diferença entre a data de nascimento e a data da matrícula;
- Tempo do ensino médio à graduação (em anos): Representa o intervalo de tempo, em anos, entre a conclusão do ensino médio e o início da graduação. Calculada como a diferença em anos entre a conclusão do ensino médio e a data da matrícula.

Ao final dessa etapa, das 174 colunas iniciais, restaram 25 variáveis selecionadas, focadas em características sociodemográficas e acadêmicas disponíveis no momento do ingresso do estudante no curso, a saber: Período letivo de ingresso, Descrição do curso, Sexo, Forma de ingresso, Área de procedência da escola de origem, Pai falecido, Mãe falecida, Profissão, Escola do 2º grau, Estado civil, Tipo de escola de origem, Nível de ensino, Cota, Modalidade do curso, Turno inicial, Cor/Raça, Cidade, Escola de origem, Renda familiar per capita, Renda familiar, Naturalidade, Turno, Matriz curricular, Tempo do ensino médio à graduação (em anos) e Idade de ingresso.

4.3.4 *Transformação dos dados*

A transformação proporciona que os dados sejam interpretáveis pelos algoritmos, ajustando escalas e formatos para evitar distorções em modelos sensíveis a variações numéricas ou categóricas [Géron, 2019]. Nesse contexto, foram aplicadas as seguintes etapas de transformação dos dados:

- Codificação de variáveis categóricas: Conversão dos valores das variáveis categóricas (por exemplo, sexo e cota) em valores numéricos, permitindo a aplicação de algoritmos que requerem entradas quantitativas. Esse processo foi feito utilizando a ferramenta LabelEncoder presente na biblioteca *scikit-learn*;
- Normalização e padronização de variáveis numéricas: Aplicadas para redimensionar valores das variáveis numéricas em 0 e 1, preservando distribuições originais sem afetar a interpretabilidade. Este procedimento foi utilizado apenas nos algoritmos KNN e MLP, que são sensíveis a escalas discrepantes.

4.4 Modelagem

A modelagem constitui a etapa central do estudo, na qual foram avaliados cinco algoritmos de aprendizado de máquina supervisionado: DT, RF, *XGBoost*, KNN e MLP. A implementação foi feita usando as bibliotecas *scikit-learn* e *XGBoost*.

O conjunto de dados foi dividido por hold-out, estratificado na proporção de 80% para treino e 20% para teste, preservando a distribuição original das classes no conjunto de teste.

4.4.1 Ajuste de hiperparâmetros

A busca pelos melhores hiperparâmetros foi realizada mediante a técnica de *grid search* com validação cruzada estratificada de 5 *folds* sobre o conjunto de treino, utilizando o *F1-score* como métrica de referência.

Para cada algoritmo foram explorados hiperparâmetros específicos, a definição dos intervalos testados seguiu um processo iterativo partindo de experimentações preliminares. Inicialmente, foram adotados valores mínimos e máximos. A partir desses testes iniciais, os intervalos foram ajustados conforme o comportamento observado dos modelos, permitindo concentrar a busca nas faixas mais promissoras em termos de estabilidade e desempenho. A seguir são apresentados os hiperparâmetros explorados.

- DT: investigaram-se variações na profundidade máxima (max_depth) da árvore (valores entre 5 e 20), o número mínimo de amostras para divisão interna ($min\_samples\_split$), com valores de 2, 5 e 10, número mínimo de amostras por nó folha ($min\_samples\_leaf$), com valores de 1, 2 e 4, e o critério de impureza ($criterion$) adotado (*gini* e *entropy*).
- RF: número de árvores ($n\_estimators$), com valores de 200, 300, 500 estimadores, profundidade máxima (max_depth) variando entre 5 e 20, além de diferentes valores para o número mínimo de amostras para divisão ($min\_samples\_split$), com valores de 2, 5, e 10 e para o número mínimo de amostras por folha ($min\_samples\_leaf$), com valores de 1, 2 ou 4.
- *XGBoost*: foram exploradas combinações envolvendo diferentes profundidades máximas das árvores (max_depth) entre 5 e 20, taxas de aprendizagem ($learning\_rate$) de 0,01, 0.1 e 0,2 e número de estimadores ($n\_estimators$) de 200, 300 e 500.
- KNN: foram avaliados valores de k entre 3 e 11, considerando diferentes esquemas de ponderação dos vizinhos (uniform e distance) e duas métricas de distância: Manhattan e Euclidiana.
- MLP: foram testadas arquiteturas com 1 e 2 camadas ocultas com variação na quantidade de neurônios por camada (16 a 64 neurônios por camada), camadas: ((32), (32,16), (64,32)), função de ativação ReLU, regularização α de 0.0001, 0.001 e 0.01, e taxa de aprendizagem ($learning\_rate$) entre 0,01 e 0,1.

Ao final da *grid search*, os parâmetros que resultam nos melhores valores foram então utilizados para treinar os modelos finais e empregados na etapa subsequente de análise dos resultados.

4.4.2 Balanceamento de dados

Devido ao desbalanceamento das classes, cada algoritmo foi treinado e avaliado em três cenários independentes: (i) Sem balanceamento (distribuição original dos dados); (ii) SMOTE; e (iii) *Random Oversampler*.

A inclusão desses cenários permitiu analisar a sensibilidade dos modelos ao desbalanceamento e selecionar a estratégia com os melhores resultados.

4.4.3 Avaliação

A avaliação dos modelos foi realizada utilizando o conjunto de teste (20% dos dados). Nesse contexto, as métricas consideradas nesta etapa foram: acurácia, *F1-score* de cada classe e AUC.

Essas métricas foram computadas para cada algoritmo em cada cenário de balanceamento, permitindo comparações quantitativas. Além disso, a matriz de confusão foi empregada como instrumento complementar de análise.

4.4.4 Análise de importância das variáveis

A fim de identificar as *features* mais influentes na predição, extraiu-se a importância das variáveis nos modelos RF e *XGBoost*. No RF, obteve-se diretamente a importância das *features*, enquanto que no *XGBoost*, empregou-se o método de ganho (*gain*), que mede a contribuição relativa de cada variável para a melhoria do modelo.

5 Resultados e discussão

Esta seção apresenta e discute os melhores os resultados obtidos pelos cinco algoritmos avaliados aplicados sob três cenários distintos de balanceamento de dados, conforme descrito na seção de metodologia.

5.1 Hiperparâmetros finais

A partir da aplicação da estratégia de *grid search*, obteve-se os seguintes hiperparâmetros para cada algoritmo:

- DT: observou-se variação clara entre os cenários. Sem balanceamento, o melhor modelo utilizou $criterion = gini$ e $max\_depth = 20$, reforçando a tendência à maior profundidade em dados desbalanceados. Com SMOTE, a melhor configuração foi $criterion = entropy$, $max_depth = 10$, $min\_samples\_split = 10$ e $min\_samples\_leaf = 1$. Já no *Random Oversampler*, obteve-se um modelo mais compacto ($max_depth = 5$), também com $criterion = entropy$.
- RF: apresentou, no cenário sem balanceamento, seu melhor desempenho com árvores mais profundas ($max\_depth = 20$) e maior número de estimadores ($n\_estimators = 500$), além de $min\_samples\_split = 2$ e $min\_samples\_leaf = 1$. Com SMOTE, o modelo ótimo reduziu a profundidade ($max\_depth = 10$) e utilizou $min\_samples\_split = 10$, $min\_samples\_leaf = 2$ e $n\_estimators = 200$. No *Random Oversampler*, a profundidade permaneceu em $max\_depth = 10$, mas a melhor configuração incluiu $min\_samples\_split = 2$, $min\_samples\_leaf = 4$ e $n\_estimators = 500$, sugerindo maior estabilidade com o uso de mais árvores após o balanceamento.
- *XGBoost*: o cenário sem balanceamento apresentou melhor desempenho com um modelo mais complexo, combinando taxa de aprendizado elevada e grande profundidade ($learning\_rate = 0.2$, $max\_depth = 10$, $n\_estimators = 500$). Nos cenários balanceados, tanto com SMOTE, quanto com *Random Oversampler*, o modelo utilizou $learning\_rate = 0.01$, $max_depth = 5$ e $n\_estimators = 300$ (SMOTE) e 500 (*Random Oversampler*.

- KNN: manteve-se o padrão de ponderação por distância em todos os cenários, com variação no número de vizinhos: 3 no cenário sem balanceamento, 11 com SMOTE e 7 com *Random Oversampler*.
- MLP: os resultados indicaram arquiteturas distintas conforme o balanceamento. O modelo sem reamostragem apresentou melhor desempenho com ativação ReLU e arquitetura composta por duas camadas ocultas com 32 e 16 neurônios, respectivamente, utilizando o otimizador *Limited-memory Broyden-Fletcher-Goldfarb-Shanno* (LBFGS), taxa de aprendizado inicial $learning\_rate = 0{,}001$ e regularização $\alpha = 0{,}001$. Em SMOTE, a configuração ideal manteve a ativação ReLU e a mesma estrutura de duas camadas (32, 16). No *Random Oversampler*, o melhor ajuste também preservou essa arquitetura compacta, reforçando que redes menos profundas foram suficientes para capturar os padrões presentes nos dados.

5.2 Visão geral do desempenho dos modelos

A Tabela 2 mostra o desempenho comparativo de cada algoritmo sob diferentes cenários de balanceamento, considerando as métricas de precisão, *recall*, *F1-score* para cada uma das classes ("Sem Êxito" e "Êxito"), assim como acurácia, F1 macro, *F1 weighted* e AUC para ambas as classes. Os resultados destacados em negrito representam os melhores desempenhos observados entre os algoritmos avaliados. Neste estudo, o critério principal adotado para a definição do melhor cenário foi *F1 macro*, *F1 weighted*, AUC e acurácia para ambas as classes. Essa escolha foi motivada pela necessidade de avaliar não apenas o desempenho global do modelo, mas também sua capacidade de manter um comportamento equilibrado entre as classes. Dessa forma, a seleção do melhor cenário buscou privilegiar modelos com desempenho consistente e equilibrado, evitando decisões baseadas exclusivamente na acurácia global.

De modo geral, observou-se que os cenários com balanceamento não elevaram significativamente a acurácia global dos modelos, que permaneceu em valores próximos aos obtidos sem balanceamento. Entretanto, houve ganhos consistentes no *F1-score* da classe minoritária (Êxito), indicando maior capacidade dos modelos em identificar corretamente os estudantes com êxito acadêmico.

Considerando que o cenário é desbalanceado, o melhor desempenho global é avaliado quanto ao algoritmo que obtém os melhores resultados de *F1 macro*, *F1 weighted*, AUC e acurácia. Portanto, o modelo *Random Forest* no cenário de balanceamento dos dados com SMOTE apresentou o melhor desempenho global entre os algoritmos avaliados. Destaca-se que os resultados desse cenário são muito próximos (em algumas métricas, com valores similares) ao cenário de balanceamento com *Random Oversampler*. Nesse cenário, o modelo alcançou acurácia de aproximadamente 70%, F1-score de 0,76 para a classe "Sem Êxito" e de 0,60 para a classe "Êxito", além de AUC de cerca de 0,75. Em termos de precisão e *recall*, foram obtidos valores de 0,76 e 0,75 para a classe "Sem Êxito", e de 0,59 e 0,60 para a classe "Êxito", respectivamente, indicando um desempenho mais equilibrado entre a classe majoritária e a classe minoritária quando comparado aos demais cenários.

A matriz de confusão correspondente indica que, de um total de 463 estudantes pertencentes à classe "Êxito", o modelo classificou corretamente 279, enquanto 184 foram incorretamente rotulados como "Sem Êxito". Para a classe majoritária, 585 estudantes foram corretamente classificados, de um total de 779. Esses resultados evidenciam que o uso do SMOTE contribuiu para uma melhor distribuição do desempenho entre as classes, reduzindo vieses em favor da classe majoritária e resultando em um modelo mais adequado ao contexto do problema de evasão, no qual a correta identificação da classe minoritária é particularmente relevante.

O algoritmo *XGBoost* também apresentou desempenho competitivo, especialmente no cenário com SMOTE, no qual atingiu acurácia de 69%, *F1-score* de 0,76 para a classe "Sem Êxito" e de 0,59% para a classe "Êxito", além de *F1 macro* de 0,67, *F1 weighted* de 0,68 e AUC de 0,74.

A Árvore de Decisão (DT) apresentou seus melhores resultados no cenário com *SMOTE*, alcançando *F1-score* de 0,73 para a classe "Sem Êxito"e de 0,55 para a classe "Êxito", além de Acurácia de 0,66 para ambas as classes, *F1 macro* de 0,64, *F1 weighted* de 0,66 e AUC de 0,68. Embora esse desempenho seja inferior ao obtido pelos modelos baseados em ensembles, observa-se uma melhoria em relação ao cenário sem balanceamento, especialmente nas métricas de *recall* da classe minoritária.

Os algoritmos KNN e MLP apresentaram desempenho inferior de forma consistente em todos os cenários avaliados. Mesmo com a aplicação de técnicas de balanceamento, os valores de *F1-score* da classe "Êxito" permaneceram abaixo de 55%, e as AUCs situaram-se em torno de 63% a 65%. Esses resultados indicam limitações desses métodos na captura de padrões mais complexos presentes nos dados educacionais analisados, quando comparados aos modelos baseados em árvores e *ensembles*.

Os resultados de desempenho obtidos neste estudo mostram-se consistentes com os resultados da literatura apresentada. Assim como observado por Primão [2022], modelo baseado em *ensembles*, como *XGBoost* apresentou desempenho superior em relação a abordagens mais simples, como DT e MLP, na capacidade de capturar padrões em evasão escolar. De modo semelhante ao trabalho de Araújo [2025], o *Random Forest* demonstrou bom equilíbrio entre métricas, destacando-se particularmente na acurácia e *recall*.

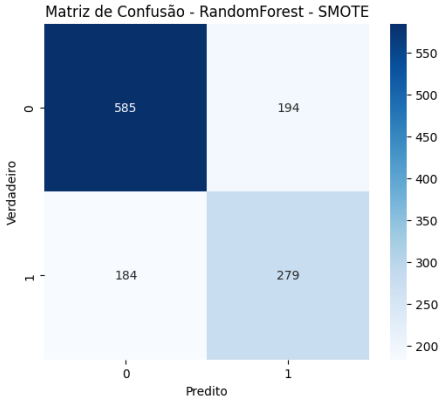

Figura 2. Matriz de confusão do RF para o cenário *SMOTE*.

Tabela 2. Resultados de desempenho dos modelos avaliados em diferentes cenários de balanceamento.

Modelo	Cenário	Sem Êxito			Êxito			Acurácia	*F1 Macro*	*F1 Weighted*	AUC
		Precisão	*Recall*	F1	Precisão	*Recall*	F1				
DT	Sem bal.	0,70	0,75	0,73	0,53	0,47	0,50	0,64	0,61	0,64	0,62
	SMOTE	0,73	0,73	0,73	0,55	0,55	0,55	0,66	0,64	0,66	0,68
	Random Oversampler	0,77	0,56	0,65	0,49	0,73	0,59	0,62	0,62	0,63	0,70
RF	Sem bal.	0,73	0,82	0,77	0,61	0,48	0,54	0,69	0,65	0,68	0,73
	SMOTE	0,76	0,75	0,76	0,59	0,60	0,60	**0,70**	**0,68**	**0,70**	**0,75**
	Random Oversampler	0,78	0,71	0,75	0,58	0,66	0,62	0,69	**0,68**	**0,70**	**0,75**
XGBoost	Sem bal.	0,73	0,78	0,75	0,58	0,51	0,54	0,68	0,65	0,68	0,71
	SMOTE	0,75	0,76	0,76	0,59	0,59	0,59	0,69	0,67	0,69	0,74
	Random Oversampler	0,78	0,68	0,72	0,55	0,68	0,61	0,68	0,67	0,68	0,74
KNN	Sem bal.	0,70	0,69	0,70	0,49	0,50	0,50	0,62	0,60	0,62	0,63
	SMOTE	0,71	0,63	0,67	0,48	0,57	0,52	0,61	0,59	0,61	0,64
	Random Oversampler	0,72	0,61	0,66	0,48	0,59	0,53	0,60	0,59	0,61	0,64
MLP	Sem bal.	0,68	0,83	0,75	0,54	0,35	0,42	0,65	0,58	0,63	0,63
	SMOTE	0,71	0,64	0,67	0,48	0,56	0,51	0,61	0,59	0,61	0,63
	Random Oversampler	0,73	0,62	0,67	0,49	0,62	0,55	0,62	0,61	0,63	0,65

5.3 Análise das variáveis mais influentes na predição

A análise da importância das variáveis fornece uma visão clara sobre os fatores que mais influenciaram as predições dos modelos com melhor desempenho: *XGBoost* e RF.

No *XGBoost*, a variável mais influente foi Renda Familiar, seguida por Período letivo de ingresso, Nível de ensino, Estado civil e Sexo. Esse conjunto indica que fatores socioeconômicos associados à trajetória acadêmica no momento do ingresso exercem papel central na modelagem, contribuindo significativamente para distinguir estudantes com maior predisposição ao êxito daqueles mais propensos à evasão.

No RF, a variável de maior peso também foi Renda familiar, seguida por Descrição do curso, Matriz curricular, Idade de ingresso e Renda familiar per capita. Embora apresentem diferentes pesos entre si, ambos os modelos convergem ao ressaltar a relevância de fatores socioeconômicos e acadêmicos estruturais na previsão do êxito ou evasão estudantil, reforçando a influência combinada entre aspectos acadêmicos e socioeconômicos na determinação dos padrões de evasão.

As Figuras 3 e 4 apresentam o *ranking* das variáveis mais influentes nos modelos *XGBoost* e RF, respectivamente. A comparação entre os dois algoritmos revela convergência quanto à importância da renda familiar, evidenciando o peso do contexto socioeconômico na permanência estudantil. As divergências observadas nos demais atributos refletem diferenças na forma como cada modelo captura relações entre características individuais, estruturais e institucionais.

Embora a DT tenha apresentado desempenho inferior em relação aos modelos de *ensemble*, sua natureza interpretável permite uma análise qualitativa complementar. A seguir é apresentado um recorte, de profundidade três, da árvore gerada no cenário com balanceamento com SMOTE (Figura 5). A árvore evidencia que Renda Familiar constitui o critério inicial de decisão do modelo, funcionando como principal eixo de separação entre os estudantes. Nos níveis subsequentes, variáveis como Sexo, Descrição do Curso e Período Letivo de Ingresso aparecem de forma recorrente, indicando sua relevância na diferenciação entre perfis de permanência e evasão.

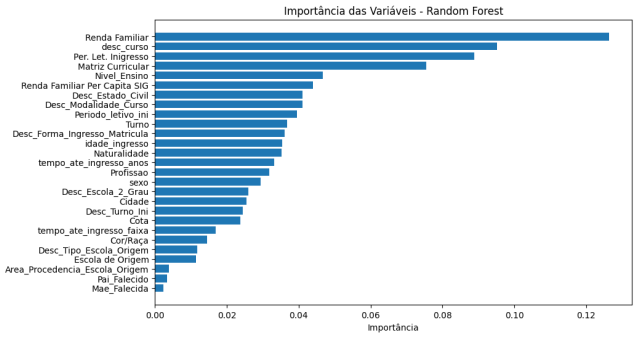

Figura 3. Variáveis mais importantes para a predição da evasão escolar, considerando o RF.

À medida que a árvore se aprofunda, atributos como Cota, Modalidade do Curso, Matriz Curricular e Renda Fa-

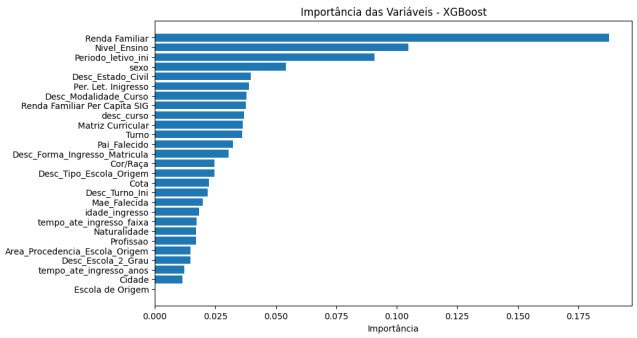

Figura 4. Variáveis mais importantes para a predição da evasão escolar, considerando o *XGBoost*.

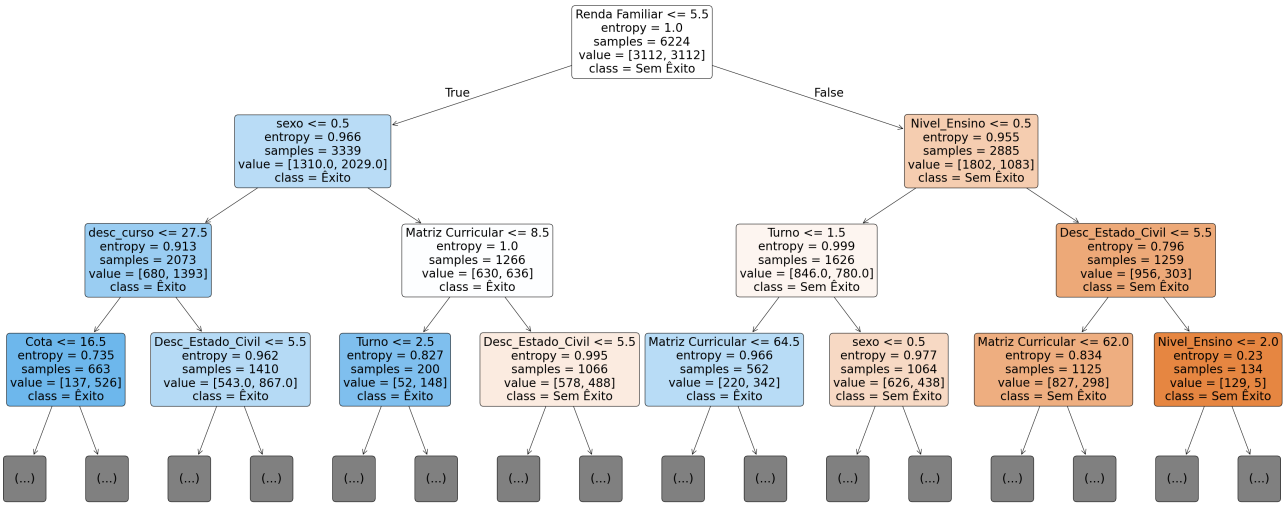

Figura 5. Recorte da árvore gerada pelo algoritmo DT no cenário de balanceamento com SMOTE.

miliar Per Capita surgem como refinamentos importantes, contribuindo para a formação de regiões mais homogêneas no espaço de decisão. A presença desses atributos nos níveis superiores da árvore reforça a centralidade de fatores socioeconômicos e acadêmicos no processo decisório do classificador.

De forma geral, o destaque das variáveis socioeconômicas, em especial da renda familiar, sugere que a evasão não pode ser compreendida apenas como resultado de fatores estritamente acadêmicos, mas também como reflexo de condições materiais que afetam a permanência estudantil. A influência de variáveis como renda, renda familiar per capita, idade de ingresso, estado civil e sexo aponta para diferentes níveis de vulnerabilidade, que podem dificultar a continuidade no curso. Sob a perspectiva institucional, esses resultados reforçam a importância de políticas de assistência estudantil, apoio à permanência e acompanhamento mais direcionado já nos períodos iniciais da trajetória acadêmica. Também indicam a necessidade de ações preventivas voltadas a grupos com maior exposição a risco, permitindo que a instituição utilize essas evidências para orientar intervenções mais estratégicas, integrando suporte financeiro, pedagógico e psicossocial.

6 Considerações finais

Este trabalho investigou a predição da evasão estudantil por meio de técnicas de aprendizado de máquina aplicadas a dados acadêmicos e sociodemográficos de estudantes ingressantes no IFCE *campus* Tianguá. A abordagem adotada contemplou desde a preparação dos dados, com tratamento, engenharia de atributos, balanceamento de dados até a avaliação de diferentes algoritmos de aprendizagem de máquina supervisionada.

Os resultados evidenciaram que modelos de *ensemble*, especialmente o RF e o *XGBoost*, apresentaram maior capacidade de discriminação entre as classes, confirmando o potencial dessas técnicas no apoio à compreensão de fatores associados ao risco de evasão. Considerando um cenário real que os dados são desbalanceados, as estratégias de balanceamento de dados aumentam levemente o desempenho dos modelos.

Além disso, a análise de importância das variáveis reforçou a presença de padrões socioeconômicos e acadêmicos que se relacionam à permanência e ao desempenho dos estudantes, identificando que variáveis como a renda familiar, o sexo e o estado civil têm alto nível de importância para a predição da evasão escolar.

Como limitações deste estudo, destaca-se, primeiramente, a utilização de dados provenientes de um único *campus*, o que pode restringir a representatividade dos resultados frente à diversidade de contextos institucionais. Além disso, não foram consideradas variáveis comportamentais ou informações de desempenho acadêmico ao longo do curso, o que pode limitar a compreensão mais dinâmica dos fatores associados à evasão. Por fim, tais aspectos impõem restrições à generalização dos resultados, que devem ser interpretados com cautela ao serem extrapolados para outras instituições ou realidades educacionais distintas.

Para trabalhos futuros, sugere-se incorporar novos dados institucionais à medida que se tornem disponíveis, incluir variáveis adicionais relacionadas à vida acadêmica e ao acompanhamento pedagógico, assim como aprofundar a avaliação da importância das variáveis na evasão escolar, de técnicas de IA Explicável e de outros modelos de aprendizagem de máquina. Ademais, na análise da importância das variáveis, sugere-se adotar uma abordagem individualizada, identificando, para cada estudante, quais fatores mais influenciam seu risco de evasão, tais como renda, distância da residência ou outros aspectos relevantes, possibilitando intervenções mais direcionadas e eficazes. Também se destaca como possibilidade a criação e integração de um sistema de monitoramento automatizado dentro da instituição, permitindo o uso prático dos modelos e a evolução contínua das análises.

Declarações complementares

Contribuições dos autores

EA contribuiu para a concepção deste estudo e realização dos experimentos. FRVS contribuiu na orientação do trabalho e ACO na coorientação do trabalho. Todos os autores leram e aprovaram o manuscrito final.

Conflitos de interesse

Os autores declaram que não têm nenhum conflito de interesses.

Referências

Amorim, C. C., Branco, U. V. C., and Dias Júnior, J. J. L. (2025). Evasão e retenção no ensino superior: ampliando a compreensão teórico-metodológica. *Revista da FAEEBA – Educação e Contemporaneidade*, 34(79). DOI: 10.21879/faeeba2358-0194.2025.v34.n79.p77-96.

Arantes, A. R., Rodrigues, L. B., Kagimura, R., Cardoso, B. G. d. S., and Junqueira, M. P. (2021). Evasão e retenção no ensino superior: abordagem baseada em taxas quantitativas. *Revista Contemporânea de Educação*, 16(36):4–21. DOI: 10.20500/rce.v16i36.42914.

Araújo, G. B. (2025). Análise e discussão da predição de um caso de evasão universitária: Uma abordagem baseada em data mining e machine learning. Master's thesis, Universidade Federal do Ceará, Fortaleza. Disponível em: `https://repositorio.ufc.br/bitstream/riufc/81274/1/2025_tcc_gbaraujo.pdf`. Acesso em: 20 dez. 2025.

Baggi, C. A. d. S. and Lopes, D. A. (2011). Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. *Avaliação: Revista da Avaliação da Educação Superior*, 16(2). DOI: 10.1590/S1414-40772011000200007.

Bastos, N., Gomes, L., Silveira, R., and Oliveira, C. (2025). Mapeamento sistemático da mineração de dados educacionais no combate à evasão escolar no brasil. In *Anais do IX Congresso sobre Tecnologias na Educação*, pages 177–187, Porto Alegre, RS, Brasil. SBC. DOI: 10.5753/ctrle.2025.12545.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. DOI: 10.1023/A:1010933404324.

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357. DOI: 10.1613/jair.953.

Chen, T. and Guestrin, C. (2016). Xgboost: A scalable tree boosting system. page 785–794. DOI: 10.1145/2939672.2939785.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning*. MIT Press, Cambridge.

Géron, A. (2019). *Mãos à obra: Aprendizado de máquina com Scikit-learn & TensorFlow – conceitos, ferramentas e técnicas para a construção de sistemas inteligentes*. Alta Books, Rio de Janeiro, 1 edition. Tradução de Rafael Contatori.

Han, J., Kamber, M., and Pei, J. (2012). *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Francisco, 3 edition.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning*. Springer, Nova York, 2 edition.

He, H. and Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284. DOI: 10.1109/TKDE.2008.239.

IFCE (2017). Plano estratégico de permanência e Êxito dos estudantes do ifce. Fortaleza: IFCE. Disponível em: `https://portal.ifce.edu.br/documents/3515/Plano_Estrat%C3%A9gico_Institucional_para_Perman%C3%AAncia_e_%C3%8Axito_dos_Estudantes_do_IFCE.pdf`. Acesso em: 6 dez. 2025.

INEP (2024). Apresentação do censo da educação superior 2024. Disponível em: `https://download.inep.gov.br/educacao_superior/censo_superior/documentos/2024/apresentacao_censo_da_educacao_superior_2024.pdf`. Acesso em: 8 dez. 2025.

James, G., Witten, D., Hastie, T., Tibshirani, R., and Taylor, J. (2023). *An Introduction to Statistical Learning: with Applications in Python*. Springer, New York.

Lemaître, G., Nogueira, F., and Aridas, C. K. (2016). Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. *The Journal of Machine Learning Research*. DOI: https://dl.acm.org/doi/10.5555/3122009.3122026.

Linhares, J. F., Veras, E. d. A., Oliveira, R. d. S., Silveira, R., and Oliveira, C. T. d. (2025). Desenvolvimento de uma ferramenta para mapear os índices acadêmicos de uma instituição de ensino. *Anais do 17º Encontro Unificado de Computação do Piauí (ENUCOMPI)*, pages 29–38. DOI: 10.5753/enucompi.2025.9582.

Malerba, A. (2024). Previsão de evasão universitária com aprendizado de máquina. Master's thesis, Universidade Federal de Itajubá, Itajubá. Dissertação de Mestrado em Ciência e Tecnologia da Computação. Disponível em: `https://repositorio.unifei.edu.br/jspui/handle/123456789/4072`. Acesso em: 20 dez 2025.

Murphy, K. P. (2022). *Probabilistic Machine Learning: An Introduction*. MIT Press, Cambridge.

Primão, A. P. (2022). Uso de algoritmos de machine learning para prever a evasão escolar no instituto federal de santa catarina (ifsc). Master's thesis, Universidade Federal de Santa Catarina, Florianópolis. Dissertação de Mestrado em Administração. Disponível em: `https://repositorio.ufsc.br/handle/123456789/238320`. Acesso em: 20 dez. 2025. Universitária.

Saccaro, A., França, M. T. A., and Jacinto, P. A. (2024). 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*. DOI: 10.1590/0101-41614925amp.

SEMESP (2025). Mapa do ensino superior no brasil: 15ª edição. Disponível em: `https://www.semesp.org.br/wp-content/uploads/2025/02/mapa-do-ensino-superior-no-brasil-2025.pdf`. Acesso em: 6 dez. 2025.

Silva, P. T. d. F. e. and Sampaio, L. M. B. (2022). Políticas de permanência estudantil na educação superior: reflexões de uma revisão da literatura para o contexto brasileiro. *Revista de Administração Pública*, 56(5):603–631. DOI: 10.1590/0034-761220220034.