

PREVISÃO DA EVASÃO ESCOLAR ATRAVÉS DA ANÁLISE DE DADOS E APRENDIZAGEM DE MÁQUINA: Um estudo de caso

PREDICTING SCHOOL EVASION THROUGH DATA
ANALYSIS AND MACHINE LEARNING: A case study

Dyego Barbosa

dhbb@discente.ifpe.edu.br

Luciano Cabral

luciano.cabral@jaboatao.ifpe.edu.br

Resumo. A evasão escolar é um desafio global significativo, que compromete o desenvolvimento educacional, econômico e social de diversas comunidades. Este trabalho propõe o uso de métodos automáticos baseados em aprendizado de máquina para prever e monitorar alunos em risco de abandono escolar, com foco nos estudantes do Instituto Federal de Pernambuco - Campus Jaboatão dos Guararapes. O risco avaliado refere-se à possibilidade de os alunos interromperem seus estudos antes da conclusão, devido a fatores acadêmicos e sociodemográficos identificados nos dados coletados. Foram testados diversos algoritmos, e o Random Forest apresentou o melhor desempenho, superando os demais modelos com base nas métricas utilizadas. Os resultados mostram que é viável usar modelos preditivos para apoiar decisões e desenvolver políticas para reduzir a evasão escolar.

Palavras-chave: Evasão escolar; Análise inteligente; Aprendizado de máquina; Algoritmo Random Forest; Modelos preditivos;

Abstract. School dropout is a significant global challenge, which compromises the educational, economic and social development of different communities. This work proposes the use of automatic methods based on machine learning to predict and monitor students at risk of dropping out of school, focusing on students at the Instituto Federal de Pernambuco - Campus Jaboatão dos Guararapes. The assessed risk refers to the possibility of classes interrupting their studies before completion, due to academic and sociodemographic factors identified in the data collected. Several algorithms were tested, and Random Forest presented the best performance, surpassing the other models based on the analyzes used. The results show that it is feasible to use predictive models to support decisions and develop policies to reduce school dropout rates.

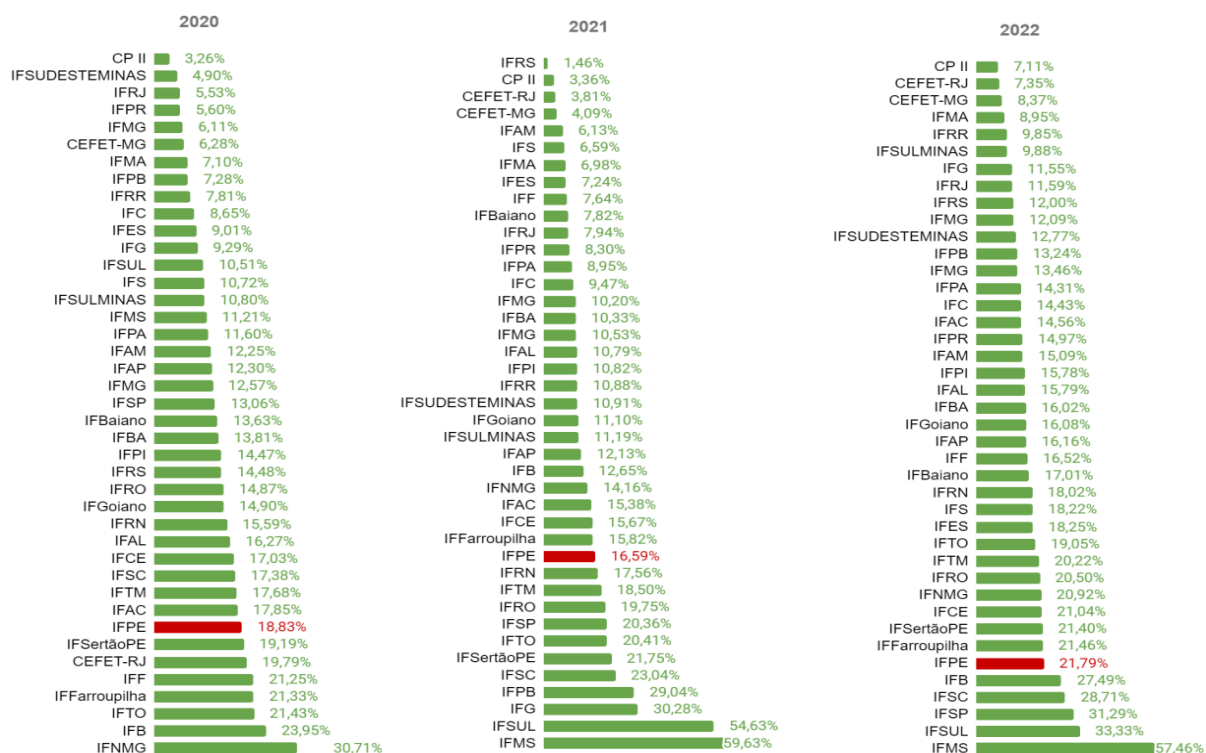
Keywords: School dropout; Intelligent analysis; Machine learning; Random Forest Algorithm; Predictive models;

1 INTRODUÇÃO

A Evasão escolar pode ser definida como o ato de fuga ou abandono da escola em razão da realização de outras atividades, conforme descrito por Riffel (2010). Esse fenômeno é uma questão complexa que vai além do contexto específico de uma única instituição e se manifesta em universidades, escolas e instituições de ensino em todo o mundo. A evasão impacta diretamente a trajetória educacional dos alunos, comprometendo seu desenvolvimento pessoal e profissional, e representa um desafio significativo para os sistemas educacionais globais na busca por soluções que possam minimizar suas consequências.

Trazendo consequências diretas para o desenvolvimento econômico e social. Este fenômeno compromete não apenas o futuro dos estudantes, que abandonam suas trajetórias de formação, mas também representa uma ameaça ao retorno do investimento em políticas públicas voltadas para a educação e à manutenção da qualidade do ensino. No Brasil, esse cenário é alarmante, especialmente nas instituições federais de ensino. A Nota de Auditoria (IFPE, 2024) destaca que o Instituto Federal de Pernambuco (IFPE) experimentou um aumento notável na taxa de evasão, que atingiu 21,79%, resultando em sua queda para a 36ª posição no ranking das instituições da Rede Federal de Educação Profissional, como mostra no Gráfico 1. Esses dados reforçam a necessidade urgente de estratégias eficazes que possam identificar precocemente os estudantes em risco e implementar intervenções que favoreçam sua permanência e sucesso acadêmico.

Gráfico 1 – Classificação das instituições que compõem Rede Federal de Educação Profissional, Científica e Tecnológica pelo critério da taxa de evasão 2020-2022



Fonte: Plataforma Nilo Peçanha (2022)

A necessidade de estudos que abordem estratégias eficazes para combater as altas taxas de evasão escolar é evidente, especialmente em instituições de ensino que enfrentam desafios crescentes nesse aspecto. Uma das abordagens mais promissoras para enfrentar essa questão é o uso de técnicas de aprendizado de máquina, que permitem a análise de grandes volumes de dados acadêmicos para prever comportamentos e identificar estudantes em risco de abandono. O aprendizado de máquina pode ser definido como uma forma de inteligência artificial que permite aos sistemas aprenderem e melhorarem automaticamente a partir de experiências passadas sem serem explicitamente programados. Conforme descrito pela *International Business Machines Corporation*.

O aprendizado de máquina pode usar conjuntos de dados rotulados, também conhecidos como aprendizado supervisionado, para informar seu algoritmo, mas não exige necessariamente um conjunto de dados rotulados. Essa tecnologia possibilita a criação de modelos preditivos que apoiam a tomada de decisões e a implementação de políticas voltadas para a redução da evasão.

Diante desse cenário desafiador, este artigo busca contribuir com a sociedade, apresentando uma solução potencial para minimizar o problema da evasão escolar. A pesquisa foi conduzida no IFPE com base em um modelo de aprendizado de máquina, treinado a partir de dados acadêmicos e sociodemográficos, visando identificar parâmetros essenciais para a previsão da evasão escolar.

2 REVISÃO DA LITERATURA

No presente tópico, serão abordados conceitos fundamentais que sustentam o estudo realizado, com destaque para a evasão escolar, a mineração de dados e o aprendizado de máquina. Esses temas são explorados para contextualizar e fundamentar a aplicação de técnicas analíticas para a previsão da evasão.

2.1 Evasão escolar

A evasão escolar é um fenômeno complexo que afeta significativamente o ambiente educacional e as perspectivas de desenvolvimento individual e social. Segundo Johann.

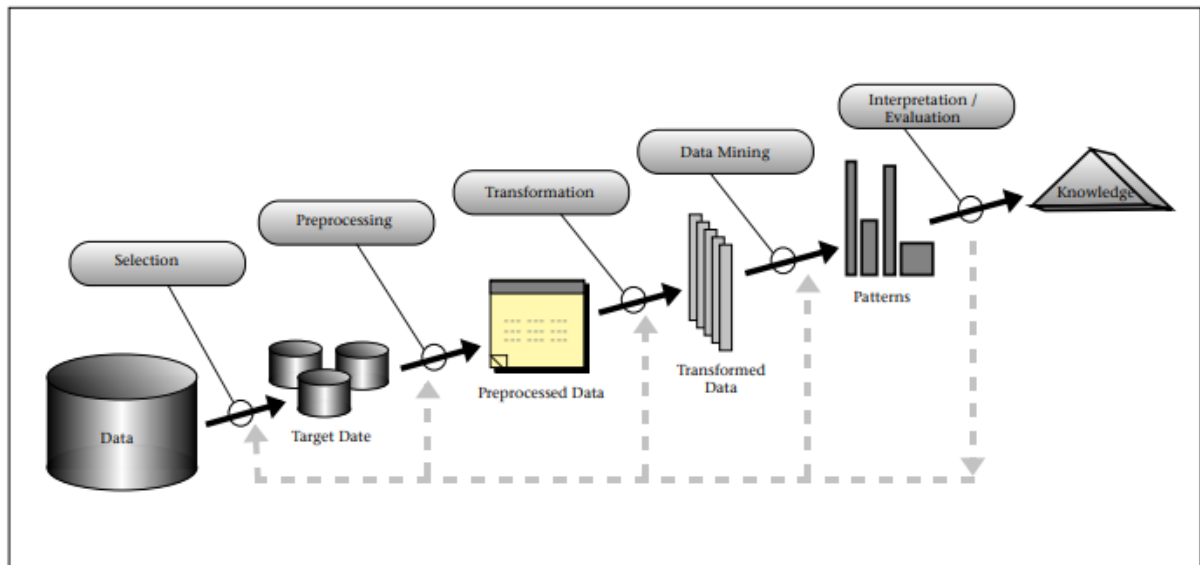
A evasão é um fenômeno que se caracteriza pelo abandono do curso, resultando no rompimento do vínculo jurídico estabelecido com a instituição de ensino. Esse rompimento ocorre quando o aluno não renova sua matrícula ou manifesta a intenção de continuar no curso. Nesse contexto, a evasão é compreendida como um abandono definitivo, já que a ausência de renovação da matrícula formaliza a interrupção do vínculo entre o aluno e a escola.

Nesse contexto, a evasão é interpretada como um abandono definitivo, caracterizando uma decisão muitas vezes irreversível por parte do aluno, que deixa de renovar sua matrícula e, assim, interrompe formalmente o vínculo estabelecido com a instituição de ensino.

2.2 Mineração de dados

A mineração de dados é uma área essencial no contexto da análise educacional, sendo parte do processo mais amplo conhecido como *Knowledge Discovery in Databases* (KDD). Conforme descrito por Fayyad (1996), o KDD é composto por diversas etapas interligadas, que incluem desde a seleção e preparação dos dados até a interpretação dos resultados. A Figura 1 abaixo ilustra essas etapas de forma detalhada, evidenciando o fluxo sistemático necessário para transformar grandes volumes de dados em conhecimento útil.

Figura 1 - Etapas do Processo KDD



Fonte: Fayyad (1996)

Conforme Silva (2022), todo o processo KDD ocorre de forma iterativa e interativa, uma vez que cada etapa está conectada às demais, envolvendo decisões e tarefas que precisam ser realizadas pelos usuários. Caso uma das etapas não alcance os resultados esperados, é possível retroceder e ajustar etapas anteriores, como a preparação dos dados, para garantir a eficácia e a consistência do processo na totalidade. Essa flexibilidade é fundamental para assegurar a qualidade dos modelos gerados e a confiabilidade das análises.

2.3 Aprendizado de máquina

O aprendizado de máquina, também conhecido pelo termo em inglês *machine learning* (ML), refere-se a um campo da inteligência artificial que se concentra no desenvolvimento de algoritmos capazes de aprender e realizar previsões a partir de dados. Segundo Kohavi.

Diferentes algoritmos de aprendizado de máquina podem ser aplicados a uma mesma base de dados, alguns destes podem apresentar melhores resultados que outros. Entretanto, não existe garantia de que um determinado algoritmo seja melhor ou que terá um melhor desempenho, sendo necessários estudos experimentais para descobrir qual se adapta melhor ao problema analisado.

Essa perspectiva ressalta a importância de experimentar diferentes abordagens para selecionar o modelo mais eficaz para o problema em questão, o que é especialmente relevante no contexto educacional, onde a precisão das previsões pode impactar diretamente na elaboração de estratégias de intervenção.

3 TRABALHOS RELACIONADOS

A evasão escolar é um tema amplamente explorado em diversas pesquisas, dada sua relevância e os impactos que gera no sistema educacional. Segundo Donna (2023), algumas das principais causas da evasão estudantil incluem a localização da instituição, problemas estruturais no curso, ausência de laços afetivos com a instituição, falta de integração entre a universidade e a educação básica, metodologia de ensino adotada, escolhas inadequadas de carreira, razões econômicas e a falta de informações sobre o curso antes do ingresso. Esses aspectos evidenciam a multiplicidade de fatores que influenciam na decisão de abandono escolar e demonstram dados que podem ser levados em consideração neste trabalho.

Outro trabalho relevante que explora a relação entre os dados coletados e a evasão escolar é o estudo de Oliveira (2024). Nesse estudo, os objetivos foram alcançados por meio da aplicação de Mineração de Dados Educacionais (MDE), um processo que utiliza métodos de mineração de dados e aprendizado de máquina no contexto educacional. Esse enfoque evidencia que, para compreender e mitigar a evasão escolar, não basta apenas coletar dados; é fundamental aplicar técnicas apropriadas que permitam ao algoritmo analisar padrões e realizar previsões precisas. Essa abordagem reforça a importância do uso de métodos avançados para transformar dados brutos em informações úteis para a tomada de decisões estratégicas no ambiente educacional.

Colpo (2021) estudaram diferentes formas de representar o comportamento ligado à evasão escolar na criação de modelos preditivos, com foco na educação universitária. Durante o trabalho, eles trataram detalhadamente os dados e testaram vários algoritmos. Eles descobriram que analisar o comportamento dos alunos ao longo de todos os semestres, de forma acumulativa e progressiva, ajuda bastante a melhorar os modelos. Dos algoritmos testados, a *Decision Tree* e o *Random Forest* se destacaram, com o *Random Forest* sendo o mais preciso, alcançando uma precisão de 80,1%.

Outro ponto importante com relação aos modelos preditivos descrito por Rodrigues (2022) destaca a seguinte estratégia, a importância de reconstruir cada modelo do zero usando a melhor combinação de hiperparâmetros. Após isso, os modelos são treinados com 80% dos dados disponíveis e testados com os 20% restantes. Essa abordagem é essencial para evitar problemas como *overfitting* e *underfitting*, comuns no desenvolvimento de modelos de aprendizado de máquina. O *overfitting* acontece quando o modelo aprende em excesso os dados de treinamento, prejudicando sua capacidade de generalizar. O *underfitting*, por outro lado, mostra um ajuste ruim, levando a um desempenho ruim tanto nos dados de treino quanto nos de teste. A estratégia descrita busca um equilíbrio, garantindo resultados confiáveis e consistentes.

Um estudo relacionado, realizado no Instituto Federal de Minas Gerais (IFMG), onde de acordo com Melo (2022), que utilizou o modelo baseado no algoritmo *Random Forest* para prever a evasão escolar entre os estudantes atualmente

matriculados no Campus Bambuí. A análise foi aplicada em um total de 744 instâncias de estudantes, dos quais 382 (51,3%) foram classificados como evadidos e 362 (48,7%) como formados. O modelo apresentou um percentual médio de confiabilidade de 87%, com um desvio padrão de 4%, utilizando os cálculos realizados pela biblioteca *PyCaret*. Esses resultados reforçam a eficácia da aplicação de algoritmos de aprendizado de máquina na identificação de padrões de evasão, contribuindo para o desenvolvimento de estratégias preventivas mais assertivas.

Outro projeto relevante no campo da análise de aprendizado foi o de Beltrán (2019), que desenvolveu um sistema voltado para monitorar e sinalizar alunos com risco de evasão. A ferramenta permite listar os alunos classificados como “situação normal” ou “risco de evasão”, além de detalhar seus dados e compará-los ao perfil ideal para o curso. O sistema também oferecia funcionalidades voltadas à gestão acadêmica, como a possibilidade de agendar reuniões pedagógicas, sugerir materiais de apoio, indicar suporte de tutoria e notificar professores para um acompanhamento mais próximo. Essas funcionalidades facilitaram a identificação de riscos e colaboraram na implementação de medidas preventivas, destacando a importância de ferramentas práticas na gestão educacional.

Os trabalhos relacionados apresentados neste tópico destacam a relevância e a complexidade do problema da evasão escolar, evidenciando a multiplicidade de fatores que levam ao abandono educacional. Os estudos abordados demonstram a importância da coleta, análise e aplicação de dados educacionais por meio de técnicas de aprendizado de máquina. Essas pesquisas reforçam a necessidade de estratégias que combinem metodologias robustas e tecnologias avançadas para prever e mitigar a evasão escolar.

4 PROBLEMA DA PESQUISA

A evasão escolar persiste como um desafio complexo e impactante no âmbito educacional do Instituto Federal de Pernambuco - Campus Jaboatão dos Guararapes e também em outras instituições de ensino, pois envolve uma intersecção de fatores multifacetados. A complexidade emerge da interação entre elementos socioeconômicos, psicológicos e acadêmicos, tornando difícil a atribuição de causas específicas. Além disso, a evasão impacta não apenas os alunos individualmente, mas reverbera no sistema educacional como um todo, comprometendo metas de inclusão e equidade. As implicações socioeconômicas, como falta de recursos financeiros e acesso limitado a oportunidades, somam-se às questões psicológicas, incluindo desmotivação e falta de apoio emocional. O desafio reside na necessidade de abordagens holísticas que compreendam a complexidade desses fatores para desenvolver estratégias eficazes de prevenção e intervenção.

Assim, o presente estudo visa analisar a problemática da evasão escolar, utilizando os dados fornecidos e recorrendo a algoritmos de inteligência artificial. A proposta é realizar uma análise de padrões nas evasões, buscando antecipar com sucesso casos de evasão antes de sua concretização.

5 METODOLOGIA

Para a realização deste estudo, utilizou-se uma abordagem quantitativa baseada na análise de dados e aplicação de técnicas de aprendizado de máquina. A base de

dados foi obtida do sistema de gestão acadêmica Q-Acadêmico, contendo informações de alunos matriculados no período entre 2014 e 2021 no campus do Instituto Federal de Pernambuco (IFPE). Após uma análise preliminar, foram selecionados 14 atributos relevantes, tais como sexo, estado civil, tipo de cota e turno, para construção e treinamento dos modelos preditivos.

O processo seguiu as etapas do *Knowledge Discovery in Databases* (KDD), incluindo pré-processamento dos dados, tratamento de valores ausentes, normalização, codificação de variáveis categóricas e divisão da base entre conjuntos de treino e teste. Os dados foram classificados em duas categorias: estudantes regulares e evadidos, para treinar os modelos preditivos.

6 DESENVOLVIMENTO

O desenvolvimento deste estudo incluiu várias etapas estruturadas para alcançar os objetivos propostos. Primeiramente, realizou-se a coleta dos dados sócio-demográficos e a situação acadêmica dos alunos a cada período, utilizando o sistema acadêmico institucional. Em seguida, aplicaram-se técnicas de pré-processamento, incluindo limpeza e organização dos dados, bem como uma análise exploratória para identificar padrões e tendências. Na etapa seguinte, foram selecionados algoritmos de aprendizado de máquina considerados mais adequados ao problema, com base em sua capacidade de lidar com os dados disponíveis.

Os modelos foram treinados e avaliados utilizando métricas clássicas como Acurácia, F1-Score, *Recall*, Kappa e Precisão, que possuem ampla aceitação em trabalhos similares. Por fim, os resultados dos modelos foram comparados para identificar o desempenho mais eficaz e os fatores determinantes para a previsão da evasão. Na Tabela 1, apresenta-se uma descrição detalhada de cada métrica utilizada no processo de avaliação dos modelos.

Tabela 1 – Descrição das métricas de avaliação de desempenho

Métrica	O que mede	Como interpretar
<i>Accuracy</i> (Acurácia)	Refere-se à precisão geral do modelo, ou seja, qual a proporção de previsões corretas em relação ao total.	Quanto mais próximo de 1, melhor. Indica a eficácia geral do modelo em acertar as previsões.
F1 Score (Pontuação F1)	Uma média harmônica entre precisão e recall.	É útil quando há um desequilíbrio nas classes. Quanto mais próximo de 1, melhor equilíbrio entre precisão e recall.
<i>Recall</i> (Revocação)	A proporção de casos positivos reais que foram corretamente identificados pelo modelo.	Quanto mais próximo de 1, melhor o modelo está em capturar todos os casos positivos.
Cohen's Kappa (Kappa de Cohen)	Avalia a concordância entre as previsões do modelo e os resultados reais, levando em conta a concordância esperada ao acaso.	Quanto mais próximo de 1, melhor a concordância do modelo, considerando a concordância esperada ao acaso.
<i>Precision</i> (Precisão)	A proporção de casos positivos previstos pelo modelo que são verdadeiramente	Quanto mais próximo de 1, menor é a taxa de falsos positivos.

Com base nessas informações, foram criados modelos de inteligência artificial utilizando os algoritmos *Decision Tree Classifier*, *Random Forest* e *XGBoost*, com o propósito de estimar a probabilidade de evasão dos alunos. A seleção desses modelos considerou critérios como desempenho, facilidade de interpretação e habilidade para trabalhar com diferentes tipos de dados.

6.1 Pré-processamento da base de dados

Foi reunida informações abrangentes sobre os alunos matriculados nos cursos do IFPE, totalizando mais de 100 atributos distintos. Após uma análise detalhada dos campos disponíveis, foram selecionadas 14 colunas consideradas essenciais para as análises e o treinamento dos modelos. O processo de seleção e tratamento, descrito previamente, resultou na definição dos atributos apresentados na Tabela 2, vale ressaltar que os dados dos alunos não foram divulgados devido às questões LGPD (Lei Geral de Proteção de Dados Pessoais).

Tabela 2 - Nomenclatura das colunas selecionadas.

Coluna	Descrição
Coeficiente_Rendimento	Coeficiente de rendimento
Desc_Tipo_Escola_Origem	Tipo de escola de origem
Percentual_Frequencia	Percentual de frequência
Nivel_Ensino	Nível de ensino
Desc_Cor	Cor ou raça declarada
Desc_Modalidade_Curso	Modalidade do curso (concomitante, subsequente ou outros)
Desc_Cota	Tipo de cota que o aluno se beneficiou para ingressar
Desc_Estado_Civil	Situação conjugal
Sexo	Sexo biológico declarado
Desc_Renda_Per_Capita	Renda familiar bruta per capita declarada
Desc_Naturalidade	Naturalidade declarada
Desc_Turno	Turno de matrícula do curso
Desc_Sit_Matricula	Situação acadêmica (matriculado, concluído, evadido ou outras);

Fonte: Autor

6.2 Exploração e análise dos dados relacionados à evasão

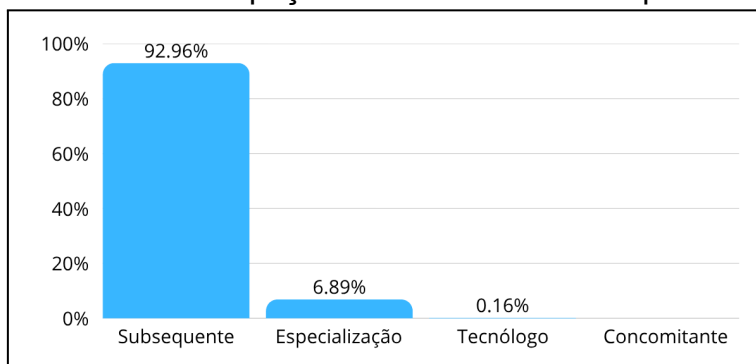
Esta seção apresenta uma leitura descritiva da base com o objetivo de identificar

concentrações de evasão por grupo (sexo, estado civil, cor/raça, modalidade, curso, turno e tipo de cota). As leituras e gráficos a seguir têm caráter exclusivamente descritivo e servem apenas para contextualizar a distribuição dos casos na amostra; não discutem causas, não propõem intervenções e não estabelecem vínculos diretos com a etapa preditiva. Esta observação aplica-se a todas as subseções, gráficos e tabelas desta seção.

6.2.1 Análise por modalidade e turno dos cursos

O Gráfico 2 apresenta a participação de cada modalidade no total de evasões, oferecendo uma visão descritiva da concentração de casos entre as diferentes ofertas de curso. No recorte analisado, a modalidade subsequente concentrou 92,96% do total de evasões, enquanto as demais modalidades somaram 7,04%. Essa leitura identifica modalidades com maior peso relativo na evasão observada, sem inferir causalidade.

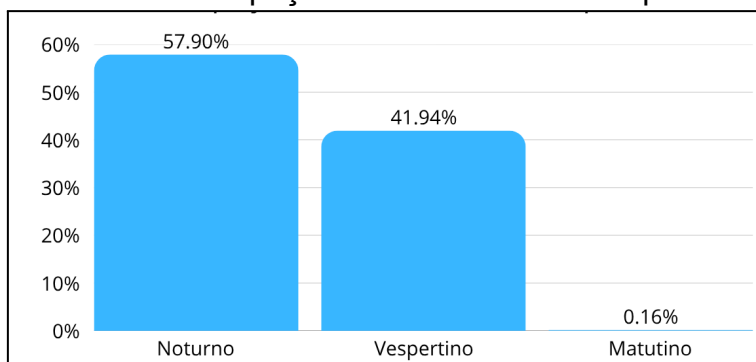
Gráfico 2 - Participação no Total de Evasões por Modalidade de Curso (%).



Fonte: Autor

O Gráfico 3 mostra a participação de cada turno no total de evasões, possibilitando a comparação direta entre períodos de oferta quanto ao volume relativo de desligamentos.

Gráfico 3 - Participação no Total de Evasões por Turno (%).



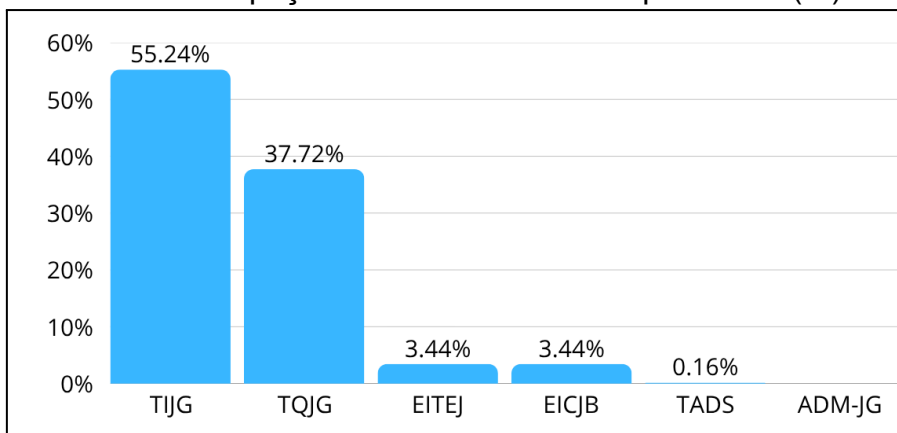
Fonte: Autor

6.2.2 Análise por curso

O Gráfico 4 apresenta a participação de cada curso no total de evasões, oferecendo uma visão descritiva da concentração de desligamentos por área de ensino. A leitura permite identificar cursos com maior peso relativo na evasão

observada, sem inferir causalidade e sem propor explicações para as diferenças registradas.

Gráfico 4 - Participação no Total de Evasões por Curso (%).



Fonte: Autor.

Tabela 3 - Descrição dos cursos.

Abreviação	Curso
TIJG	Técnico em Informática para Internet - Subsequente
TQJG	Técnico em Qualidade - Subsequente
EITEJ	Especialização em Desenvolvimento, Inovação e Tecnologias Emergentes
EICJB	Especialização em Gestão e Qualidade em Tecnologia da Informação e Comunicação
TADS	Tecnologia em Análise e Desenvolvimento de Sistemas
ADM-JG	Assistente Administrativo - PROEJA - SEE/PE

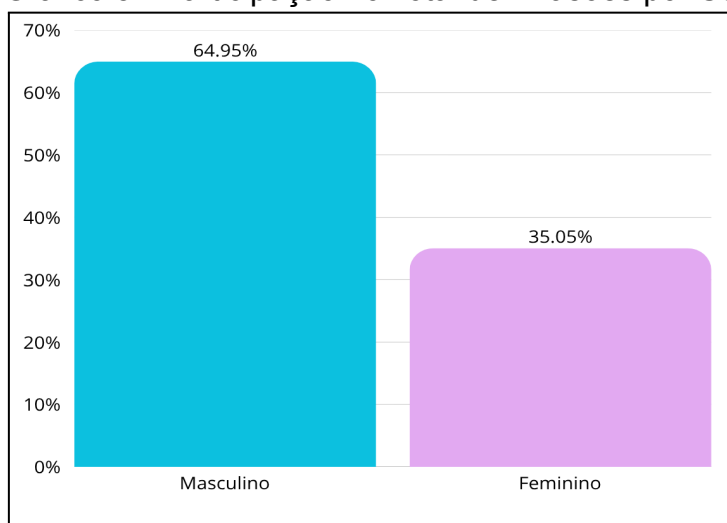
Fonte: Autor.

Conclui-se, nesta leitura descritiva, que no período analisado o curso TIJG concentrou 55,24% das evasões e o TQJG 37,72%, enquanto EITEJ, EICJB, TADS e ADM-JG somaram os 7,04% restantes; esses resultados servem apenas para contextualizar a distribuição dos casos na amostra e não constituem evidência causal.

6.2.3 Análise por sexo e estado civil

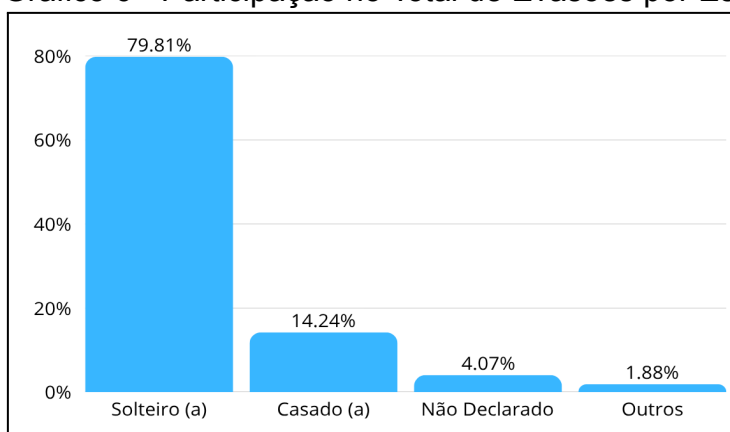
Os Gráficos 5 e 6 apresentam a participação de cada categoria no total de evasões, respectivamente por sexo e por estado civil. No recorte analisado, o sexo masculino respondeu por 64,95% das evasões e o feminino por 35,05%. Quanto ao estado civil, estudantes solteiros concentraram 79,81% das evasões e casados 14,24%, enquanto as categorias “não declarado” e “outros” compõem o percentual remanescente.

Gráfico 5 - Participação no Total de Evasões por Sexo (%).



Fonte: Autor

Gráfico 6 - Participação no Total de Evasões por Estado Civil (%).

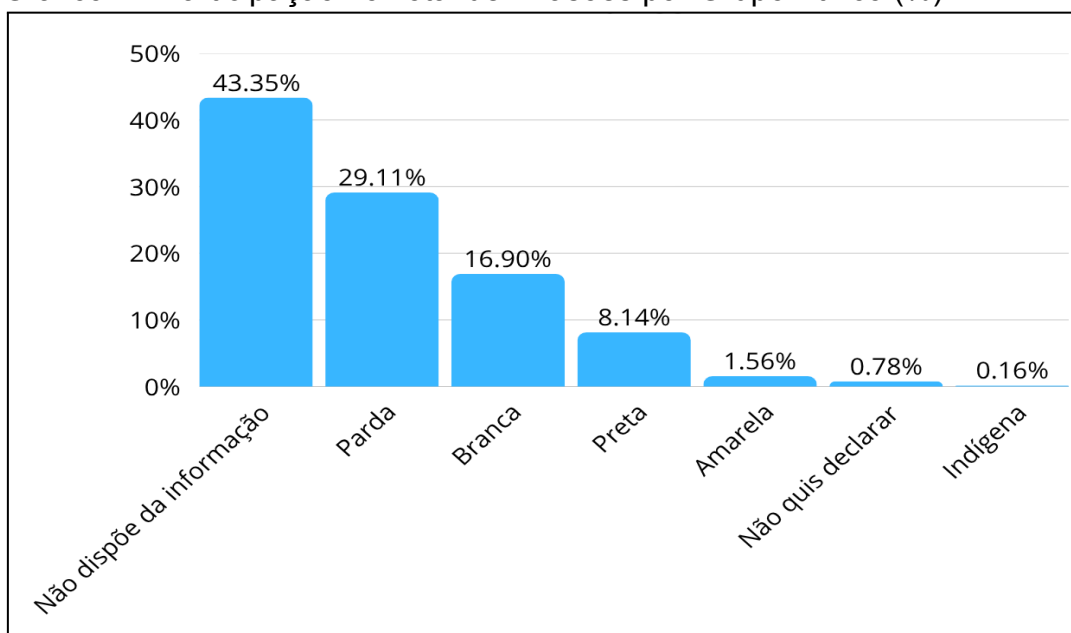


Fonte: Autor.

6.2.4 Análise por grupo étnico e sistema de reserva de vagas (cotas)

O Gráfico 7 destaca o percentual de evasão de estudantes conforme os diferentes grupos étnicos. Essa análise permite identificar possíveis desigualdades relacionadas à permanência nos cursos e aponta a importância de investigar fatores socioculturais que possam influenciar essas diferenças.

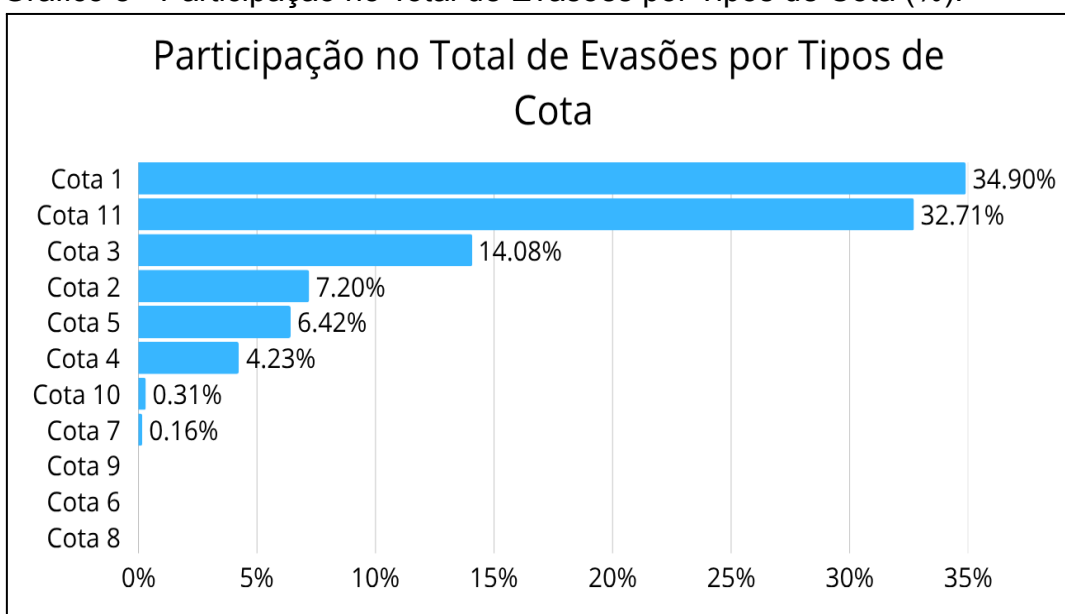
Gráfico 7 - Participação no Total de Evasões por Grupo Étnico (%).



Fonte: Autor.

Por outro lado, o Gráfico 8 apresenta o percentual de evasão por tipo de cota utilizada no processo seletivo dos cursos. Essa abordagem contribui para avaliar como os diferentes sistemas de reserva de vagas impactam a permanência dos estudantes, sendo a descrição detalhada de cada tipo de cota fornecida na Tabela 4.

Gráfico 8 - Participação no Total de Evasões por Tipos de Cota (%).



Fonte: Autor.

Tabela 4 - Descrição das cotas do sistema de reserva de vagas.

Abreviação	Descrição do tipo da cota
Cota 1	Não possui
Cota 2	Aluno de Escola Pública com renda $\leq$ 1,5 SM por pessoa
Cota 3	Aluno de Escola Pública com renda $\leq$ 1,5 SM por pessoa, autodeclarado preto, pardo ou indígena
Cota 4	Aluno de Escola Pública com renda $>$ 1,5 SM por pessoa
Cota 5	Aluno de Escola Pública com renda $>$ 1,5 SM por pessoa, autodeclarado preto, pardo ou indígena
Cota 6	Aluno de Escola Pública, com deficiência e com renda $\leq$ 1,5 SM por pessoa
Cota 7	Aluno de Escola Pública, com deficiência e com renda $\leq$ 1,5 SM por pessoa, autodeclarado preto, pardo ou indígena
Cota 8	Aluno de Escola Pública, com deficiência e com renda $>$ 1,5 SM por pessoa
Cota 9	Alunos de Escola Pública, sem comprovação renda
Cota 10	Alunos de Escola Pública, sem comprovação renda, autodeclarado preto, pardo ou indígena
Cota 11	Ampla Concorrência

Fonte: Autor.

O Gráfico 8 revela que os estudantes que ingressaram por ampla concorrência ou sem cotas apresentam uma taxa de evasão superior àqueles que ingressaram por meio de cotas. Esse dado sugere que a ausência de benefícios associados aos sistemas de reserva de vagas pode resultar em maiores dificuldades para esses alunos, incluindo desafios como desmotivação, falta de identificação com o ambiente acadêmico, necessidade de conciliar estudos com trabalho, baixa autoestima ou ausência de suporte institucional. Pesquisas futuras podem explorar esses aspectos mais detalhadamente, contribuindo para uma compreensão aprofundada das razões que levam à evasão escolar em diferentes grupos de ingresso, conforme descrito na Tabela 4.

6.3 Tratamento dos dados para treinamento dos algoritmos

Nesta seção, são detalhados os procedimentos adotados para preparar os dados de evasão dos alunos para aplicação nos modelos de inteligência artificial. O primeiro passo foi a conversão de valores categóricos (*strings*) em valores numéricos inteiros. Para isso, foram aplicadas regras simples, atribuindo valores inteiros às categorias textuais de forma lógica ou alfabética. As colunas convertidas incluem: Sexo, Estado civil, Cor, Tipo de curso, Curso, Turno e Tipo de cota. Por exemplo, na coluna "Sexo", os valores "masculino" foram substituídos por 1 e "feminino" por 0.

Na sequência, foram excluídas as colunas relacionadas à situação da matrícula, uma vez que estas representavam a variável alvo (dependente) do estudo. Nesse contexto, o valor 0 indicava uma situação regular, enquanto o valor 1 correspondia à evasão. Posteriormente, os dados foram divididos em variáveis independentes (X) representando todas as colunas restantes, e a variável dependente (Y), que é o *target* (variável que deseja prever). Por fim, os dados foram particionados em conjuntos de treino e teste, utilizando uma proporção de 80% para o treinamento dos modelos e 20% para a avaliação de desempenho.

7 Análise de desempenho dos modelos de aprendizado de máquina

Nesta seção são apresentados os resultados do treinamento e da avaliação de três modelos de aprendizado de máquina: Decision Tree Classifier, XGBoost e Random Forest. A avaliação baseou-se nas métricas do classification report, complementadas pelo índice de Cohen's Kappa, amplamente adotadas em estudos dessa natureza.

Para aperfeiçoar o desempenho, aplicou-se GridSearchCV apenas aos modelos Decision Tree e XGBoost, com validação cruzada k-fold (5, embaralhada) e accuracy como métrica de seleção. Em cada um, avaliou-se uma grade de hiperparâmetros pertinentes (p.ex., max_depth, min_samples_split e min_samples_leaf na Decision Tree; n_estimators, learning_rate, max_depth, min_child_weight, gamma no XGBoost), selecionando-se a melhor combinação pelo desempenho médio na validação; o modelo escolhido foi então ajustado e avaliado no conjunto de teste. O Random Forest foi mantido sem busca exaustiva de hiperparâmetros.

O GridSearch (GridSearchCV) é um procedimento sistemático de otimização de hiperparâmetros: testa-se exaustivamente uma grade pré definida de valores, estima-se o desempenho por validação cruzada e escolhe-se a combinação que maximiza a métrica definida (neste trabalho, accuracy). Esse processo reduz o risco de escolhas ad hoc de hiperparâmetros e tende a produzir modelos mais estáveis e com melhor desempenho geral.

Como métricas de avaliação, utilizaram-se Accuracy, Precision, Recall, F1-Score, Cohen's Kappa e a área sob a curva ROC (ROC-AUC). A Accuracy indica a proporção global de acertos e é mais informativa quando as classes estão equilibradas; a Precision mede, entre os exemplos previstos como evasão, quantos de fato o são (controla “alarmes falsos”); o Recall mede, entre os que realmente evadiram, quantos foram corretamente identificados (evita “casos perdidos”); o F1-Score é a média harmônica entre Precision e Recall, sintetizando o equilíbrio entre ambos, útil em cenários com desbalanceamento; o Cohen's Kappa corrige a acurácia pelo acordo esperado ao acaso, oferecendo uma leitura mais “justa” em bases desbalanceadas; por fim, a ROC-AUC resume, em um único número (0,5 a 1,0), o poder de separação do modelo ao longo de todos os limiares de decisão — quanto maior a área, melhor a capacidade do modelo de ranquear corretamente casos de evasão acima de não-evasão.

7.1 Decision Tree Classifier (DTC)

O *Decision Tree Classifier* utiliza uma abordagem baseada em árvores de decisão, aplicando o método de divisão e conquista para realizar a classificação dos exemplos. Cada nó representa uma regra de decisão fundamentada nos atributos dos dados, permitindo a organização e separação dos exemplos de maneira lógica e eficiente.

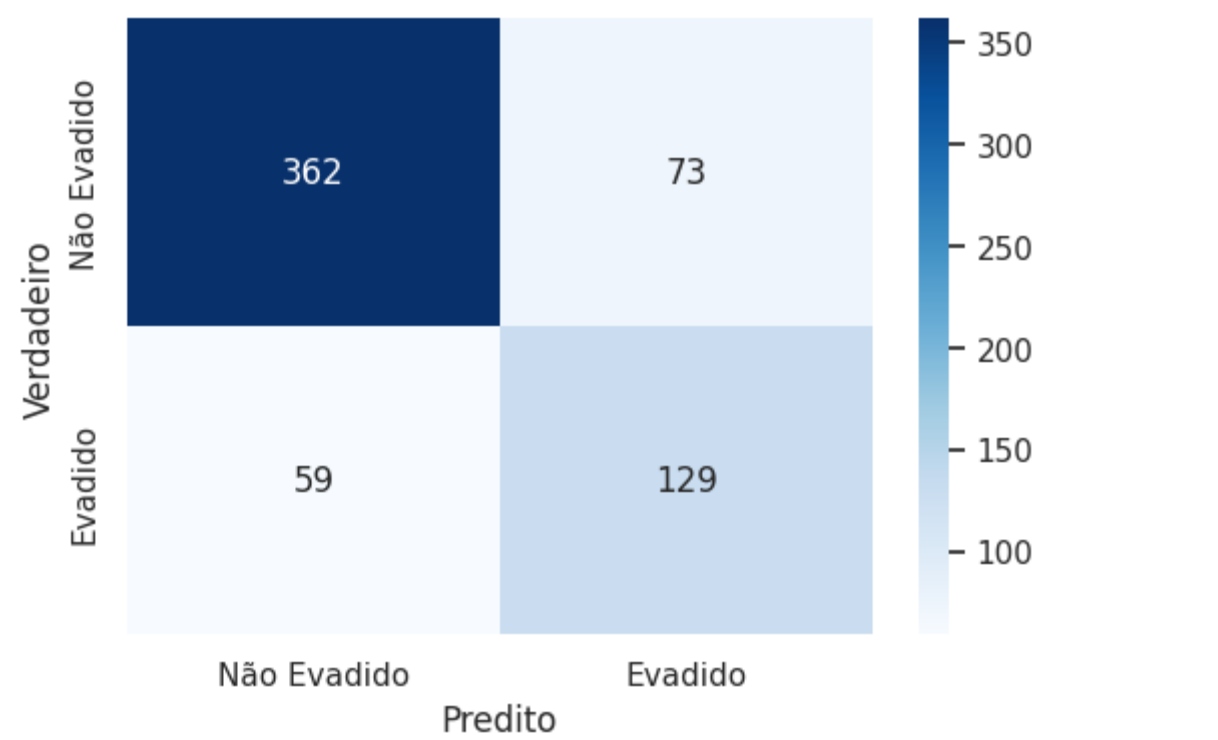
Os principais pontos fortes desse modelo são sua simplicidade, facilidade de interpretação e rapidez de execução. Para avaliar seu desempenho, foram realizados dois experimentos distintos, modelo base (sem GridSearchCV) utilizando hiperparâmetros padrão. Modelo otimizado (com GridSearchCV) ajustado via validação cruzada k-fold (5 divisões), com busca exaustiva sobre parâmetros como max_depth, min_samples_split e min_samples_leaf.

Tabela 5 - Resultados do processamento do *Decision Tree Classifier* sem *gridSearch*.

Acurácia	<i>Precision</i>	<i>Recall</i>	F1-Score	Cohen's Kappa
0,79	0,64	0,79	0,79	0,51

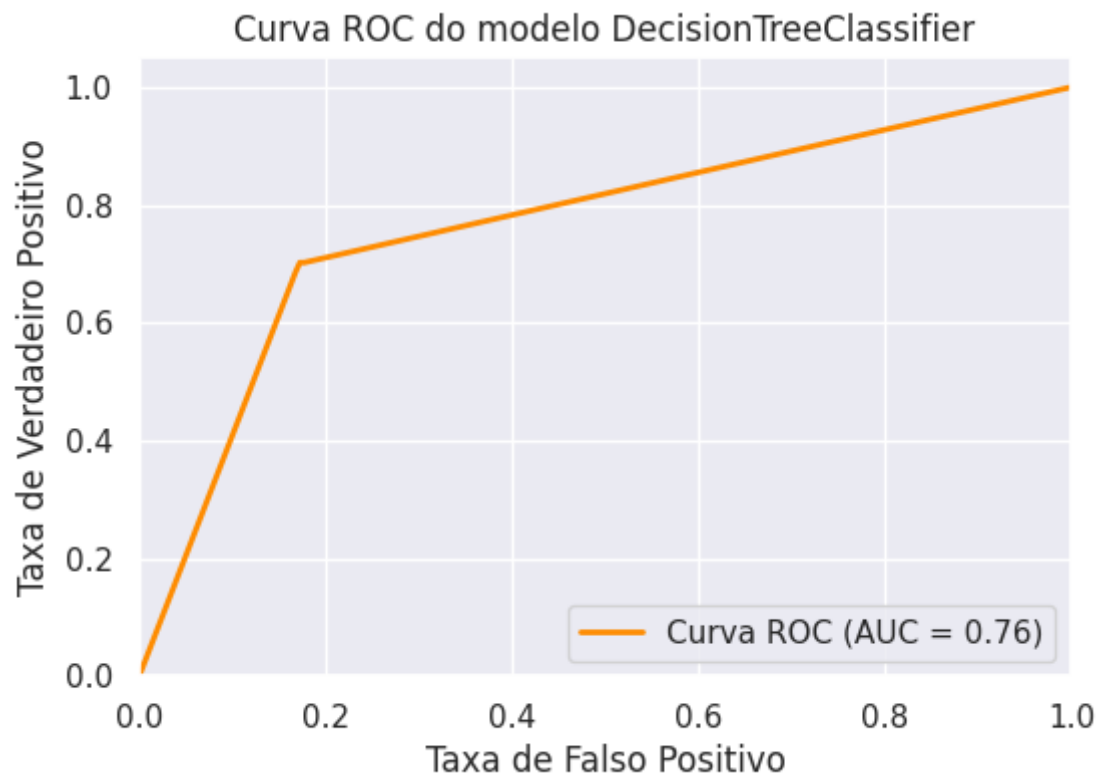
Fonte: Autor.

Gráfico 9 - Matriz de confusão do *Decision Tree Classifier* sem *gridSearch*.



Fonte: Autor.

Gráfico 10 - Curva ROC do *Decision Tree Classifier* sem *gridSearch*.



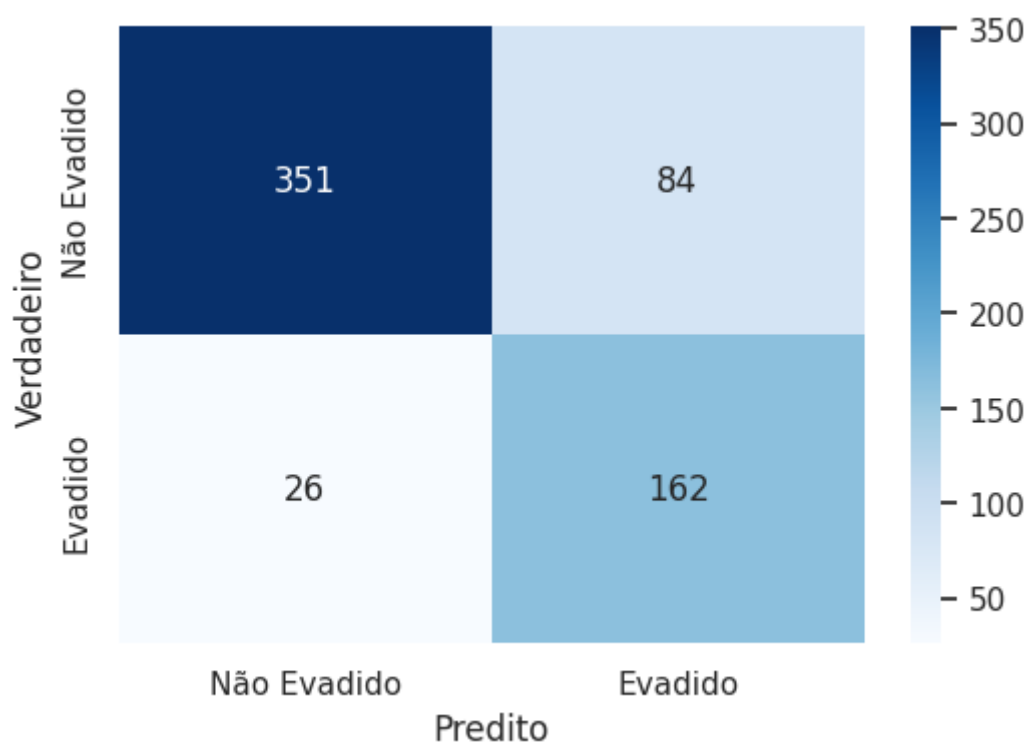
Fonte: Autor.

Tabela 6 - Resultados do processamento do *Decision Tree Classifier* com *gridSearch*.

Acurácia	<i>Precision</i>	<i>Recall</i>	F1-Score	Cohen's Kappa
0,82	0,66	0,82	0,83	0,61

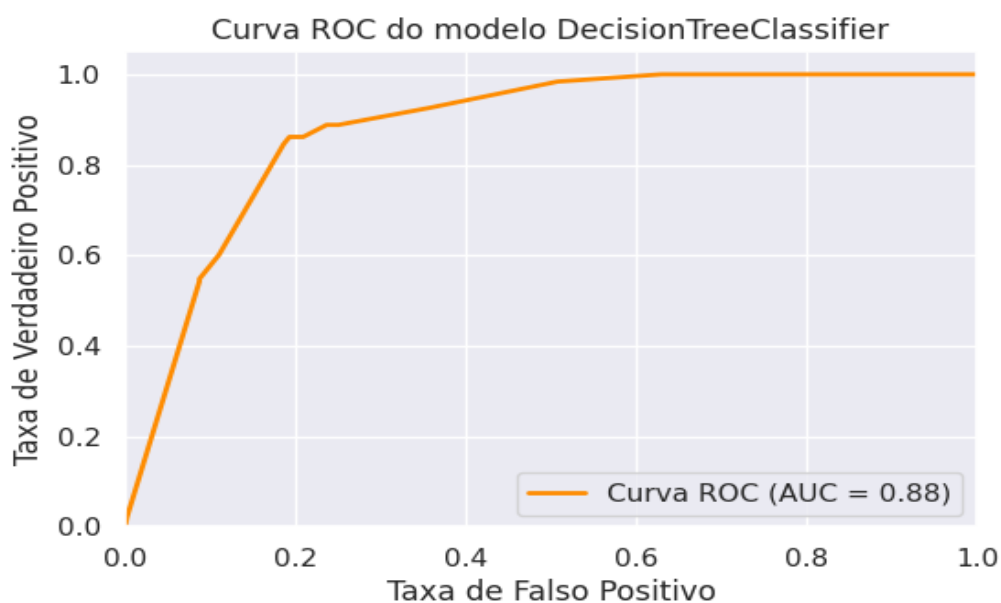
Fonte: Autor.

Gráfico 11 - Matriz de confusão do *Decision Tree Classifier* com *gridSearch*.



Fonte: Autor.

Gráfico 12 - Curva ROC do *Decision Tree Classifier* com *gridSearch*.



Fonte: Autor.

7.2 XGBoost

O *XGBoost* é um algoritmo avançado baseado em árvores de decisão que emprega o método de *boosting*, combinando diversos modelos fracos para formar um modelo robusto e eficiente. Segundo Teixeira (2024), sua principal estratégia consiste em corrigir iterativamente os erros dos modelos anteriores, aprimorando progressivamente o desempenho preditivo.

Sua popularidade decorre da alta eficiência em bases desbalanceadas, do bom

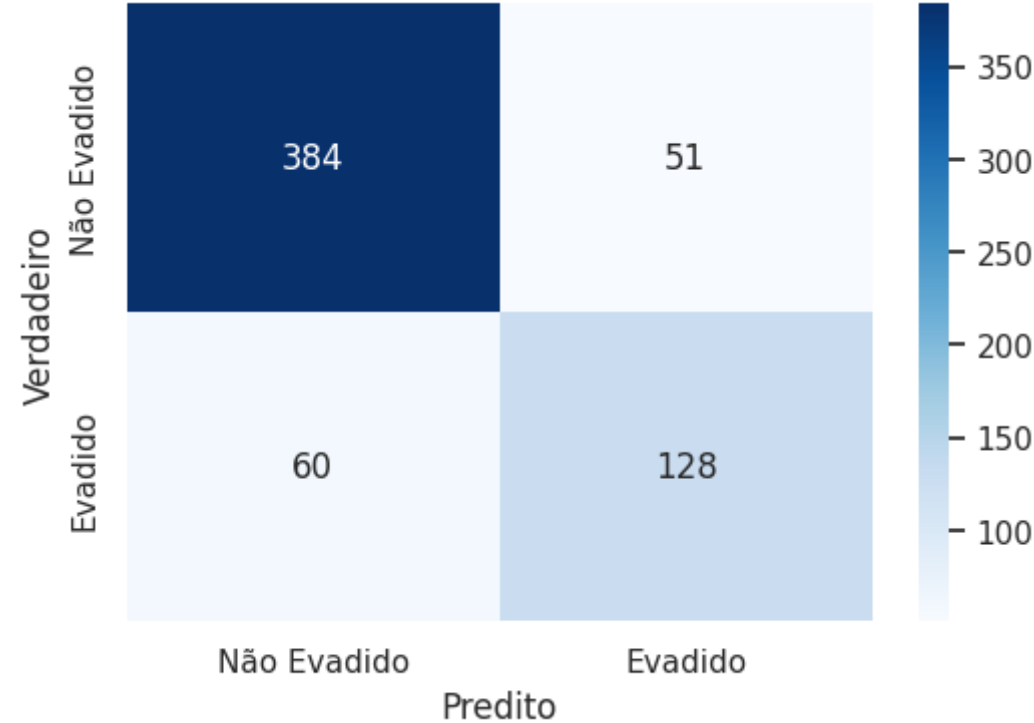
desempenho em competições de *machine learning* e da capacidade de generalização em diferentes contextos. Para avaliar o desempenho do *XGBoost*, foram realizados dois experimentos, modelo base (sem *GridSearchCV*), utilizando hiperparâmetros padrão. Modelo otimizado (com *GridSearchCV*), ajustado via validação cruzada k-fold (5 divisões), com busca exaustiva sobre parâmetros como *n_estimators*, *learning_rate*, *max_depth*, *min_child_weight* e *gamma*.

Tabela 7 - Resultados do processamento do *XGBoost* sem *gridSearch*.

Acurácia	<i>Precision</i>	<i>Recall</i>	F1-Score	Cohen's Kappa
0,82	0,72	0,82	0,82	0,57

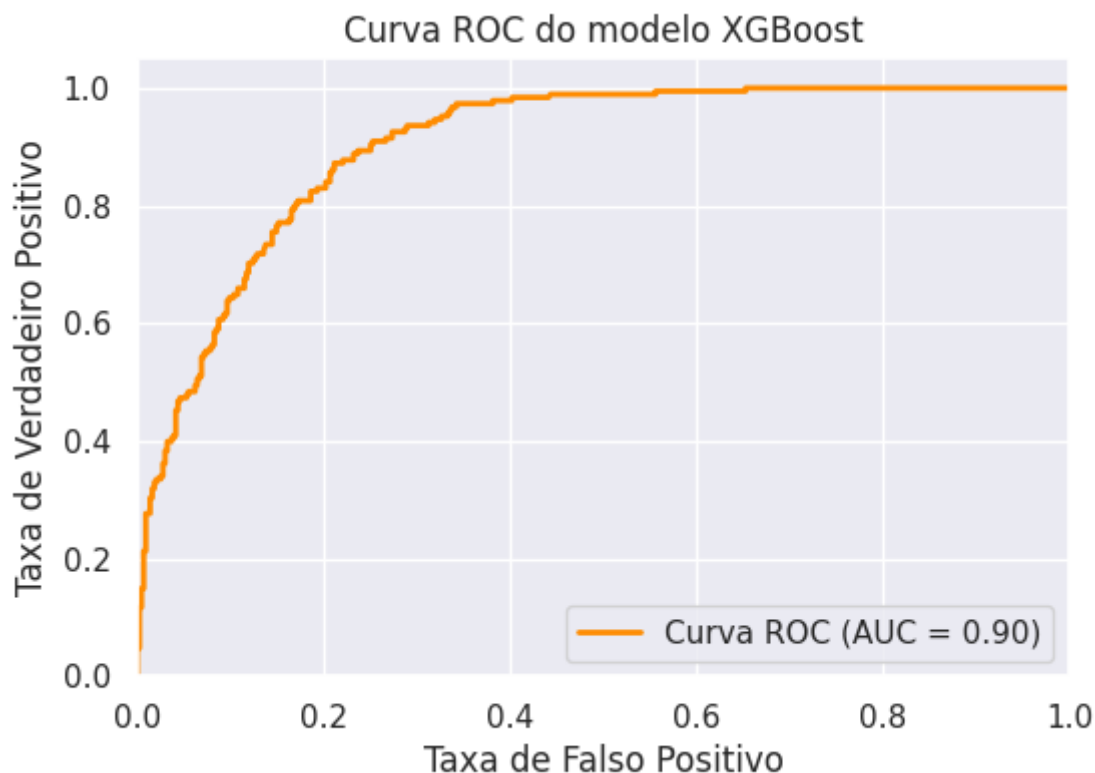
Fonte: Autor.

Gráfico 13 - Matriz de confusão do *XGBoost* sem *gridSearch*.



Fonte: Autor.

Gráfico 14 - Curva ROC do *XGBoost* sem *gridSearch*.



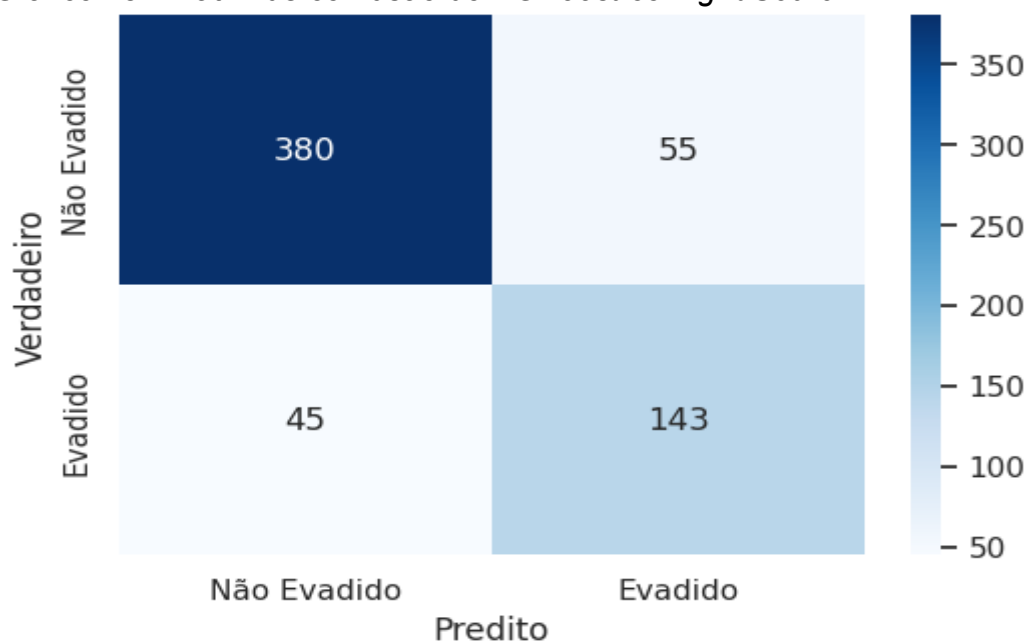
Fonte: Autor.

Tabela 8 - Resultados do processamento do *XGBoost* com *gridSearch*.

Acurácia	<i>Precision</i>	<i>Recall</i>	F1-Score	Cohen's Kappa
0,84	0,72	0,84	0,84	0,62

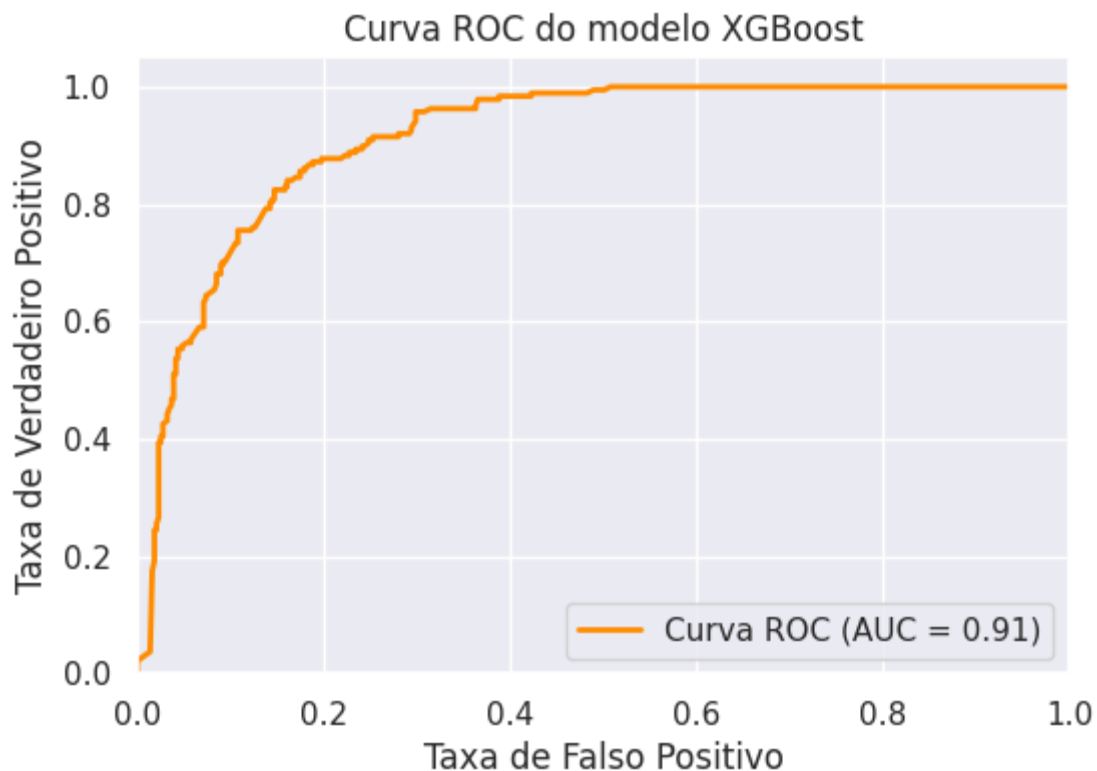
Fonte: Autor.

Gráfico 15 - Matriz de confusão do *XGBoost* com *gridSearch*.



Fonte: Autor.

Gráfico 16 - Curva ROC do *XGBoost* com *gridSearch*.



Fonte: Autor.

7.3 Random Forest

O *Random Forest* é um algoritmo baseado em árvores de decisão que aplica o método de *bagging*, combinando múltiplas árvores treinadas de forma paralela para formar um modelo robusto e estável. Conforme descrito por Breiman (2001), essa técnica utiliza amostras aleatórias com reposição para treinar diferentes árvores de decisão e, em seguida, realiza uma votação majoritária (no caso de classificação) ou média (em regressão) para determinar o resultado final.

A principal vantagem do *Random Forest* é sua resistência ao *overfitting*, por reduzir a variância do modelo sem aumentar o viés, além de apresentar alto desempenho em bases desbalanceadas.

Neste estudo, o modelo foi configurado com `n_estimators = 200` e `random_state = 42`, parâmetros definidos empiricamente com base na literatura e em testes preliminares, não foi utilizada busca exaustiva de hiperparâmetros (*GridSearchCV*), visto que o objetivo era avaliar o comportamento padrão do algoritmo sob uma configuração robusta e consistente.

Tabela 9 - Resultados do processamento do *Random Forest*.

Acurácia	<i>Precision</i>	<i>Recall</i>	F1-Score	Cohen's Kappa
0,8347	0,834	0,8347	0,8343	0,6059

Fonte: Autor.

8 Interpretação dos Resultados

Os resultados das métricas de avaliação (*precision*, *recall*, F1, *accuracy* e Cohen's kappa) permitem comparar a qualidade dos modelos testados. Para *Decision Tree* e *XGBoost*, reportamos os melhores desempenhos obtidos após otimização via *GridSearchCV* (validação cruzada k-fold), a fim de comparar cada algoritmo em sua configuração ótima. Já o *Random Forest* foi avaliado sem *GridSearch*, utilizando *n_estimators* = 200 e *random_state* = 42, mantendo uma configuração robusta e consistente. Nestas condições, o *Random Forest* apresentou o melhor desempenho global no conjunto de teste. Para facilitar a comparação direta, a Tabela 10 consolida os resultados.

Tabela 10 - Comparativo consolidado entre os algoritmos experimentados (*Decision Tree* e *XGBoost* com *GridSearchCV*; *Random Forest* com *n_estimators*=200 e *random_state*=42).

Modelo	<i>Precision</i>	<i>Recall</i>	F1	<i>Accuracy</i>	Kappa
XGBoost	0,72	0,84	0,84	0,84	0,62
RandomForest	0,83	0,83	0,83	0,83	0,60
Decision Tree	0,66	0,82	0,83	0,82	0,61

Fonte: Autor.

Outro estudo que também destacou o modelo *Random Forest* foi o de Silva (2022). Conforme apresentado na Tabela 11, ao utilizar as mesmas métricas de avaliação, mas aplicando-as em um *dataset* diferente, o desempenho superior desse modelo continuou evidente, reforçando sua eficácia em cenários variados.

Tabela 11 - Resultados comparativos de Silva (2022).

Métricas	Naise Bayer	Regressao Logistica	Random Forest	Xgboost
Acurácia	0.26235678	0.77123456	0.73678789	0.77234771
Precisão	0.47356890	0.69346780	0.70567845	0.70345781
Recall	0.49452367	0.72235689	0.75147890	0.75346745
F1 Score	0.23235689	0.70341256	0.71467945	0.71345790

Fonte: Silva (2022)

9 Implementação do sistema

Foi desenvolvida uma aplicação web para operacionalizar a pesquisa e facilitar a reprodução dos procedimentos. A solução materializa a arquitetura proposta em camadas, oferecendo um fluxo unificado do carregamento de bases à execução de análises e à visualização de métricas e previsões, com registro de histórico e rastreabilidade das execuções. Embora não fosse requisito do estudo, a aplicação foi concebida como prova de conceito para uso acadêmico e institucional, preservando o escopo analítico original. A descrição do funcionamento e dos

principais componentes encontra-se nas subseções 9.1 a 9.3.

9.1 Funcionamento do Back-end

O funcionamento do back-end é dividido entre dois módulos principais, Evasion-Watch e Evasion-Pred. O Evasion-Watch é o núcleo do sistema, responsável por receber, validar e armazenar dados estudantis, agendar análises, controlar usuários (autenticação via JWT), disparar tarefas via Kafka e consolidar histórico e status de análises. O Evasion-Pred, por sua vez, é dedicado ao processamento analítico e preditivo: recebe tarefas via Kafka, manipula os dados recebidos do banco MariaDB, executa rotinas de *machine learning* (*Random Forest* e outros), salva os resultados de predição, métricas de treinamento e atualiza o estado das tarefas conforme processamento.

9.2 Funcionamento do Front-end

O front-end do sistema (Evasion-WebApp) foi desenvolvido em Angular e organiza a interface em módulos e componentes voltados a gerenciamento de análises, agendamento, treinamento de modelos e histórico de resultados. A camada de apresentação utiliza Angular Material para oferecer navegação responsiva, pesquisa e visualização de dados. A aplicação consome o back-end por meio de uma API REST (HTTP/JSON), através da qual realiza operações como agendamento e execução de análises, upload de arquivos, treinamento de modelos de aprendizado de máquina e consulta/manipulação de históricos.

Imagem 1 - Tela de análise.

Próx. Execução: 09/10/2025 Mensal - Dia 15

Próx. Execução: 09/10/2025 Semestral - Dia 1

Próx. Execução: 11/10/2025 Anual - Dia 10

Pesquisar

Matricula	Nome	Data da análise	Situação
12345	Maria Silva	01/10/2025	Em risco
67890	João Souza	05/10/2025	Regular
11223	Ana Costa	10/10/2025	Em risco
44556	Pedro Lima	15/10/2025	Regular
78901	Lucas Fernandes	20/10/2025	Regular

Items per page: 5 0 of 0

Fonte: Autor.

Imagem 2 - Tela de Treinamento da IA.



Fonte: Autor.

9.3 Integração entre as camadas

A integração entre as camadas do sistema é baseada em comunicação via API REST e mensageria assíncrona. O front-end (Angular) acessa as funcionalidades do back-end Java (Evasion-Watch) por meio de requisições HTTP, transmitindo dados no formato JSON e utilizando autenticação JWT para garantir segurança e controle de acesso. O Evasion-Watch, por sua vez, é responsável por orquestrar os fluxos principais: consulta e armazenamento dos dados, gerenciamento de usuários, agendamento e disparo de análises e treinamentos.

Ao acionar rotinas de *machine learning* ou previsões preditivas, o Evasion-Watch envia mensagens usando Apache Kafka para o módulo Evasion-Pred (Python), que consome essas solicitações, processa os dados diretamente no banco e devolve os resultados ao próprio back-end. Todo o gerenciamento de estados (tarefas, históricos, agenda) é realizado e centralizado pelo back-end, permitindo que o front-end possa acompanhar e exibir em tempo real o status das operações solicitadas.

Assim, o sistema combina protocolos HTTP (REST), autenticação JWT, JSON e mensageria Kafka para garantir uma arquitetura escalável, segura e modular, que separa claramente responsabilidades entre as camadas de apresentação, processamento, orquestração e análise de dados.

10 Considerações finais

Este estudo apresentou uma aplicação de inteligência artificial voltada para a previsão da evasão escolar entre alunos de cursos técnicos e superiores de um Instituto Federal. Foram empregadas técnicas de aprendizado de máquina para o treinamento e avaliação de três modelos: *XGBoost*, *Random Forest* e *Decision Tree Classifier*. Os dados utilizados consistiram em 2.077 registros extraídos do sistema Q-Acadêmico, incluindo variáveis sociodemográficas e acadêmicas.

Entre os modelos analisados, o *Random Forest* se destacou, alcançando os melhores resultados em todas as métricas avaliadas, mesmo sem ajustes adicionais. Os achados demonstram que os modelos desenvolvidos foram capazes de prever a evasão escolar com elevada acurácia e precisão, apresentando ainda um equilíbrio satisfatório entre recall e *F1-score*. Tais resultados oferecem suporte significativo

para o desenvolvimento de estratégias mais eficazes no enfrentamento da evasão escolar.

10.1 Limitações

Como trabalhos futuros, propõe-se ampliar a base de dados em volume e diversidade, incluindo novas coortes e atributos relevantes e segmentação temporal que diferencie períodos pré, durante e pós-pandemia, de modo a reduzir viés conjuntural e fortalecer a generalização. Recomenda-se explorar de forma sistemática os insights da seção 6.2, convertendo achados descritivos em hipóteses testáveis e em engenharia de atributos, por exemplo a combinação de modalidade e turno, indicadores de participação no total de evasões e taxas internas por grupo, bem como segmentações que auxiliem a seleção de variáveis. Sugere-se evoluir o sistema de treinamento já implementado com ingestão contínua de dados, melhores dashboards e melhores personalizações de treinamento e previsão da evasão. Na modelagem, além dos algoritmos atualmente utilizados: XGBoost, Random Forest e Decision Tree, recomenda-se avaliar CatBoost, LightGBM, regressão logística regularizada, SVM e redes neurais multicamadas. Por fim, recomenda-se adotar mais métricas alinhadas ao problema, incluindo PR-AUC e Brier score e também realizar validação externa em dados de outras instituições para avaliar a robustez e a transferência dos resultados.

10.2 Trabalhos futuros

Para pesquisas futuras, planeja-se explorar técnicas adicionais de aprendizado de máquina, como redes neurais artificiais e deep learning, que possuem potencial para lidar com maior complexidade e identificar padrões mais avançados nos dados. Além disso, propõe-se ampliar o conjunto de dados, tanto em número de instâncias quanto em variedade de características, o que pode enriquecer as análises e melhorar a capacidade preditiva dos modelos. Outra direção importante seria aplicar os algoritmos desenvolvidos em dados de outras instituições de ensino, visando avaliar a sua capacidade de generalização e a validade dos resultados em diferentes contextos educacionais.

REFERÊNCIAS

Breiman, L. Random Forests. *Machine Learning. Spring Nature Link* 45, 5–32 (2001). Disponível em: <https://doi.org/10.1023/A:1010933404324> Acesso em: 10 de setembro de 2024.

Beltrán, C. A. R.; Xavier, J. C.; Barreto, C. A. S.; Neto, C. A. O. Plataforma de Aprendizado de Máquina para Detecção e Monitoramento de Alunos com Risco de Evasão. Anais do Simpósio Brasileiro de Informática na Educação. 2019. Disponível em: <http://milanesa.ime.usp.br/rbie/index.php/sbie/article/view/8892/6445> Acesso em: 13 de setembro de 2024.

COLPO, M.; PRIMO, T.; AGUIAR, M. Predição da evasão estudantil: uma análise comparativa de diferentes representações de treino na aprendizagem de modelos genéricos. In: Anais do XXXII **Simpósio Brasileiro de Informática na Educação**. Porto Alegre: SBC, 2021. p. 873–884. DOI: 10.5753/sbie.2021.218517. Disponível em: <https://sol.sbc.org.br/index.php/sbie/article/view/18114>. Acesso em: 20 de setembro de 2024.

DONNA, G. T. Análise da evasão estudantil com estudo de caso o IFES Campus Cachoeiro de Itapemirim utilizando recursos do Machine Learning. 2023. 57 f. Trabalho de conclusão de curso (Bacharelado em Sistemas de Informação) – **Instituto Federal do Espírito Santo**, Campus Cachoeiro de Itapemirim, Cachoeiro de Itapemirim. Disponível em: <https://repositorio.ifes.edu.br/handle/123456789/3461>. Acesso em: 02 de outubro de 2024.

FAYYAD, U.; PIATETSKY-SHAPIO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37, 1996. DOI: 10.1609/aimag.v17i3.1230. Disponível em: <https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1230>. Acesso em: 11 de outubro de 2024.

IBM (International Business Machines Corporation). Machine learning: o que é, como funciona e por que é importante. **IBM tópico Machine Learning** Disponível em: <https://www.ibm.com/br-pt/topics/machine-learning>. Acesso em: 12 de outubro de 2024.

INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE PERNAMBUCO (IFPE). **Ministério da Educação** Nota de Auditoria nº 005/2024. Disponível em: <https://portal.ifpe.edu.br/wp-content/uploads/2024/04/Nota-de-Auditoria-no-005.2024.pdf>. Acesso em: 15 de outubro de 2024.

JOHANN, C. C. Evasão escolar no Instituto Federal Sul-Rio-Grandense: um estudo de caso no campus Passo Fundo. 2012. Dissertação (Mestrado em Educação) – **Universidade de Passo Fundo**, Passo Fundo. Disponível em: <http://tede.upf.br/jspui/handle/tede/739?mode=full#preview-link0>. Acesso em: 17 de outubro de 2024.

KOHAVI, R.; SOMMERFIELD, D.; DOUGHERTY, J. Data mining using MLC++: A Aprendizado de Máquinas library in C++. **International Journal on Artificial Intelligence Tools**, 1997. Disponível em: https://www.academia.edu/18484450/Data_Mining_Using_a_Machine_Learning_Library_in_C. Acesso em: 22 de outubro de 2024.

MELO, E. C.; SOUZA, F. H.; SANTOS, E. B. Predição da evasão escolar nos cursos superiores do IFMG – Campus Bambuí com o apoio de técnicas de aprendizado de máquina. **Revista Eletrônica de Sistemas de Informação e Gestão Tecnológica**, v. 12, n. 1, 2022. Disponível em: <http://periodicos.unifacef.com.br/resiget/article/view/2342>. Acesso em: 25 de outubro de 2024.

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**, v. 32, p. 1–21, 2024. DOI: 10.5753/rbie.2024.3542. Disponível em: <https://journals-sol.sbc.org.br/index.php/rbie/article/view/3542>. Acesso em: 31 de outubro de 2024.

RIFFEL, S. M.; MALACARNE, V. Evasão escolar no ensino médio: o caso do Colégio Estadual Santo Agostinho no município de Palotina – PR, 2010. **Dia a Dia Educação** Disponível em: <http://www.diaadiaeducacao.pr.gov.br/portals/pde/arquivos/1996-8.pdf>. Acesso em: 02 de novembro de 2024.

RODRIGUES, E. H. A. Análise de caracterização quantitativa e predição da evasão escolar nos cursos da área de computação do ICEA por meio de técnicas de Data Science. 2022. 71 f. Monografia (Graduação em Engenharia de Computação) – **Instituto de Ciências Exatas e Aplicadas**, Universidade Federal de Ouro Preto, João Monlevade. Disponível em: <https://monografias.ufop.br/handle/35400000/4414>. Acesso em: 03 de novembro de 2024

SILVA, M. A. C. Sistema de análise preditiva da evasão discente utilizando métodos de aprendizado de máquina. **C.E.S.A.R – Centro de Estudos e Sistemas Avançados do Recife**, 2022.

TEIXEIRA, J. S. T. Uma comparação entre WaveNet e XGBoost para propagação direta de ondas tradicionais e inversão sísmica usando modelos de camada horizontal. Orientador: Dr. João Medeiros de Araújo. 2024. 92f. Tese (Doutorado em Ciência e Engenharia de Petróleo) - **Centro de Ciências Exatas e da Terra**, Universidade Federal do Rio Grande do Norte, Natal, 2024. Disponível em: <https://repositorio.ufrn.br/handle/123456789/60485> Acesso em: 07 de novembro de 2024.