

Instituto Federal de Brasília
Campus Brasília
Tecnologia em Sistemas para Internet

Milena de Araujo Andrade e Rayra Feitosa Alves

UMA ANÁLISE DOS DADOS DE EVASÃO NOS CURSOS TÉCNICOS DO EIXO DE INFORMAÇÃO E COMUNICAÇÃO DO INSTITUTO FEDERAL CAMPUS BRASÍLIA

Brasília - DF
2025

Milena de Araujo Andrade e Rayra Feitosa Alves

UMA ANÁLISE DOS DADOS DE EVASÃO NOS CURSOS TÉCNICOS DO EIXO DE INFORMAÇÃO E COMUNICAÇÃO DO INSTITUTO FEDERAL DE BRASÍLIA CAMPUS BRASÍLIA

Trabalho de Conclusão de Curso apresentado ao curso de Tecnologia em Sistemas para Internet do Instituto Federal de Brasília do Campus Brasília, como parte da exigência para obtenção do título de Tecnólogo em Sistemas para Internet.

Orientador(a): Me. Fernando Wagner Brito Hortêncio Filho
Computação/IFB

Brasília - DF
2025

A554 Andrade, Milena de Araújo.
Uma análise dos dados de evasão nos cursos técnicos do eixo de informação e comunicação do Instituto Federal campus Brasília / Milena de Araújo Andrade, Rayra Feitosa Alves. — Brasília, 2025.
68 f. : il. color.

Orientador: Fernando Wagner Brito Hortêncio Filho.
Trabalho de conclusão de curso (Graduação) — Instituto Federal de Educação, Ciência e Tecnologia de Brasília, Tecnologia em Sistemas para Internet, 2025.

1. Evasão escolar. 2. Aprendizagem de máquina. 3. Cursos técnicos. I. Alves, Rayra Feitosa. II. Hortêncio Filho, Fernando Wagner Brito (orient.). III. Título.

CDU: 004

Ficha catalográfica elaborada com os dados fornecidos pelas autoras.

Milena de Araujo Andrade e Rayra Feitosa Alves

UMA ANÁLISE DOS DADOS DE EVASÃO NOS CURSOS TÉCNICOS DO EIXO DE INFORMAÇÃO E COMUNICAÇÃO DO INSTITUTO FEDERAL CAMPUS BRASÍLIA

Trabalho de Conclusão de Curso apresentado ao curso de Tecnologia em Sistemas para Internet do Instituto Federal de Brasília do Campus Brasília, como parte da exigência para obtenção do título de Tecnólogo em Sistemas para Internet.

Aprovado em 10 de Dezembro de 2025.

BANCA EXAMINADORA

_____

Me.Fernando Wagner Brito Hortêncio Filho
Computação IFB
Orientador

_____

Dra.Sylvana Karla da Silva de Lemos Santos
Computação IFB
Doutora

_____

Me.Tiago Henrique Faccio Segato
Computação IFB
Mestre

RESUMO

A evasão escolar é um desafio crescente para as principais instituições de ensino, sendo necessário estudos e soluções eficazes para sua redução. Assim, nesse contexto, este trabalho busca responder a seguinte pergunta de pesquisa: “De que maneira os modelos preditivos podem ser usados para indicar a evasão de estudantes na área de informação e comunicação?”. A pesquisa será guiada com base em dados provenientes de bases do meio acadêmico, especificamente relacionados aos cursos Técnico em Informática e Técnico em Desenvolvimento de Sistemas, da modalidade presencial, coletados ao término do curso. O instrumento de coleta contempla tanto variáveis demográficas (faixa etária, renda familiar per capita, cidade de residência, necessidades específicas, etnia e gênero) quanto variáveis acadêmicas (forma de ingresso, tipo de vaga, situação de matrícula e ano de ingresso). O objetivo a partir desses dados é desenvolver análises técnicas e estatísticas utilizando algoritmos de aprendizagem de máquina para prever padrões de evasão. Os resultados obtidos demonstraram que a utilização de aprendizado de máquina pode ser útil para o acompanhamento do perfil de evasão dos alunos com base nas variáveis mais frequentes, contribuindo para tomadas de decisão dos gestores, coordenadores e profissionais da educação na mitigação de problemas de evasão relacionados aos fatores sociais, econômicos, acadêmicos e emocionais que comprometem a trajetória do aluno.

Palavras-chave: Evasão escolar; Curso técnico;Aprendizagem de máquina.

ABSTRACT

School dropout is a growing challenge for leading educational institutions, requiring effective studies and solutions to reduce it. Thus, in this context, this study seeks to answer the following research question: “How can predictive models be used to indicate student dropout in the field of information and communication?” The research will be guided by data from academic databases, specifically related to classroom-based courses in Computer Technology and Systems Development, collected at the end of the course. The collection instrument includes both demographic variables (age group, per capita family income, city of residence, specific needs, ethnicity, and gender) and academic variables (form of admission, type of vacancy, enrollment status, and year of admission). The objective based on this data is to develop technical and statistical analyses using machine learning algorithms to predict dropout patterns. The results obtained demonstrated that the use of machine learning can be useful for monitoring student dropout profiles based on the most frequent variables, contributing to decision-making by managers, coordinators, and education professionals in mitigating dropout problems related to social, economic, academic, and emotional factors that compromise the student's trajectory.

Keywords: Student dropout; Certificate programs; Machine learning.

LISTA DE FIGURAS

Figura 1 - Taxa de rendimento(abandono).....16
Figura 2 - Taxa de Evasão.....17
Figura 3 - Metodologia CRISP-EDM.....30
Figura 4 - K-fold Cross Validation.....35
Figura 5 - Situação de matrícula.....38
Figura 6 - Distribuição por faixa etária.....39
Figura 7 - Distribuição por gênero.....39
Figura 8 - Distribuição por etnia.....40
Figura 9 - Distribuição por etnia e localização.....40
Figura 10 - Distribuição por renda.....41
Figura 11 - Curva AUC-ROC.....48
Figura 12 - Importância das variáveis.....50
Figura 13 - Curva AUC-ROC.....58
Figura 14 - Nível de importância das variáveis.....58

LISTA DE TABELAS

Tabela 1 - Dados.....31
Tabela 2 - Dados de situação matrícula.....32
Tabela 3 - Matriz de confusão utilizada na avaliação dos modelos.....36
Tabela 4 - Matrizes de confusão sem GridSearch.....43
Tabela 5 - Matrizes de confusão após a aplicação do Grid Search.....45
Tabela 6 - Resultados dos modelos.....47
Tabela 7 - Matrizes de confusão sem Grid Search.....52
Tabela 8 - Matrizes de confusão com Grid Search.....54
Tabela 9 - Avaliação dos modelos (EMI).....56

LISTA DE ABREVIATURAS E SIGLAS

ANN	Artificial Neural Networks
AB	AdaBoost
CRISP-EDM	CRoss-Industry Standard Process for Educational Data Mining
DT	Decision Tree
EDM	Educational Data Mining
EMI	Ensino Médio Integrado
EPT	Educação Profissional e Tecnológica
GB	Gradient Boosting
IFB	Instituto Federal de Educação, Ciência e Tecnologia de Brasília
IFCE	Instituto Federal do Ceará
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
KNN	K-nearest Neighbors
LR	Logistic Regression
NB	Naive Bayes
NEIS	National Education Information System
RF	Random Forest
PPE	Plano Estratégico para Permanência e Êxito
SVM	Support Vector Machines
RF	Random Forest
TDS	Técnico Subsequente em Desenvolvimento de Sistemas

SUMÁRIO

1 INTRODUÇÃO..... 13
1.1 Tema..... 15
1.2 Problema..... 15
1.2.1 Justificativa..... 18
1.2.2 Objetivo geral..... 19
1.2.3 *Objetivos específicos*..... 19
1.4 Estrutura do TCC..... 20
1.5 Classificação da pesquisa..... 20
2 CONCEITOS GERAIS E REVISÃO DE LITERATURA..... 21
2.1 Evasão escolar..... 21
2.2 Estatísticas descritivas..... 22
2.3 Machine learning..... 23
2.3.1 Support Vector Machine (SVM)..... 24
2.3.2 Random Forest (RF)..... 24
2.3.3 K-Nearest Neighbors (KNN)..... 25
2.3.4 Logistic Regression (LR)..... 25
2.3.5 Gradient Boosting (GB)..... 26
2.3.6 AdaBoost (AB)..... 26
2.4 Trabalhos relacionados..... 26
2.4.1 Trabalhos Nacionais..... 27
2.4.1.1 CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais..... 27
2.4.1.2 Análise de Classificadores para Predição de Evasão de Campi de uma Instituição de Ensino Federal..... 27
2.4.2 Trabalhos Internacionais..... 28
2.4.2.1 Machine Learning Predicts Upper Secondary Education Dropout as Early as the End of Primary School..... 28
2.4.2.2 Application of the performance of machine learning techniques as support in the prediction of school dropout..... 29
3 METODOLOGIA..... 29
3.1 Entendimento e coleta dos dados..... 31
3.2 Pré-processamento..... 32
3.3 Ferramentas a serem utilizadas..... 33
3.4 Treinamento e teste..... 34
3.5 Métricas para desempenho de classificação..... 35
4 RESULTADOS E DISCUSSÃO..... 38
4.1 Caracterização da amostra..... 38
4.2 Resultados após o pré-processamento dos dados..... 41
4.3 Avaliação de Desempenho - Técnico subsequente em Desenvolvimento de Sistemas..... 43

4.4 Avaliação de Desempenho - Técnico em Informática Integrado ao Ensino Médio..... 51
5 CONCLUSÃO..... 61
REFERÊNCIAS..... 62

1 INTRODUÇÃO

O Instituto Federal de Brasília (IFB) é uma instituição pública de ensino que, há mais de 15 anos, oferece cursos de educação profissional e tecnológica (EPT) no Distrito Federal, formando e qualificando profissionais para o mercado de trabalho. Faz parte da Rede Federal de Educação Profissional e Tecnológica, criada em 2008, que reúne mais de 600 unidades distribuídas por todo o Brasil (BRASIL, 2026).

Dentre as possibilidades de cursos ofertados, os atrelados ao ensino técnico desempenham um papel fundamental e importante na formação profissional e na inserção de jovens e adultos no mercado de trabalho. Uma pesquisa realizada com gestores de recursos humanos aponta que questões relacionadas ao conhecimento técnico, em conjunto com competências socioemocionais, estão entre os principais critérios considerados e analisados em contratações de nível médio ou técnico. A comunicação, em especial, é destacada como a habilidade mais valorizada, e o diploma técnico é visto como um indicativo de comprometimento e foco por parte dos candidatos (Fundação Roberto Marinho *et al*., 2021).

A oferta de cursos técnicos, no entanto, vem acompanhada de desafios. Dentre estes, destaca-se a evasão escolar, um fenômeno que desperta a atenção de gestores, professores e demais profissionais envolvidos com a educação. Trata-se de um problema não exclusivo do IFB, mas presente em diversas instituições de ensino do país. Como aponta Corrêa e Côrrea (2024), a evasão escolar não se limita a algumas unidades escolares, sendo uma questão nacional (e até mesmo internacional) que ocupa papel relevante nas discussões e pesquisas educacionais brasileiras. Ainda segundo os autores, fatores como a ausência de políticas públicas governamentais, falta de incentivo familiar e desigualdades sociais contribuem para exclusão educacional e social.

Segundo Silva e Araújo (2017), a evasão escolar é caracterizada pelo abandono do curso por parte do estudante, sem que haja retorno ao sistema de ensino. Esta é uma das principais problemáticas enfrentadas pelas instituições de ensino superior, pela educação profissional e tecnológica, e pela educação de jovens e adultos. É um fenômeno complexo, que pode ser influenciado por múltiplos fatores, sejam eles sociais, econômicos, acadêmicos e emocionais, comprometendo

assim a trajetória educacional dos estudantes, o investimento público realizado e a qualidade do ensino oferecido.

De acordo com a UNESCO (2020), esse abandono não impacta só a vida do aluno, mas também o desenvolvimento social e econômico das comunidades. Por isso, o combate à evasão tem sido uma preocupação em várias partes do mundo, e buscar formas de evitar que os alunos desistam dos estudos se tornou uma prioridade em muitas políticas educacionais.

Os impactos da evasão escolar são multifacetados, atingindo não apenas o estudante, mas também a instituição de ensino e a sociedade como um todo. Quando um aluno abandona um curso, perde-se uma oportunidade de formação profissional, de mobilidade social e de inserção qualificada no mercado de trabalho. Para as instituições, segundo Silva et al. (2020) a evasão compromete indicadores de desempenho e os recursos como infraestrutura, tecnologia e equipe previamente destinados, acabam sendo subutilizados, o que pode gerar desperdícios. Já para a sociedade, a evasão contribui para a manutenção de desigualdades sociais, além de impactar negativamente o desenvolvimento econômico e social, ao limitar a qualificação da força de trabalho disponível.

Neste contexto, estratégias de aprendizagem de máquina vêm sendo utilizadas como uma abordagem promissora para ajudar a mitigar o problema da evasão escolar. Conforme apontam Magalhães et al. (2025), técnicas de inteligência artificial têm sido amplamente investigadas na literatura para a predição da evasão, demonstrando potencial na identificação precoce de estudantes em risco. Essas técnicas permitem analisar grandes volumes de dados educacionais para identificar padrões e fatores que indicam maior risco de abandono. Com base nessas informações, é possível gerar modelos preditivos capazes de antecipar quais estudantes têm maior probabilidade de evasão, possibilitando intervenções preventivas mais direcionadas e eficazes.

Considerando a relevância do tema, este trabalho tem por finalidade investigar os dados acerca do êxito e da evasão nos cursos do Ensino Médio Integrado em Informática (EMI) e no Técnico Subsequente em Desenvolvimento de Sistemas (TDS) do IFB Campus Brasília. Para isso, serão gerados e avaliados

modelos de aprendizado de máquina, com foco na predição do risco de evasão discente, buscando verificar o potencial uso prático dessas ferramentas para apoiar a instituição na tomada de decisões mais eficazes, contribuindo para a permanência e sucesso dos estudantes.

1.1 Tema

No que concerne ao eixo de informação e comunicação, o Instituto Federal de Brasília - Campus Brasília oferece vagas tanto para o curso técnico integrado ao ensino médio quanto para o técnico subsequente a este. A evasão nesses cursos preocupa a instituição, pois quando o aluno evade há um desperdício de recursos e menos qualificação de pessoas para o mercado de trabalho.

Com a utilização de modelos preditivos pode-se identificar padrões e estimar quais são os fatores que mais contribuem para a evasão escolar. Por meio de uma análise estatística descritiva, será possível examinar de forma detalhada os dados relacionados aos estudantes, como idade, turno, forma de ingresso, permitindo assim traçar um panorama mais claro da evasão nos cursos analisados. Essa abordagem possibilita identificar perfis mais suscetíveis ao abandono e reconhecer possíveis padrões que se repetem ao longo dos anos. Com base nessas informações, educadores, coordenadores e gestores poderão compreender melhor as características do público atendido e adotar medidas mais eficazes de acompanhamento, intervenção e prevenção, com foco na permanência e no êxito escolar.

1.2 Problema

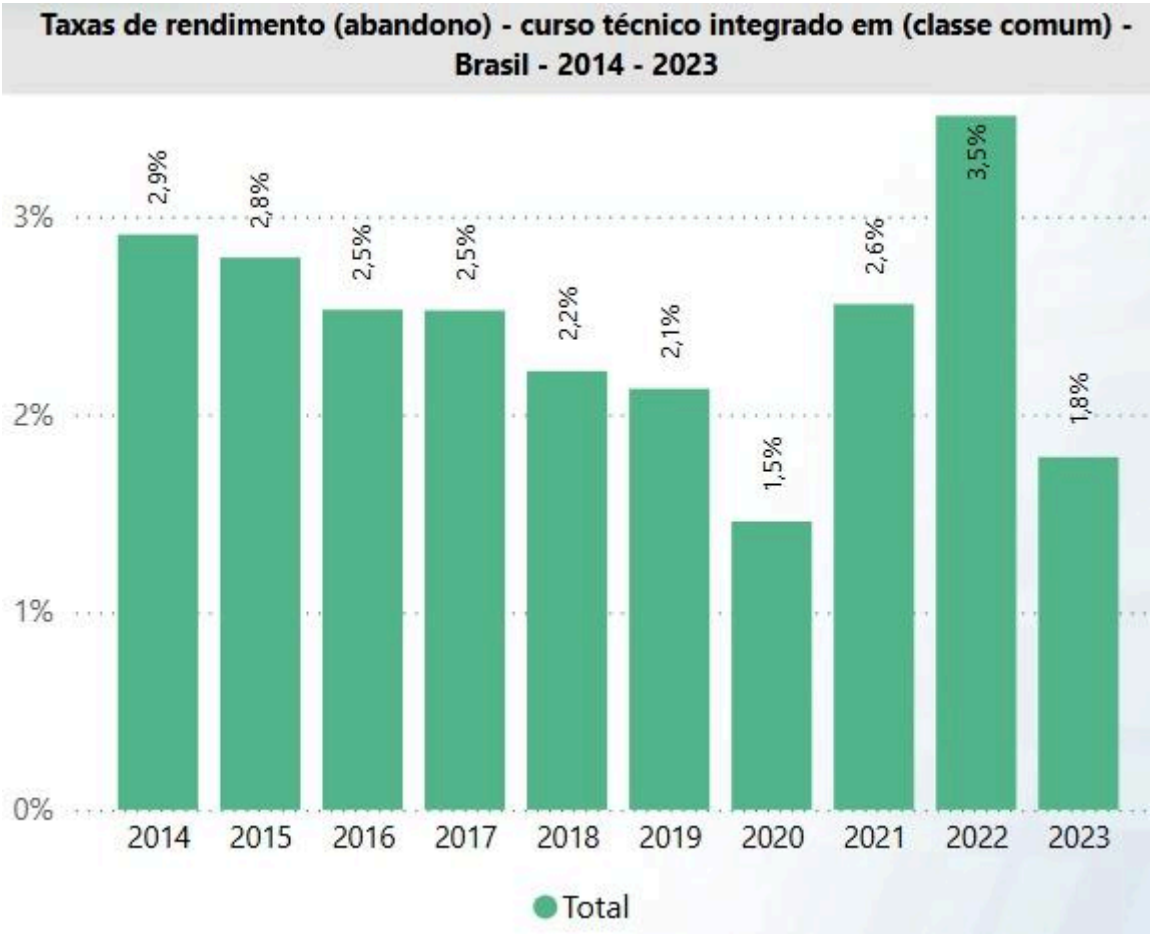
A evasão escolar nos cursos técnicos no Brasil é um um problema que vem sendo estudado no sentido de encontrar soluções que possam reduzir as ocorrências desse fenômeno. Na pesquisa realizada por Almeida e Ferreira (2014) a evasão escolar está relacionada à falta de interesse dos alunos. Também como as dificuldades enfrentadas pelos estudantes para conciliar escola e trabalho.

Segundo o Plano Estratégico para Permanência e Êxito dos Estudantes (PPE) do Instituto Federal do Ceará (IFCE), (Vieira *et al.*, 2017), há diversas variáveis que podem colaborar para a evasão dos estudantes como fatores individuais, que dizem respeito às características pessoais dos alunos, os fatores internos à instituição,

relacionados com problemas como infraestrutura, currículo, gestão e práticas pedagógicas, e fatores externos, relacionados a dificuldades sociais, financeiras ou profissionais que influenciam a permanência do estudante.

De acordo com dados do Painel Estatístico do Censo Escolar da Educação Básica, disponibilizado pelo Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP), a taxa de abandono em cursos técnicos integrados ao ensino médio alcançou 3,5% em 2022. A Figura 1 ilustra o salto de evasão no ano de 2022.

Figura 1 — Taxa de rendimento(abandono)

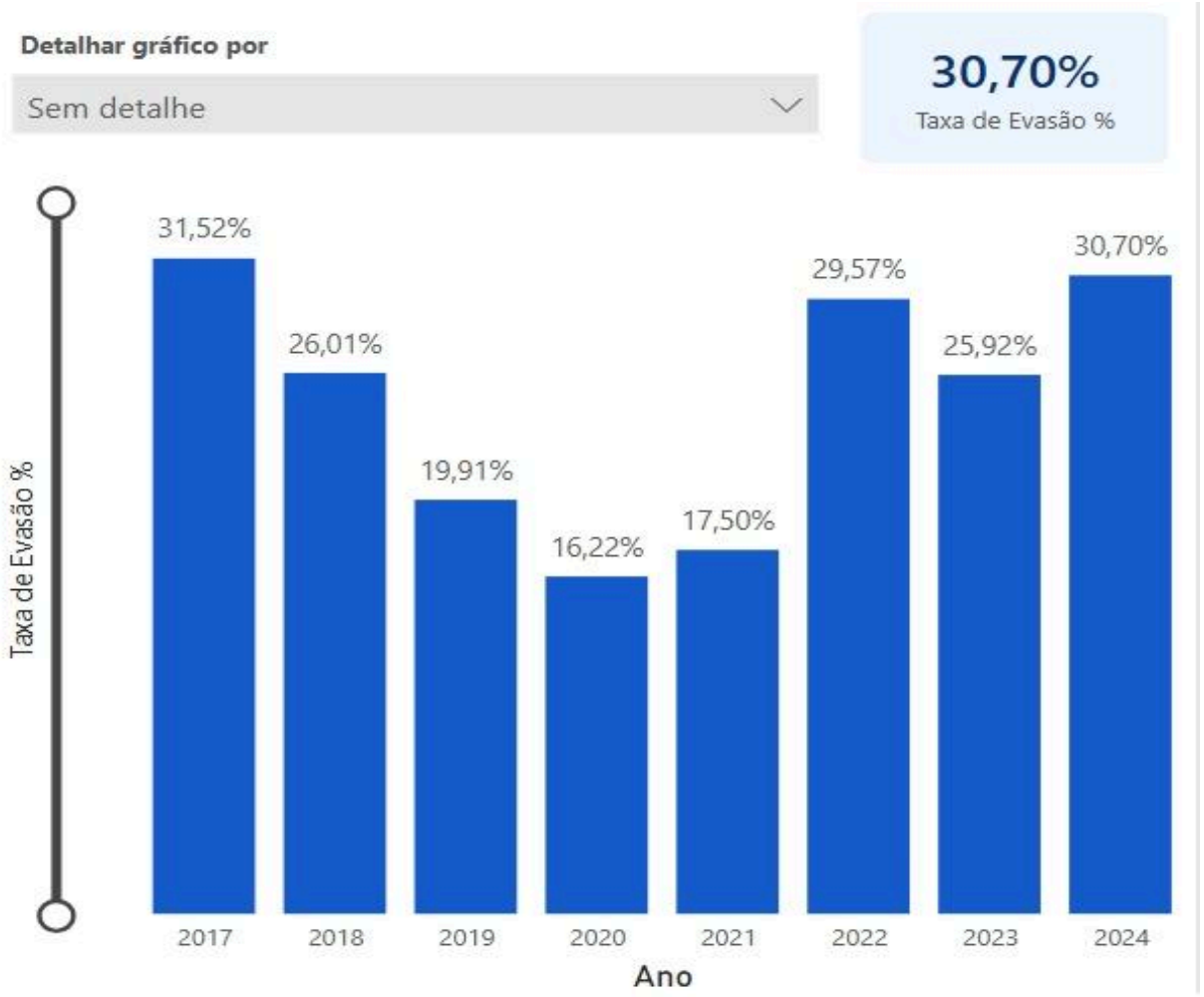


Fonte: via Censo Escolar da Educação Básica, (2024).

Já nas modalidades subsequente e concomitante, os índices de evasão apresentam níveis elevados em comparação ao ensino integrado. Conforme apresentado na Figura 2, com base em dados extraídos da Plataforma Nilo Peçanha - Extrator PNP, a taxa de evasão nessas modalidades variou ao longo dos anos, registrando valores entre 16,22% em 2020 e o seu pico de 31,52% em 2017, com um aumento significativo a partir de 2022, quando atingiu 29,57%. Em 2024, até o

momento, a taxa de evasão é de 30,70%, mantendo-se em patamar elevado e indicando a vulnerabilidade dessas modalidades. Essa tendência reflete os desafios enfrentados pelos estudantes, que frequentemente precisam conciliar os estudos com o trabalho e a vida pessoal, contribuindo diretamente para os elevados índices de abandono escolar.

Figura 2 — Taxa de Evasão



Fonte: via Plataforma Nilo Peçanha - Extrator PNP, (2024).

Mesmo com os esforços da instituição para ajudar os estudantes a continuarem seus estudos, os cursos técnicos do eixo de Informação e Comunicação do IFB - Campus Brasília ainda enfrentam alguns casos de evasão (BRASIL, 2025). Isso afeta não só o desempenho da escola, mas também a formação de profissionais qualificados para o mercado de trabalho. Por isso, é importante entender melhor por que tantos alunos desistem e buscar formas de mudar essa realidade. Diante desse

cenário, este trabalho busca responder à seguinte pergunta: De que maneira os modelos preditivos podem ser usados para indicar a evasão de estudantes nos cursos técnicos do IFB?

1.2.1 Justificativa

Compreender os fatores que levam estudantes a abandonarem cursos técnicos é uma demanda urgente para instituições de ensino, como o Instituto Federal de Brasília (IFB), que buscam garantir não apenas a formação qualificada dos seus alunos, mas também a efetividade do investimento público em educação. A evasão discente representa uma perda significativa em termos pedagógicos, sociais e econômicos, afetando a qualidade do ensino, os indicadores institucionais e o desenvolvimento profissional dos estudantes.

Pesquisas na área da educação profissional apontam que a desistência está relacionada a diversos fatores, como a necessidade de trabalhar, problemas pessoais e familiares, dificuldades de transporte e alimentação, falhas no acompanhamento multiprofissional, dificuldades de aprendizagem, relação professor-aluno e falta de equipamentos para a realização de atividades remotas (Moreira; Cardoso, 2024).

Diante deste cenário, torna-se essencial realizar estudos que monitorem e investiguem as causas da evasão escolar dos cursos técnicos, a fim de compreender os motivos que levam os alunos a abandonarem sua formação. Somente com esse entendimento é possível desenvolver estratégias eficazes para conter o avanço dos índices e até mesmo prever quais estudantes apresentam maior risco de evasão.

Neste contexto, este trabalho propõe a utilização de técnicas de Aprendizado de Máquina para a construção de modelos preditivos capazes de identificar padrões associados à evasão, considerando como estudo de caso os cursos técnicos do eixo de informação e comunicação do IFB - Campus Brasília. Essa abordagem visa oferecer à instituição uma ferramenta de apoio à tomada de decisão, permitindo intervenções mais assertivas e personalizadas, contribuindo para a permanência e o sucesso dos estudantes.

1.2.2 Objetivo geral

Gerar, analisar e avaliar modelos preditivos, gerados por técnicas de algoritmos de aprendizado de máquina, utilizando dados de alunos dos cursos Técnico Subsequente em Desenvolvimento de Sistemas e Técnico Integrado em Informática do Instituto Federal de Brasília Campus Brasília.

Neste trabalho, por meio da análise dos dados dos estudantes, serão desenvolvidos modelos preditivos que possam identificar padrões e traçar perfis de alunos com maior risco de evasão, gerando *insights* que irão auxiliar a instituição na adoção de medidas preventivas mais eficazes.

1.2.3 Objetivos específicos

- Fazer uma análise estatística descritiva dos dados dos alunos para entender melhor o cenário da evasão nos dois cursos estudados.
- Mapear o perfil dos alunos evadidos para compreender padrões demográficos, educacionais e comportamentais.
- Criar modelos preditivos usando técnicas de aprendizagem de máquina para identificar alunos com maior ou menor risco de evasão.
- Avaliar o desempenho desses modelos usando métricas que visam elucidar os níveis de acerto das predições.
- Comparar os resultados entre os dois cursos e discutir os fatores que mais influenciam a evasão em cada um.

1.4 Estrutura do TCC

Este projeto de trabalho de conclusão de curso foi organizado em dois capítulos principais, o primeiro apresenta a introdução do trabalho, o tema, o problema, o objetivo geral, os objetivos específicos e a estrutura do TCC. O segundo capítulo apresenta o referencial teórico de diferentes fontes que tratam acerca do tema e dos métodos e algoritmos a serem gerados. O terceiro capítulo trata-se da metodologia utilizada no trabalho, o detalhamento dos dados acadêmicos, as

técnicas de pré-processamento e normalização, ferramentas a serem utilizadas, além de técnicas para treinamento e teste, e suas métricas de avaliação. O quarto capítulo apresenta os resultados das análises estatísticas dos perfis dos estudantes, avaliações e resultados dos modelos gerados. Por fim, o quinto capítulo traz a conclusão, destaca as contribuições das pesquisas, as limitações do estudo e sugestões de técnicas e práticas para trabalhos futuros.

1.5 Classificação da pesquisa

A pesquisa realizada neste trabalho é classificada como descritiva, pois seu principal objetivo é analisar, identificar e descrever características de um conjunto de dados educacionais. A abordagem é de natureza quantitativa, que analisa variáveis numéricas e estatísticas. Esse método permite aprimorar a análise e discussão dos resultados, por meio dos números, propondo melhorias na área educacional que evitem a evasão dos estudantes.

2 CONCEITOS GERAIS E REVISÃO DE LITERATURA

2.1 Evasão escolar

A evasão escolar é um problema que preocupa as instituições de educação no Brasil. Em essência trata-se do desligamento do aluno de alguma instituição de ensino ou de um curso sem que ele o conclua.

Segundo Silva e Araújo (2017) as causas da evasão do aluno envolvem aspectos individuais dos estudantes, questões sociais, econômicas e familiares, além de problemas estruturais e pedagógicos da escola e do sistema educacional. Entre esses fatores, pode-se destacar a necessidade de trabalhar, dificuldades financeiras, violência, desmotivação, métodos de ensino ineficientes, falta de apoio familiar e escolar, e políticas educacionais instáveis.

No contexto do ensino médio Borges, Silva e Silva (2022) mostram que os fatores que levam o aluno a evadir do curso técnico integrado estão ligados à desmotivação dos estudantes, formação anterior inconsistente ou conhecimento limitado, além da reprovação e da dificuldade em acompanhar os conteúdos das disciplinas.

Para (Sacramento; Albuquerque; Cypriano, 2021) há uma variedade de fatores que interferem na permanência do discente no meio acadêmico. As que se destacaram no estudo foram fatores individuais como as diferenças do conhecimento e habilidades dos alunos quando ingressam no curso, falta de identificação com o curso escolhido por conta da pressa na escolha da profissão, fatores internos como aprendizagem, e a falta de organização curricular.

Além disso, fatores externos como a distância da casa do aluno até a instituição e a dificuldade econômica do estudante. Ainda segundo os autores, pesquisas que tratam da evasão dos cursos técnicos no Brasil têm como maior desafio a escassez de informações sobre a temática.

Portanto, compreender os fatores que levam à evasão escolar e desenvolver estratégias de prevenção fundamentadas em evidências são etapas fundamentais para promover a permanência e a formação dos estudantes nos cursos técnicos,

contribuindo para a melhoria da qualidade da educação e para o desenvolvimento social e econômico.

2.2 Estatísticas descritivas

A estatística descritiva constitui um dos principais recursos para o tratamento inicial de dados em pesquisas quantitativas. Por meio dela, é possível organizar, resumir e interpretar informações por meio de medidas de tendência central (média, mediana, moda) e de dispersão (desvio padrão, variância, amplitude), além da utilização de gráficos que facilitam a visualização de padrões (Triviños, 2017). Essas ferramentas permitem compreender a distribuição dos dados, identificar valores atípicos e estabelecer relações preliminares entre variáveis.

No contexto educacional, o uso de estatísticas descritivas tem sido fundamental para analisar o perfil dos estudantes e compreender aspectos associados ao desempenho e à permanência escolar. Segundo (Vasconcellos *et al.*, 2024), a análise inicial dos dados é essencial para a construção de modelos preditivos eficazes. Os autores destacam que indicadores como frequência, desempenho em disciplinas específicas, idade e histórico de reprovações podem ser usados para identificar padrões de comportamento associados à evasão.

Avançando para métodos mais sofisticados, o uso de algoritmos de aprendizado de máquina tem se consolidado como estratégia promissora para prever a evasão escolar. De acordo com Barbosa (2022), ao aplicar técnicas como regressão logística, árvores de decisão, *k-nearest neighbors (KNN)*, máquinas de vetor de suporte *(SVM)* e *Random Forest (RF)* em dados de uma instituição pública de ensino, foi possível prever com acurácia significativa os estudantes com maior risco de evasão. O estudo destacou, ainda, que a preparação dos dados que inclui o uso da estatística descritiva é etapa fundamental para o desempenho dos modelos preditivos.

Nesse sentido, observa-se uma integração entre os métodos estatísticos e os algoritmos computacionais. Enquanto a estatística descritiva oferece subsídios para compreender os dados e selecionar variáveis relevantes, o aprendizado de máquina

proporciona ferramentas para construir modelos preditivos capazes de auxiliar gestores na tomada de decisões estratégicas voltadas à redução da evasão escolar (Vasconcellos *et al.*, 2024).

Assim, a combinação entre estatística e ciência de dados aplicada à educação se mostra um caminho viável e promissor para desenvolver soluções que contribuam para a permanência dos estudantes, especialmente em cursos técnicos ofertados por instituições públicas.

2.3 Machine learning

Aprendizagem de máquina é um subcampo da ciência da computação focada em desenvolver algoritmos capazes de identificar padrões a partir de um grande volume de dados, também denominados dados de treinamento, com a finalidade de compreender determinados fenômenos (Ruback *et al.*, 2021).

Os algoritmos podem ser utilizados em várias áreas, como na medicina para identificar diagnósticos de doenças ou em bancos na previsão de clientes que podem deixar de utilizar os serviços. Na educação, no contexto de evasão escolar, é possível prever se um aluno irá abandonar o curso a partir dos dados de entrada(Machado et al., 2024).

O processo de desenvolvimento de algoritmo possui uma divisão em 3 fases: pré-processamento, treinamento e avaliação de modelo. A primeira fase consiste na organização dos dados, definição da pergunta de pesquisa e separar os dados de treinamento e teste. A segunda fase é o treinamento do algoritmo que pode ser supervisionada ou não supervisionada. A terceira fase é realizada a avaliação que compara os modelos com os dados de teste e gera os resultados (Paixão, 2022).

O aprendizado de máquina supervisionado o modelo é treinado com dados rotulados, em que as amostras são inseridas por seres humanos que já estão classificados. Ele é baseado na comparação entre os resultados que o modelo previu e os rótulos previamente classificados (Paixão, 2022).

Esse modelo pode ser aplicado ser aplicado para prever evasão de alunos ao utilizarem dados de alunos, que pode ser obtidos por meio de formulários, determinando uma variável alvo que pode ser uma situação de matrícula do aluno para identificar se ele está cursando, concluiu ou evadiu, é possível padronizar características de alunos que evadiram e analisar os fatos que contribuíram para este fenômeno.

2.3.1 Support Vector Machine (SVM)

Para Lorena e Carvalho, (2007) o *Support Vector Machine(SVM*) é descrito como um algoritmo de aprendizado de máquina supervisionado usado para classificação e regressão. Tendo como principal função de encontrar um hiperplano que separe as classes dos dados maximizando a margem entre elas.

Os algoritmos de SVM são utilizados em diversas áreas como problemas de classificação binária, reconhecimento de padrões em imagens médicas e faciais, detecção de spam e análise de textos para agrupamento, identificação de sequências de DNA e proteínas, além de detecção de fraudes financeiras e seguranças cibernéticas (Suthaharan *apud* Carvalho, 2025).

2.3.2 Random Forest (RF)

O método *Random Forest(RF)* consiste em um conjunto de Árvores de decisão, onde cada árvore depende dos valores de um vetor aleatório gerado de forma independente e com a mesma distribuição para todas as árvores (Breiman, 2011).

Uma Árvore de Decisão é um diagrama que contém as partes do nome da raiz que recebe todas as observações, os ramos que são nós internos e as folhas que contém a classificação final (Gutiérrez *et. al*., 2024).

O método foi utilizado para prever a evasão de alunos das Instituições de Ensino Superior com dados extraídos pelo INEP, foi o método mais eficaz obtendo uma taxa de acurácia em torno de 80% (Teodoro; Kappel, 2020).

2.3.3 K-Nearest Neighbors (KNN)

Neste método a classificação de um objeto é determinada por uma base de votos dos seus vizinhos mais próximos, com o objeto sendo atribuído à classe mais frequente entre seus *K* vizinhos mais próximos. O valor de *K* sendo um número inteiro positivo, relativamente pequeno e geralmente ímpar, a fim de evitar empates em classificações binárias (Altman *apud* Ferreira, 2019)

A técnica *K-Nearest Neighbors(KNN)* é um algoritmo que busca prever uma classe ou um valor para um novo ponto de dados com base nos K pontos mais próximos no conjunto de treinamento. É simples e eficaz, porém pode ter sua eficiência impactada dependendo do tamanho do conjunto de dados e a escolha correta do valor de K (Ferreira *et al.*, 2024).

Foi utilizado para prever tendências de sobrecarga dos recursos de infraestrutura de tecnologia da informação que dão suporte aos sistemas hospitalares. O KNN apresentou uma melhor acurácia na previsão de indisponibilidades, com 0,9603 (Luchtenberg *et al*., 2023)

2.3.4 Logistic Regression (LR)

A regressão logística é uma técnica estatística usada quando o objetivo é prever resultados que envolvem apenas duas possibilidades, como “sim” ou “não”. Diferente da regressão linear, que prevê valores numéricos contínuos, a regressão logística estima a chance de um evento acontecer, trabalhando com probabilidades que vão de 0 a 1. Por isso, é bastante utilizada em estudos que envolvem tomada de decisão com base em dados (Morais *et al*., 2021).

No campo da educação, essa técnica tem sido aplicada em pesquisas para identificar fatores que contribuem para situações como a evasão escolar. Ela permite avaliar quais variáveis influenciam mais a chance de um aluno abandonar os estudos, ajudando a criar estratégias de prevenção. Segundo (Morais *et al*. 2021), trata-se de um método eficiente por ser simples de interpretar e por permitir uma análise clara das influências de cada fator no resultado observado.

2.3.5 Gradient Boosting (GB)

Gradient Boosting (GB) é uma poderosa técnica de aprendizado de máquina que constrói um modelo preditivo combinando, sequencialmente, múltiplos modelos mais simples. Sua ideia central é focar nos erros (resíduos) dos modelos anteriores, adicionando novas previsões para corrigi-los e, assim, melhorar a precisão geral. De acordo com (Friedman, 2001), "o algoritmo constrói o modelo de forma aditiva, adicionando uma nova função a cada etapa que se ajusta aos resíduos da etapa anterior". Esse processo iterativo e aditivo permite que o algoritmo aprenda e se refine continuamente.

Este método foi utilizado para a predição de evasão no IFCE e obteve uma medida de 0.87 no campus de Quixadá. Em tese, significa que o *Gradient Boosting* possuiu maior precisão de acerto ao indicar o potencial de evasão de determinado aluno (Filho *et al*., 2020).

2.3.6 AdaBoost (AB)

AdaBoost (Adaptive Boosting) é um algoritmo fundamental de *ensemble learning* que combina múltiplos "classificadores fracos" para formar um "classificador forte". Proposto por Freund e Schapire (1996), ele funciona de forma iterativa: a cada etapa, o AdaBoost ajusta os pesos dos exemplos de treinamento, dando mais importância àqueles que foram classificados incorretamente. Isso permite que os classificadores subsequentes foquem nos erros, e os modelos com melhor desempenho recebem maior peso na decisão final. Esse processo adaptativo e focado nos erros torna o AdaBoost um método eficaz para problemas de classificação.

2.4 Trabalhos relacionados

Com o aumento da utilização de tecnologia de aprendizado de máquina em diversas áreas e o nível de importância do estudo do problema da evasão escolar, as seções abaixo discutem trabalhos relacionados a esse tema. Para a busca dos

trabalhos, foram utilizadas palavras-chave como “evasão escolar”, “*machine learning”*, “educação profissional” e “predição de evasão”. Na Seção 2.4.1 aborda trabalho de âmbito nacional, enquanto que a Seção 2.4.2 exibe trabalhos internacionais.

2.4.1 Trabalhos Nacionais

2.4.1.1 CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais.

O estudo do "CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais" de (Ramos *et al*., 2020) explora a adaptação da metodologia CRISP-DM, que é utilizada em diversas áreas para modelação de dados para cenários de dados educacionais. O modelo CRISP-EDM segue integralmente as seis etapas do modelo original, porém apresenta algumas diferenças em razão ao tipo de dados e do domínio educacional.

Os resultados das minerações utilizando o modelo CRISP-EDM oferece um padrão reconhecido amplamente na área e oferece oportunidades para que educadores e pesquisadores obtenham conhecimentos relevantes como modelagem do estudante, detecção de comportamentos, previsão de desempenho, detecção precoce de evasão, feedback e apoio ao estudante e professores. Também possibilitam a identificação de relações mais interessantes entre as variáveis em grande conjunto de dados.

2.4.1.2 Análise de Classificadores para Predição de Evasão de Campi de uma Instituição de Ensino Federal

No trabalho de Filho *et al*., (2020) foi apresentada uma análise da problemática da evasão e a construção de modelos de predição utilizando dados dos campi mais afetados do Instituto Federal do Ceará(IFCE). Os dados foram obtidos por meio da plataforma do IFCE, com alunos agregados por campus, por curso e por período em formato de planilhas eletrônicas. Para o pré-processamento dos dados foi utilizado a linguagem de programação Python e a biblioteca Pandas das quais foram possíveis a remoção de duplicidade, correção de termos,

preenchimento de valores nulos e, principalmente, o mapeamento de variáveis categóricas textuais em variáveis categóricas numéricas.

Os classificadores escolhidos foram *Naive Bayes* (NB), *Decision Tree* (DT), *Support Vector Machine* (SVM), *K-Nearest Neighbors* (KNN) e algoritmos de boosting, que consistem na composição de outros modelos, como por exemplo o *Gradient Boosting*.

Os resultados obtidos após a criação dos modelos e com a avaliação feita por eles, concluiu-se que os cinco campi com maiores taxas de evasão do IFCE, considerando dados de 2015 a 2019, são o campus Iguatu com 47,83%, campus Acopiara com 46,20%, Tianguá com 44,66% e, por fim, os campi Acaraú e Quixadá, ambos com 43,73% (Filho *et al*., 2020).

2.4.2 Trabalhos Internacionais

2.4.2.1 Machine Learning Predicts Upper Secondary Education Dropout as Early as the End of Primary School

O trabalho realizado por (Chung; Lee, 2019) aborda o desenvolvimento de modelos preditivos de aprendizado de máquina para prever alunos com maior potencial de evasão, os dados utilizados neste estudo foram do National *Education Information System* (NEIS) da Coreia do Sul. A amostra de aproximadamente 165.000 estudantes do Ensino Médio vindos de 4 cidades, incluíndo da capital da Coreia do Sul, do ano de 2014 foram extraídas. Por meio dos experimentos, os algoritmos de Random Forest mostraram uma excelente performance na predição de estudantes em termos de várias métricas de desempenho para classificação binária. Com os estudos foi possível demonstrar os benefícios do uso de Machine Learning com um grande volume de dados de estudantes na educação.

2.4.2.2 Application of the performance of machine learning techniques as support in the prediction of school dropout

O artigo de Gutiérrez, A.L.J. *et al*., (2024) apresenta um estudo que auxilia na predição de evasão escolar em instituições de ensino superior e ensino médio, comparando os modelos de técnicas de aprendizado de máquina. Os dados foram

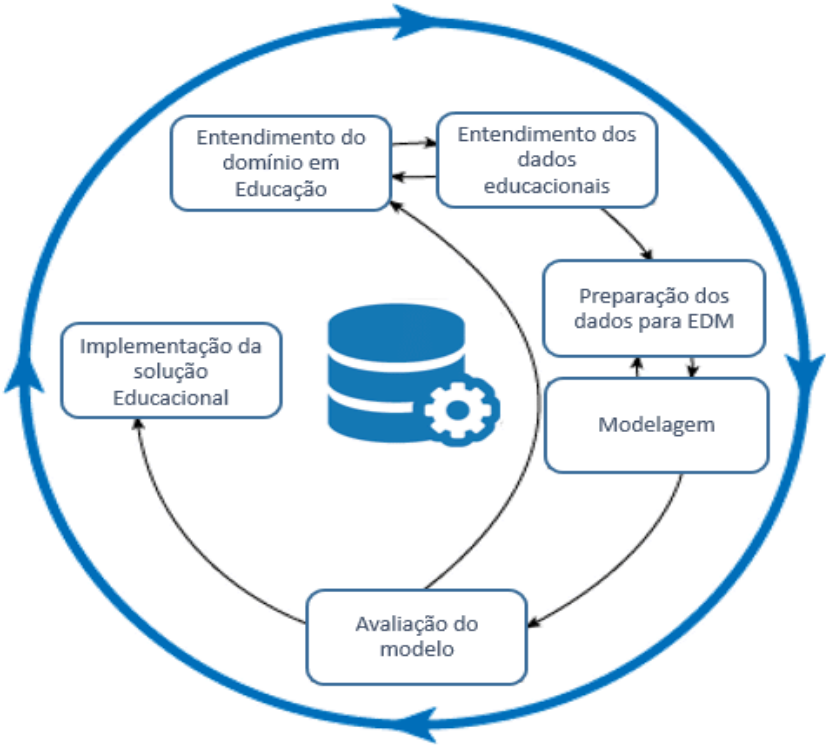
obtidos a partir da população de 32 estados do México, abrangendo pessoas com 15 anos ou mais de idade, as pesquisas foram realizadas em três períodos distintos: maio de 2010, março de 2015 e março de 2020. Após a limpeza dos dados, houve uma amostra de 1.080.782 registros no total. As técnicas aplicadas foram *Artificial Neural Networks(ANN), Support Vector Machines(SVM), Linear Ridge and Lasso Regression, Bayesian Optimization, Random Forest(RF*). O melhor resultado obtido após o desenvolvimento dos modelos foi a ANN no menor tempo com uma confiabilidade de 99%. Já os outros modelos conseguiram obter resultados com alta confiabilidade de mais de 91% .

3 METODOLOGIA

A área Mineração de Dados Educacionais (EDM – do inglês *Educational Data Mining*) é voltada para o desenvolvimento de métodos para análise de dados obtidos de ambientes educacionais. Assim, é possível identificar se o aluno está desmotivado e com isso personalizar os ambientes de ensino e melhorar as condições de aprendizagem (Baker *et al.*, 2011). Para a análise dos dados deste trabalho, a metodologia utilizada foi a *CRoss-Industry Standard Process for Educational Data Mining* (CRISP-EDM)(acrônimo de *CRoss-Industry Standard Process for Educational Data Mining*) uma adaptação do processo CRISP-DM (acrônimo de CRoss-Industry Standard Process for Data Mining) com objetivo de melhorar o contexto da mineração de dados educacionais (Ramos *et al*., 2020).

O CRISP-EDM é um processo que possui 6 etapas, que são adaptadas no contexto educacional. As fases não não possuem sequência rigorosa, sendo permitido o avanço e o retorno de etapas se necessário para melhor desenvolvimento do modelo proposto. A Figura 3 mostra o ciclo do modelo CRISP-EDM.

Figura 3 — Metodologia CRISP-EDM



Fonte: Jorge Ramos, Rodrigo Rodrigues, João Silva e Pamella Oliveira, (2020).

Na etapa 1 é feito o levantamento das problemáticas educacionais para buscar soluções por meio da mineração e a escolha das variáveis relevantes que serão coletadas no processo obtidas por meio de dados educacionais (Ramos *et al*., 2020). A etapa 2 é realizada uma análise inicial ou compreensão dos dados selecionados para a solução do problema educacional. A etapa 3 é realizada a preparação dos dados que normalmente envolve a junção das bases de dados, extração, transformação e limpeza das variáveis para a construção da tabela de análise (Ramos *et al*., 2020). A etapa 4 consiste na definição de técnicas de modelagem de dados e algoritmos de aprendizagem de máquina, iniciando com o parâmetro padrão e ajustando conforme os resultados preliminares obtidos, buscando aprimoramento nos resultados finais e comparação entre os algoritmos. Além disso, será realizada a análise da importância das variáveis (feature importance), com o objetivo de identificar quais atributos exercem maior influência na predição da evasão escolar, sendo esses resultados apresentados por meio de gráficos para facilitar a interpretação dos modelos. A etapa 5 prevê a avaliação dos modelos, considerando métricas de desempenho e também a interpretação dos resultados obtidos, incluindo a análise dos pesos ou importância das variáveis nos modelos preditivos. A etapa 6 é realizada a implementação do modelo, por meio de plugins ou dashboards que irão apresentar o resultado da aplicação dos modelos, a fim de auxiliar na tomada de decisões voltadas à melhoria dos processos educacionais.

3.1 Entendimento e coleta dos dados

Os dados utilizados neste trabalho foram disponibilizados pelo Registro Acadêmico do IFB Campus Brasília. Foi levantado o problema e a análise da problemática. A base possui um total de informação de 543 estudantes do (EMI) que foram matriculados de 2016.2 até 2023.2. Já a base do (TDS) o total é de 612 com matrículas desde o início do curso de 2016.2 até 2023.2.

A Tabela 1 exibe as variáveis que foram coletadas de cada estudante de ambos os cursos e que serão usadas para a realização da predição. Considerando

fatos sociais como sexo, etnia, renda per capita, cidade, dentre outras que foram levantadas para serem as variáveis preditoras que influenciam na evasão dos alunos.

Tabela 1 – Dados

Dado Bruto	Descrição
Nascimento	Data de nascimento do estudante
Sexo	Definição do gênero do discente, separado em masculino e feminino
Etnia	Grupo culturalmente homogêneo
Renda per capita familiar	Renda por membros da família
Cidade	Lugar geográfico onde o estudante reside
Necessidade Específica	Se o estudante é pessoa com deficiência
Forma de ingresso	Forma que o estudante entrou na instituição
Tipo de vaga	Se possui alguma cota
Curso	Curso que o estudante está matriculado
Situação da Matrícula	Status atual da matrícula
Ano de ingresso	Ano em que o estudante entrou na instituição
Semestre de ingresso	Semestre atual do estudante

Fonte: Dados da pesquisa, (2025).

A Tabela 2 mostra a variável a ser predita neste trabalho, possuía inicialmente os seguintes valores:

Tabela 2 – Dados de situação matrícula

Situação da Matrícula	Descrição
Ativa	Matrícula regular e válida
Trancada	Suspensão temporária do vínculo acadêmico
Cancelada	Interrupção do vínculo do aluno com a instituição
Evadida	Parou de frequentar a instituição
Formado	Concluiu o Curso
Transferido Externamente	Deixou a instituição de ensino para ingressar em outra instituição
Transferido Internamente	Transição que ocorre dentro da instituição

Fonte: Dados da pesquisa, (2025).

3.2 Pré-processamento

Nessa etapa, os dados foram tratados e padronizados para facilitar na utilização da metodologia CRISP-EDM. Ela é composta pela limpeza de valores nulos e padronização de dados que podem ser correlacionados, além dos processos de normalização para reduzir a redundância, e o balanceamento dos dados, processo que ajusta a distribuição dos dados de forma equilibrada. Para realizar a normalização dos dados foi necessário utilizar as técnicas de *OneHotEncoder*[1] para as colunas com variáveis categóricas e *LabelEncoder*[2] para a coluna com a variável alvo. Após a normalização das variáveis, foi realizada uma contagem dos valores das colunas “situação de matrícula” em ambos cursos. No curso EMI a variável “evadiu” se apresentou muito inferior da quantidade de “não evadiu”, sendo assim necessário realizar o balanceamento das variáveis com a técnica *oversampling*[3]*,* que consiste em duplicar ou criar dados artificiais para aumentar os valores da classe minoritária. No curso TDS houve um desbalanceamento nas variáveis de “evadiu” e “não evadiu”, sendo a classe “não evadiu” majoritária, por este motivo, foi utilizada a técnica *undersampling*[4]*,* ela diminui a quantidade dos dados da classe majoritária para balanceamento.

3.3 Ferramentas a serem utilizadas

[1] OneHotEnconder técnica de pré-processamento de dados da biblioteca *Scikit-learn* que transforma variáveis categóricas em variáveis numéricas para melhor interpretação do modelo. Disponível em <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html> Acessado em Novembro de 2025

[2] LabelEncoder técnica de pré-processamento de dados da biblioteca *Scikit-learn* utilizada para codificar os valores da variável alvo. Disponível em <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.LabelEncoder.html> Acessado em Novembro de 2025.

[3] Oversampling técnica de pré-processamento de dados para balanceamento de variáveis aumentando instâncias de classes minoritárias. Disponível em <https://imbalanced-learn.org/stable/over_sampling.html> Acessado em Novembro de 2025.

[4] Undersampling técnica de pré-processamento de dados para balanceamento de variáveis aumentando instâncias de classes majoritárias. Disponível em <https://imbalanced-learn.org/stable/under_sampling.html> Acessado em Novembro de 2025.

As ferramentas utilizadas para a limpeza e tratamento desses dados foi a linguagem de programação *Python*[5] e suas bibliotecas *Numpy*[6] *e Pandas*[7]*.* Essas ferramentas serão fundamentais para a padronização dos dados e auxiliar no entendimento. O *Scikit-learn*[8] é uma biblioteca criada em linguagem *Python* destinada a práticas e desenvolvimento de técnicas de aprendizado de máquina. Para encontrar os melhores hiperparâmetros dos modelos de aprendizados de máquina desenvolvidos para prever a predição, foi aplicado a técnica *GridSearch*[9]. Além disso, foi utilizado também a técnica *k-fold cross-validation*[10] que avalia o desempenho dos modelos de aprendizado de máquina.

3.4 Treinamento e teste

Os algoritmos escolhidos para etapa de treinamento dos modelos, foram o *Gradient Boosting* (GB), que obteve melhor precisão na previsão no trabalho de evasão realizado por *(*Filho *et al., 2020),* o *Random Forest* (RF) que também apresentou bons resultados por (Teodoro; Kappel, 2020). Além desses, também foram utilizados os algoritmos *Support Vector Machine* (SVM) , *Logistic Regression* (LR), *K-nearest Neighbors* (KNN) e *AdaBoost* (AB). Esses algoritmos foram escolhidos por serem frequentemente utilizados para previsão de evasão como nos trabalhos de (Dutra; Souza; Fernandes, 2022; Sousa *et al*., 2025; Oliveira; Medeiros, 2024) e por seus resultados satisfatórios. Essa preparação seguiu o padrão de

---

[5] Python é uma linguagem de programação de alto nível, amplamente utilizada para desenvolvimento de aplicações web, automação, análise de dados e aprendizagem de máquina. Disponível em <https://www.python.org/> Acessado em Julho de 2025..

[6] NumPy é uma biblioteca fundamental para computação científica em Python, oferecendo suporte para arrays e matrizes, e também diversas funções matemáticas. Disponível em <https://numpy.org/> Acessado em Julho de 2025.

[7] Pandas é uma biblioteca de software escrita em Python que fornece estruturas de dados e ferramentas de análise de dados, e facilita o trabalho com dados tabulares e séries temporais. Disponível em <https://numpy.org/> Acessado em Julho de 2025.

[8] Sckit-lean é uma biblioteca da linguagem Python utilizada para aplicação prática de aprendizado de máquina. Disponível em <https://scikit-learn.org/stable/> Acessado em Julho de 2025

[9] Grid Search é um algoritmo convencional de ajuste de hiperparâmetros que testa exaustivamente todas as combinações possíveis dos valores fornecidos permitindo encontrar o melhor desempenho do modelo. Disponível em <https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html>, Acessado em Julho de 2025.

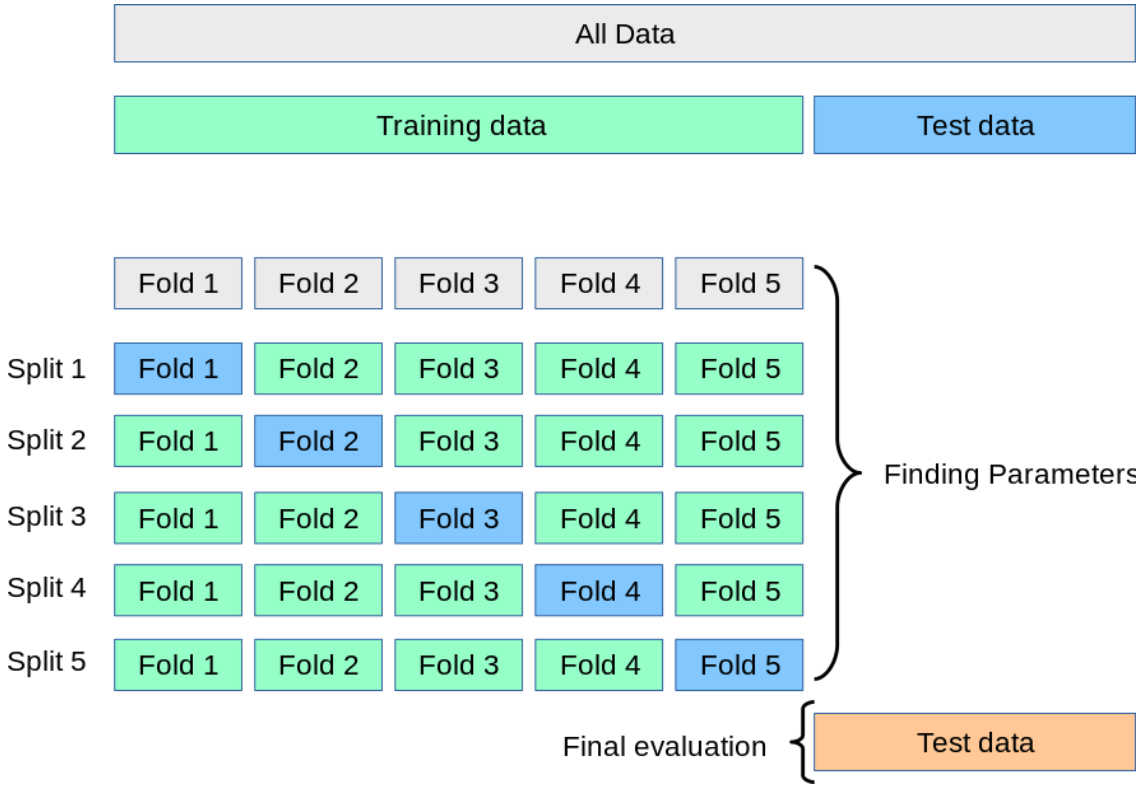
[10] K-fold Cross Validation é uma técnica utilizada para avaliar o desempenho dos modelos de machine learning. Disponível em <https://scikit-learn.org/stable/modules/cross_validation.html> Acesso em Fevereiro 2026

dividir as bases em conjuntos de treinamento e teste com 70% para treinamento e 30% para teste

A validação cruzada é uma técnica para avaliar o desempenho e precisão de um modelo com dados não vistos, ajudando a verificar a capacidade de generalização do modelo diante de um novo conjunto de dados (Oliveira; Medeiros, 2024).

A versão de validação utilizada para a geração dos modelos foi uma variação da técnica *K-fold*, com o nome de *Stratified k-Fold Cross-Validation*. Segundo (Macedo; Saporetti, 2024), o *k-fold* consiste na divisão dos dados disponíveis, que contêm N amostras, em K subconjuntos, sendo o valor de K maior que 1 (K>1). Após a partição, K−1 dos subconjuntos gerados são utilizados para o treinamento do modelo, enquanto o subconjunto restante é utilizado para à etapa de teste. Todo esse processo é repetido K vezes, garantindo que em cada iteração um conjunto de teste diferente seja empregado. Ao final, cada um dos conjuntos é avaliado por métricas pré-estabelecidas (Seção 3.5), e a estimativa final para cada métrica é obtida pela média aritmética dos K valores gerados. A vantagem da utilização do *Stratified k-Fold Cross-Validation* é que, os dados de cada grupo não são divididos aleatoriamente, e sim selecionados de maneira que mantenha a mesma proporção de classes presentes no conjunto de dados original. Assim, cada processo permanece com o mesmo balanceamento dos dados originais (Teodoro; Kappel, 2020). A Figura 4 ilustra o método K-fold cross-validation considerando K = 5.

Figura 4 — K-fold Cross Validation



Fonte: Ajitesh Kumar, (2024).

3.5 Métricas para desempenho de classificação

Nesta seção é proposta as etapas de avaliação dos modelos gerados, das quais foram categorizadas na seguinte matriz de confusão[11] demonstrada na tabela trata-se de uma matriz 2x2 onde cada uma das células armazenam, respectivamente:

- **Verdadeiro Positivo (VP):** A quantidade de casos que pertenciam à classe positiva e foram corretamente previstas como positivas pelo modelo.
- **Falso Positivo(FP):** A quantidade de casos que o modelo previu incorretamente a classe positiva, porém na verdade era negativa.
- **Verdadeiro Negativo (VN):** A quantidade de casos que realmente eram negativos e foram corretamente previstas como negativas.

[11] Matriz de confusão: Método de visualização de resultados por meio de uma tabela com colunas que representam os resultados reais dos resultados previstos para avaliar o desempenho do algoritmo classificador. Disponível em <https://scikit-learn.org/stable/modules/generated/sklearn.metrics.confusion_matrix.html> Acessado em novembro de 2025.

- **Falso Negativo (FN):** São os casos que eram positivos, mas o modelo previu como negativas.

Tabela 3 - Matriz de confusão utilizada na avaliação dos modelos.

	Não evadiu	**Evadiu**
Não evadiu	Verdadeiro negativo (VN)	Falso positivo (FP)
Evadiu	Falso negativo (FN)	Verdadeiro positivo (VP)

Fonte: Elaborado pelas autoras, (2025).

A partir da matriz de confusão, foram utilizadas as seguintes métricas para avaliação dos modelos acurácia, precisão, revocação *e F1-score*.

- A acurácia (Eq.1) indica a proporção de acertos totais do modelo.

$$A = \frac{Vp + Vn}{Vn + Fn + Vp + Fp} \quad (1)$$

- A precisão (Eq.2) mensura a quantidade de casos corretamente classificados como evasão entre todos os classificados como tal.

$$P = \frac{Vp}{Vp + Fp} \quad (2)$$

- A revocação (Eq.3) avalia a capacidade do modelo em identificar corretamente os alunos evadidos.

$$R = \frac{Vp}{Vp + Fn} \quad (3)$$

- A especificidade (Eq.4) mede a capacidade do modelo em identificar verdadeiros negativos.

$$E = \frac{Vn}{Vn + Fp} \quad (4)$$

- O F1-score (Eq.5) representa a média harmônica entre precisão e revocação, sendo especialmente útil em casos de desbalanceamento entre as classes.

$$F = 2^* \frac{Precisão * Revocação}{Precisão + Revocação} \quad (5)$$

Também será gerada a curva ROC (*Receiver Operating Characteristic*) e calculada a AUC (*Area Under the Curve*), métricas essenciais utilizadas para avaliação de desempenho de modelos de classificação. A curva ROC é uma curva em representação gráfica das métricas de desempenho de um modelo, se trata da construção da probabilidade em um gráfico que relaciona o eixo Y (vertical) a taxa de verdadeiros positivos (TPR) que possui o mesmo valor da revocação e o eixo X (horizontal) a taxa de falsos positivos (FPR) que é o complemento da especificidade (1 − E) (Filho *et al*., 2020). A AUC é a área total abaixo da curva ROC que funciona como uma métrica de separação, que reflete a capacidade do modelo distinguir entre classes positivas e negativas.

O valor da AUC varia de 0 a 1. Valores próximos a 1 indicam um desempenho excelente, que mostra que o modelo tem alta capacidade de prever corretamente as amostras em suas respectivas classes. Valores de 0.5 significam que o modelo não apresenta capacidade de discriminação se comportando como um classificador aleatório. Valores de 0 o modelo revela classificações totalmente equivocadas (Sousa, *et al.* 2025).

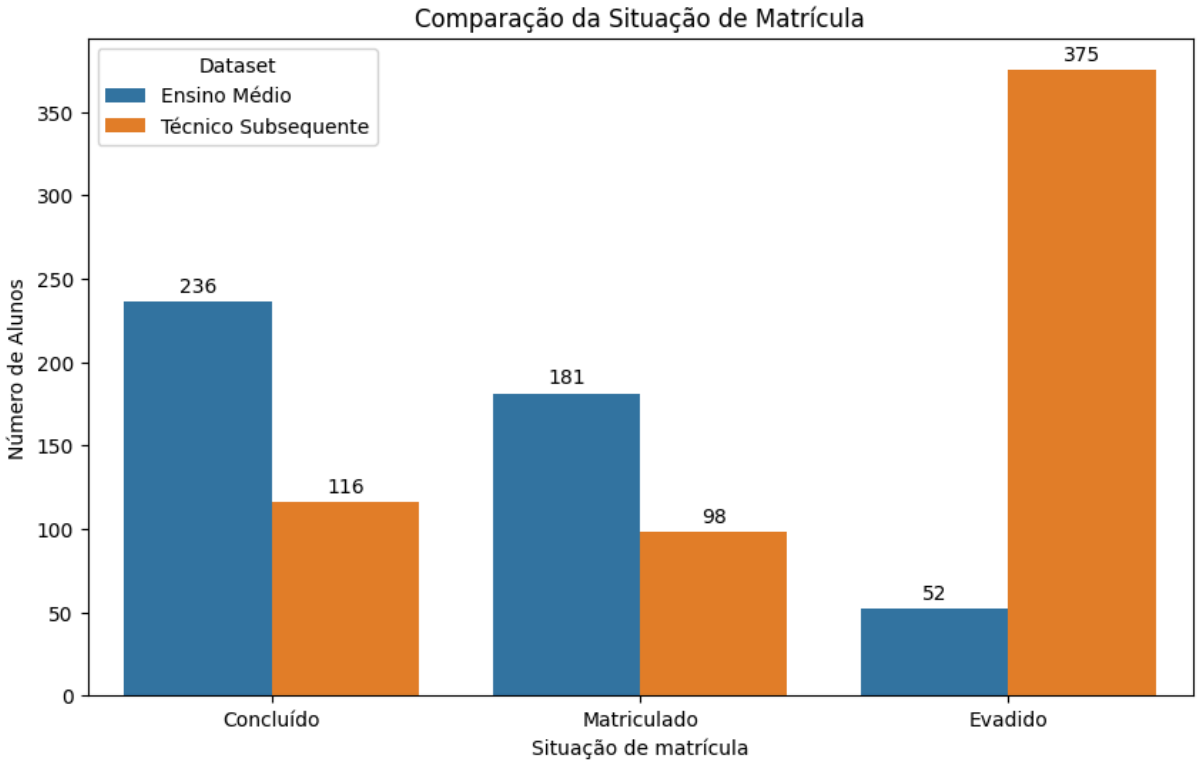
Todas as métricas foram calculadas com base no conjunto de teste, utilizando as funções da biblioteca Scikit-learn (Pedregosa *et al.*, 2011), de modo a garantir uma avaliação fiel à capacidade de generalização dos modelos.

4 RESULTADOS E DISCUSSÃO

4.1 Caracterização da amostra

Ao realizar uma análise exploratória dos dados dos alunos cadastrados no sistema do registro acadêmico do campus Brasília no período de 2016.2 até 2023.2 foi possível traçar o perfil demográfico e situação de matrícula dos estudantes. A figura 5 mostra um gráfico sobre a situação da matrícula dos cursos Técnico em Informática Integrado ao Ensino Médio (EMI) e Técnico Subsequente em Desenvolvimento de Sistemas (TDS). Em relação ao status da matrícula, 236 dos alunos do curso de EMI e 116 do TDS concluíram o curso. O curso de TDS possui 375 alunos evadidos, enquanto o EMI tem 52 alunos evadidos. É possível observar no gráfico que a taxa de evasão dos alunos do TDS é mais que 7 vezes do EMI.

Figura 5 - Situação de matrícula

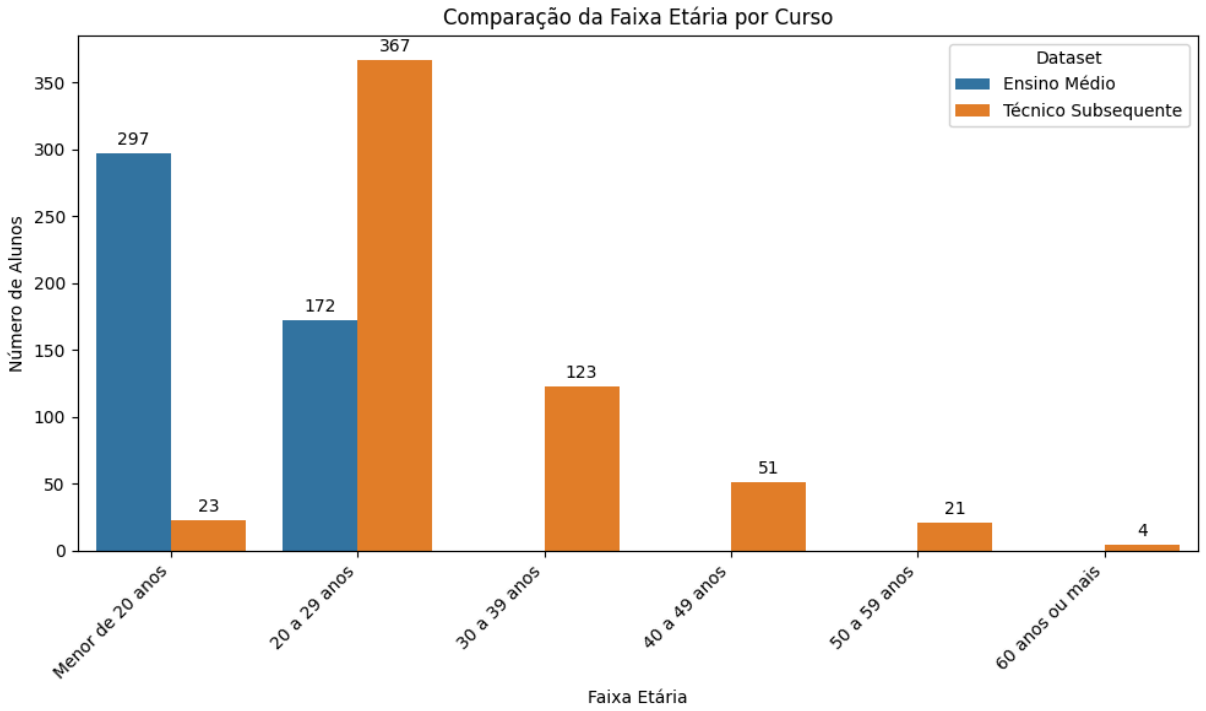


Fonte: Elaborado pelas autoras, (2025).

Os indicadores escolhidos para a análise exploratória foram: Faixa etária, gênero, etnia, localização, renda per capita. As figuras 6 e 7 exibem os indicadores de faixa etária e de gênero, no curso de EMI 297 alunos são menores de 20 anos e

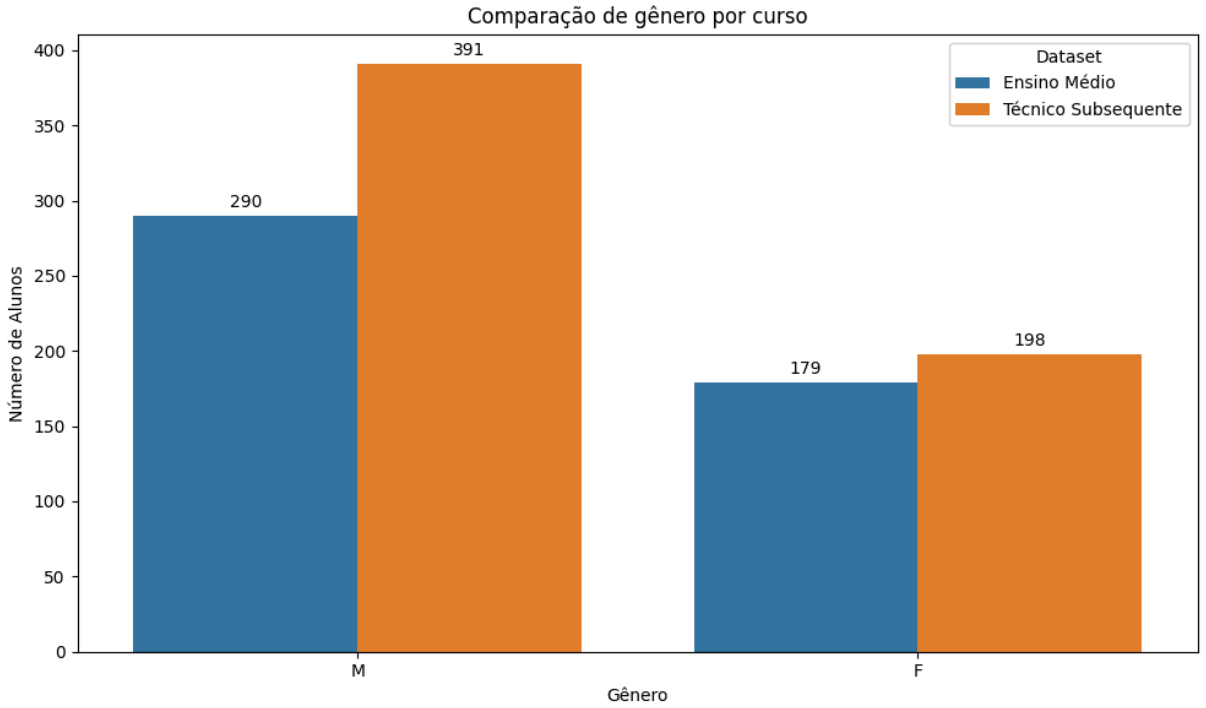
no curso de TDS 367 têm idade entre 20 a 29 anos, representando um público mais jovem. Já somados com 681 os cursos possuem maioria do sexo masculino.

Figura 6 - Distribuição por faixa etária.



Fonte: Elaborado pelas autoras, (2025).

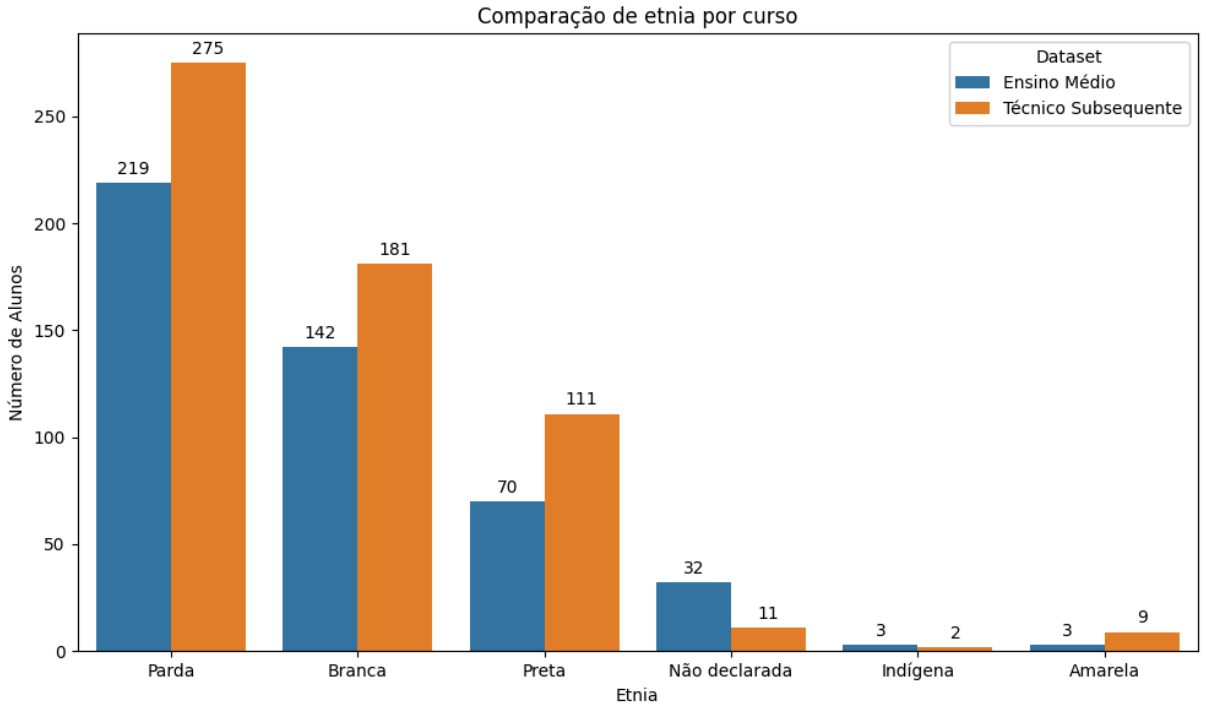
Figura 7 - Distribuição por gênero



Fonte: Elaborado pelas autoras, (2025).

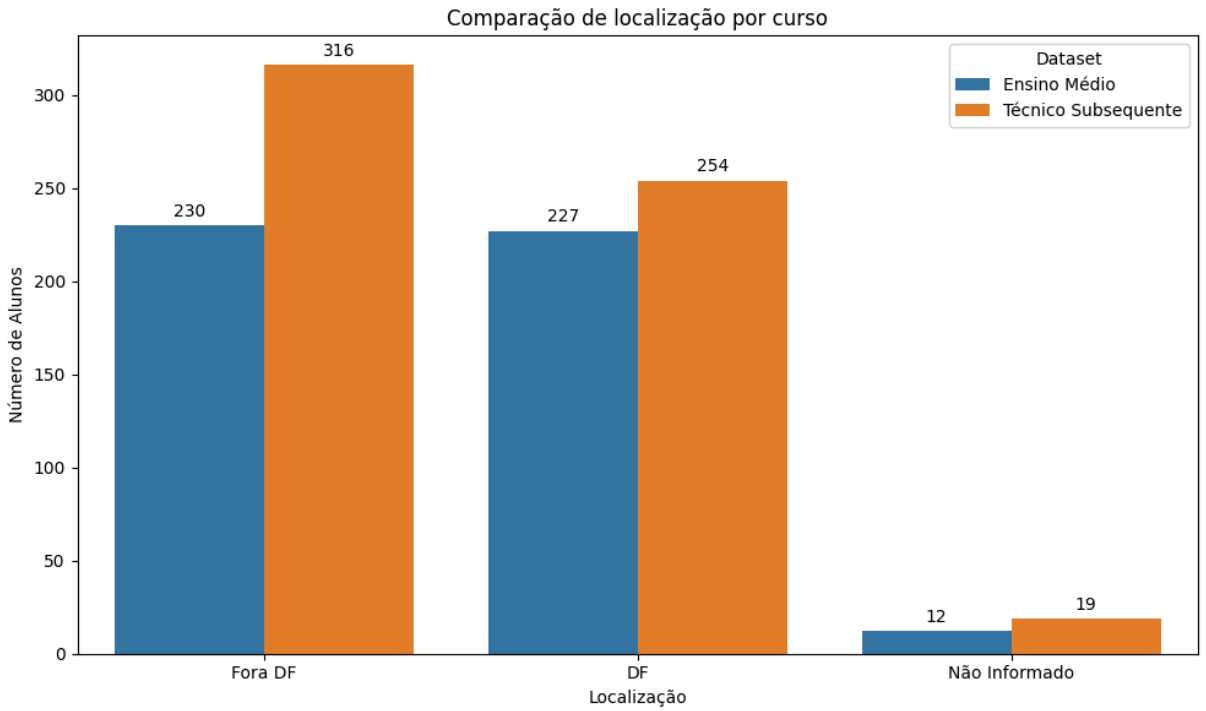
Na figura 8 e 9 são os indicadores de etnia e localização. Em ambos os cursos a maioria dos estudantes são pardos, e moram fora do Distrito Federal.

Figura 8 - Distribuição por etnia.



Fonte: Elaborado pelas autoras, (2025).

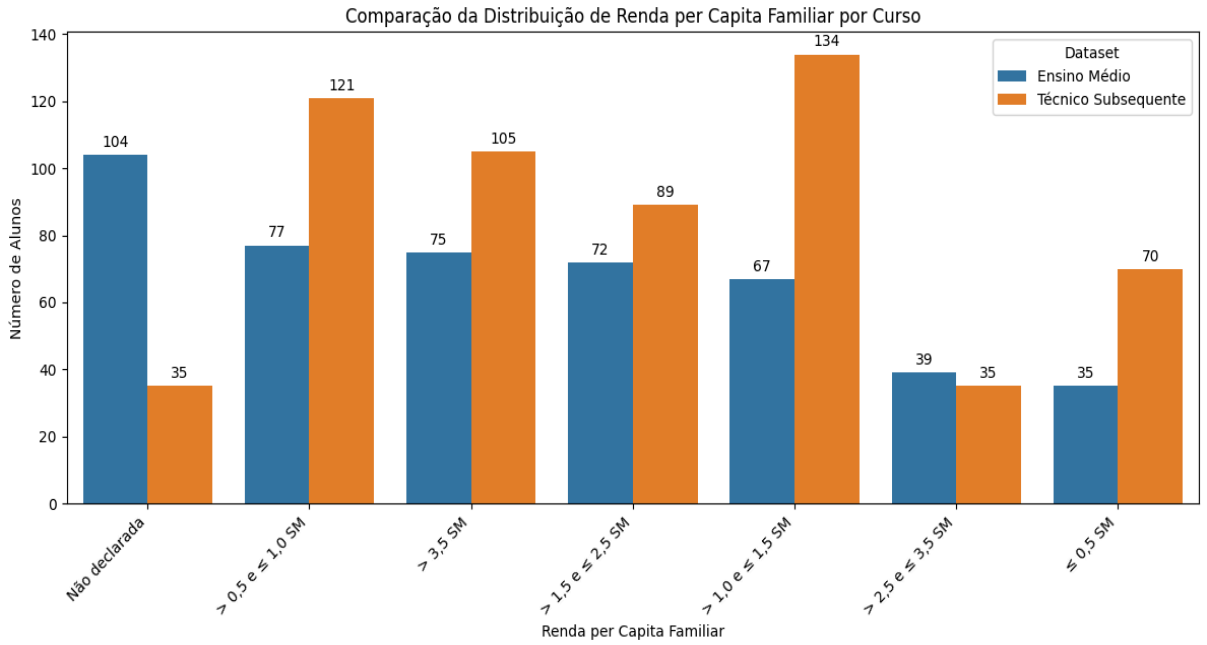
Figura 9 - Distribuição por etnia e localização.



Fonte: Elaborado pelas autoras, (2025).

A Figura 10 apresenta a distribuição dos estudantes evadidos dos cursos Ensino Médio Integrado (EMI) e Técnico Subsequente (TDS), considerando as faixas de renda per capita familiar. Observa-se que, no curso EMI, a maioria dos evadidos está concentrada nas faixas de renda mais baixas. Já no curso TDS, o maior número de discentes evadidos encontra-se nas faixas de renda superiores, especialmente entre 1,0 e 1,5 salários mínimos.

Figura 10 - Distribuição por renda



Fonte: Elaborado pelas autoras, (2025).

Após análise dos gráficos, é possível traçar o perfil dos alunos de ambos os cursos, no curso de Técnico em Informática Integrado ao Ensino Médio (EMI) a maioria dos alunos são menores de 20 anos, do sexo masculino, pardos, moram fora do Distrito Federal e possuem renda não declarada. No Técnico Subsequente em Desenvolvimento de Sistemas os estudantes possuem idade entre 20 a 29 anos, do sexo masculino, pardos, moram fora do Distrito Federal e possuem renda entre 1,0 a 1,5 salários mínimos.

4.2 Resultados após o pré-processamento dos dados

Diante dessas características apresentadas e buscando entender quais fatores mais influenciam a decisão de permanência ou evasão, foram aplicados modelos de predição. Para isso, realizou-se a preparação dos dados, que inclui o tratamento de valores faltantes, a padronização das variáveis e o balanceamento das classes utilizando *oversampling* e *undersampling*. Também foi realizada a remoção de colunas irrelevantes e restando apenas as colunas "Sexo", "Etnia", "Ano de Ingresso", "Faixa_etaria", "Localizacao", "Situacao evasao", "Renda per Capita (Faixa)", "Cotista".

Isso garantiu que o balanceamento fosse feito apenas nos dados de treino, evitando que informações do teste vazassem para o modelo.

No caso do Técnico em Informática Integrado ao Ensino Médio a divisão inicial (treino, validação e teste) resultou em um conjunto de treinamento com 328 registros, sendo 141 dados de teste. A análise mostrou que o desequilíbrio continuava: eram 292 registros da classe majoritária ('Não Evadido') e apenas 36 da minoritária ('Evadido'). Para corrigir isso, usou-se a técnica de Oversampling (SMOTE). O algoritmo criou dados sintéticos para a classe menor até que ela ficasse igual à maior, resultando em um conjunto de treino com 584 registros (292 para cada classe).

Já no Técnico subsequente em Desenvolvimento de Sistemas o conjunto de treinamento começou com 412 registros, sendo 177 dados de teste. Aqui a situação era inversa: a evasão (somada aos cancelamentos) era a maioria, com 256 exemplos, enquanto a permanência tinha 156 registros. Por causa dessa grande quantidade de evasão, aplicou-se o Undersampling, reduzindo aleatoriamente a classe maior até igualar a quantidade da menor. O resultado foi uma base de treinamento equilibrada com 312 registros, sendo exatamente 156 para cada lado.

Com os dados prontos, foram avaliados os modelos *Gradient Boosting* (GB), *Random Forest*, *Support Vector Machine* (SVM), *Logistic Regression* (LR), *K-nearest Neighbors* (KNN) e *AdaBoost* (AB). Para medir o desempenho de cada um, foram utilizadas as métricas de acurácia, precisão, revocação e *F1-score*.

4.3 Avaliação de Desempenho - Técnico subsequente em Desenvolvimento de Sistemas

A partir das análises realizadas, os seis modelos foram avaliados com base nas métricas apresentadas anteriormente, permitindo comparar o desempenho de cada algoritmo na tarefa de previsão da evasão estudantil. A Tabela 4 exibe os resultados de acertos e erros dos modelos, organizados por meio de suas respectivas matrizes de confusão.

Tabela 4 – Matrizes de confusão sem *GridSearch*

	NÃO EVADIU	EVADIU
NÃO EVADIU	KNN: 52 LR: 46 RF: 45 SVM:37 GB:44 AB:34	KNN:6 LR: 12 RF: 13 SVM:21 GB:14 AB:24
EVADIU	KNN: 66 LR: 47 RF:42 SVM:37 GB:45 AB:21	KNN:53 LR: 72 RF:77 SVM:82 GB:74 AB:98

Fonte: Elaborado pelas autoras, (2025).

A avaliação dos modelos permitiu observar como cada algoritmo se comportou na tarefa de prever a evasão. O primeiro modelo analisado foi o *K-Nearest Neighbors* , que apresentou o pior desempenho entre os seis. Apesar de ter acertado a maior parte das previsões dos alunos que permaneceram no curso (Verdadeiros Negativos – VN = 52), ele registrou um número muito alto de Falsos Negativos (FN = 66). Isso significa que muitos estudantes que realmente evadiram foram classificados como se fossem continuar, o que tornaria seu uso inviável na prática, já que a instituição deixaria de identificar vários alunos em risco e não conseguiria agir preventivamente.

O modelo de *Logistic Regression* teve um resultado intermediário. Ele acertou 46 alunos que não evadiram (VN) e 72 alunos que realmente evadiram (VP). Mesmo assim, ainda apresentou um número grande de Falsos Negativos (FN = 47), mostrando dificuldade em identificar corretamente todos os alunos que iriam evadir. Ou seja, embora tenha se saído melhor que o KNN, ainda não atende às necessidades reais do contexto.

O *Random Forest* obteve um desempenho melhor em comparação aos anteriores, alcançando 77 Verdadeiros Positivos (VP) e diminuindo os Falsos Negativos para 42. Esses resultados mostram que o modelo consegue capturar melhor os padrões relacionados à evasão e distinguir com mais precisão os alunos em risco. Apesar de ainda errar bastante, seus resultados foram mais robustos e já demonstram potencial de uso.

O *Gradient Boosting* teve um desempenho ainda mais sólido, com 74 Verdadeiros Positivos (VP) e 45 Falsos Negativos (FN). Esse modelo consegue separar bem as classes e identificar de forma mais eficiente quem realmente iria evadir, apresentando um equilíbrio maior entre acertos e erros.

O *Support Vector Machine* apresentou resultados ainda mais expressivos, alcançando 82 Verdadeiros Positivos (VP) e apenas 37 Falsos Negativos (FN). Isso mostra que o modelo foi bem eficaz em aprender com os erros e identificar os padrões mais importantes relacionados à evasão. Dessa forma, demonstrou bastante consistência e alinhamento com o objetivo central do estudo.

Já o modelo *AdaBoost* foi o modelo que apresentou o melhor desempenho geral, atingindo 98 Verdadeiros Positivos (VP) e o menor número de Falsos Negativos (FN = 21). Esse resultado mostra que o algoritmo foi o mais eficiente na identificação de alunos em risco, o que é essencial para que a instituição possa agir antes que a evasão aconteça.

De forma geral, ao comparar os seis modelos avaliados, foi possível perceber diferenças significativas no desempenho de cada um na previsão da evasão. Enquanto o KNN e a *Logistic Regression* tiveram resultados mais fracos, apresentando muitos erros críticos na identificação dos alunos que realmente evadiram, os modelos baseados em árvores, como o *Random Forest* e o *Gradient Boosting*, mostraram maior capacidade de aprendizado e redução desses erros. O SVM também apresentou um bom desempenho, destacando-se pelo equilíbrio entre acertos e erros. No entanto, o *AdaBoost* demonstrou resultados superiores em relação aos demais, alcançando o maior número de acertos na identificação de alunos em risco e o menor número de falsos negativos.

Agora com a aplicação do *GridSearch*, foi possível otimizar os hiperparâmetros dos modelos e observar mudanças relevantes em seu desempenho

preditivo. Como resultado da otimização via *GridSearch*, o algoritmo KNN foi configurado com *K* = 9, métrica de distância Euclidiana e pesos ponderados pela distância. Com essa configuração, o modelo apresentou uma leve melhora em relação ao treinamento padrão, alcançando 52 Verdadeiros Negativos (VN) e 59 Verdadeiros Positivos (VP). Apesar disso, diminuiu a quantidade de Falsos Negativos (FN = 60), evidenciando que, mesmo priorizando os vizinhos mais próximos (pesos por distância), o algoritmo continua com dificuldade em identificar os alunos que realmente se evadiram. Assim, o KNN permanece pouco adequado para a resolução deste problema.

Tabela 5 – Matrizes de confusão após a aplicação do Grid Search

	NÃO EVADIU	EVADIU
NÃO EVADIU	KNN: 52 LR: 48 RF: 42 SVM:37 GB:36 AB:30	KNN:6 LR: 10 RF:16 SVM:21 GB:22 AB:28
EVADIU	KNN: 60 LR: 43 RF:35 SVM:37 GB:29 AB:16	KNN:59 LR: 76 RF:84 SVM:82 GB:90 AB:103

Fonte: Elaborado pelas autoras, (2025).

A otimização da *Logistic Regression* definiu o uso da penalidade L2 e o valor *C*=10. Esse ajuste ajudou o modelo a filtrar melhor os dados, resultando em 48 Verdadeiros Negativos (VN) e 76 Verdadeiros Positivos (VP). Houve uma redução importante nos Falsos Negativos (FN = 43) em comparação ao teste inicial, mostrando que o modelo passou a identificar melhor os padrões de evasão. Mesmo com essa melhora, seu desempenho geral ainda é mediano quando comparado aos modelos mais complexos

O *Random Forest* teve um avanço expressivo com o *GridSearch*, que definiu o uso de 150 árvores (n_estimators) e profundidade máxima de 10 (max_depth). Com essa configuração, ele alcançou 84 Verdadeiros Positivos (VP) e reduziu os Falsos Negativos (FN) para 35. Esse resultado reforça a capacidade do modelo em capturar

melhor as relações entre as variáveis após a otimização, tornando-se uma opção mais forte tecnicamente.

O modelo *Support Vector Machine* foi um dos que mais evoluiu com o *GridSearch*, que selecionou o kernel RBF, com regularização *C*=1 e gamma *scale*. Com essa configuração, ele permaneceu com 82 Verdadeiros Positivos (VP) e 37 Falsos Negativos (FN). Comparado aos resultados sem otimização, houve permaneceu neutro na identificação de alunos que realmente evadiram. Dessa forma, o SVM se torna um dos modelos mais eficientes dentro do conjunto analisado.

O *Gradient Boosting* manteve um desempenho elevado após o *GridSearch*, que configurou o modelo com 50 estimadores, profundidade máxima de 3 e taxa de aprendizado de 0.01. Ele alcançou 90 Verdadeiros Positivos (VP) e reduziu os Falsos Negativos (FN) para 29. Embora tenha apresentado leve melhora em relação ao treinamento sem otimizar, seu desempenho final continua alto e consistente, reforçando a capacidade do método em aprender de forma sequencial e corrigir erros

Por fim, o *AdaBoost* , que já era o modelo com melhor desempenho na análise anterior, manteve sua posição após o *GridSearch*, que definiu uma taxa de aprendizado de 0.01 com 100 estimadores. Com essa configuração, ele alcançou 103 Verdadeiros Positivos (VP) e reduziu os Falsos Negativos (FN) para 16. Apesar de o *Gradient Boosting* ter mostrado melhora mais acentuada com a otimização, o *AdaBoost* ainda apresenta excelente equilíbrio e estabilidade. Isso demonstra que o modelo continua altamente eficiente e bastante confiável para uso prático.

Após essa análise também é importante a avaliação dos modelos preditivos a partir de métricas como acurácia, precisão, revocação (sensibilidade), especificidade e *F1-score* é fundamental para compreender não apenas o desempenho geral, mas também a capacidade real de cada algoritmo em identificar corretamente alunos evadidos e não evadidos, como mostra na tabela 6. Essas métricas permitem observar o equilíbrio entre erros e acertos, especialmente em um cenário de desequilíbrio de classes e alta relevância prática da classe “evasão”. Dessa forma, torna-se possível identificar qual modelo é mais adequado para apoiar a tomada de decisão no contexto educacional.

Depois do uso do *GridSearch*, observou-se uma melhora geral nos desempenhos, indicando que a otimização dos hiperparâmetros contribuiu significativamente para a eficiência preditiva dos modelos. O *Support Vector Machine* e o *Gradient Boosting* demonstraram ganhos relevantes em revocação, alcançando 0.7657 e 0.6760, respectivamente, o que significa maior capacidade de identificar corretamente alunos que realmente evadiram, um aspecto crucial para o objetivo do estudo. Entretanto, o *AdaBoost* apresentou o melhor resultado final ao analisar o conjunto das métricas, destacando-se especialmente pela maior revocação (0.8511) e pelo maior F1-score (0.7624), o que demonstra forte equilíbrio entre precisão e capacidade de captura da classe minoritária. O aumento da acurácia geral também reforça sua alta estabilidade após a otimização.

Comparativamente, algoritmos como KNN e *Logistic Regression* também apresentaram melhorias após o *GridSearch*, mas ainda permaneceram abaixo dos modelos baseados em árvores e em margem ótima quanto à capacidade de identificação da evasão. Assim, considerando o cenário completo de avaliação, o *AdaBoost* mostrou o melhor desempenho entre os modelos testados, tornando-se a opção mais indicada para aplicações práticas de detecção de risco de evasão estudantil.

Tabela 6 – Resultados dos modelos com *GridSearch*

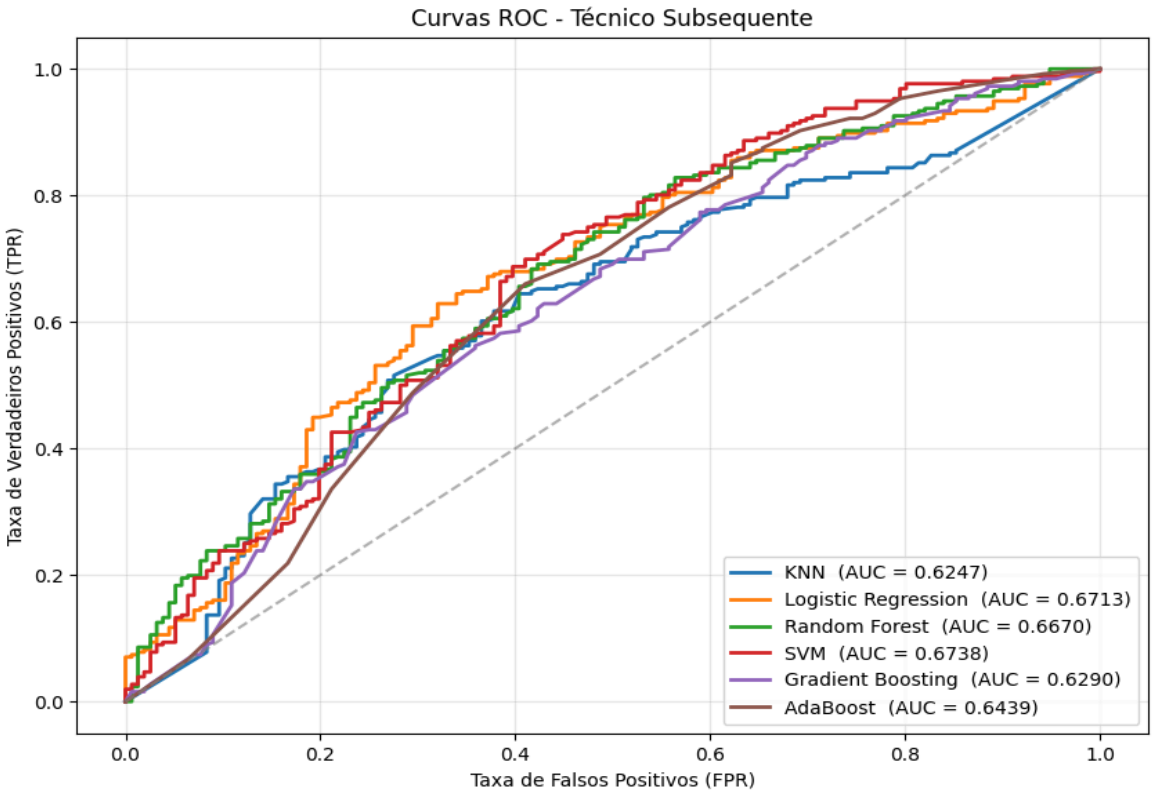
Classificador	Acurácia	Precisão	Revocação	Especificidade	F1-score
Kneighbors	0.5944	0.7540	0.5194	0.7179	0.6109
Logistic Regression	0.6504	0.7571	0.6446	0.6603	0.6959
Random Forest	0.6186	0.7220	0.6327	0.5962	0.6722
Support Vector Machine	0.6648	0.7165	0.7657	0.5000	0.7399
Gradient Boosting	0.6140	0.6986	0.6760	0.5128	0.6853

Adaboost	0.6722	0.6936	**0.8511**	0.3782	**0.7624**

Fonte: Elaborado pelas autoras, (2025).

Além das métricas já apresentadas, também foi analisada a curva AUC-ROC para avaliar a capacidade de separação dos modelos otimizados por Grid Search. A Figura 11 apresenta as curvas ROC de cada algoritmo, permitindo observar o equilíbrio entre a taxa de verdadeiros positivos e a taxa de falsos positivos. Os resultados mostraram desempenhos relativamente próximos entre os classificadores, com destaque para o Support Vector Machine (AUC = 0,6738) e a Regressão Logística (AUC = 0,6713), que obtiveram as melhores capacidades discriminativas. Na sequência aparecem o Random Forest (AUC = 0,6670), o Gradient Boosting (AUC = 0,6290) e o KNN (AUC = 0,6247). Já o AdaBoost (AUC = 0,6439) apresentou o menor valor de área sob a curva, demonstrando menor capacidade de separação entre as classes nesse cenário específico. Embora tenha apresentado bons resultados em outras métricas, o AdaBoost obteve o menor valor de AUC, indicando menor capacidade de distinção entre alunos evadidos e não evadidos em múltiplos limiares decisórios.

Figura 11 - Curva AUC-ROC

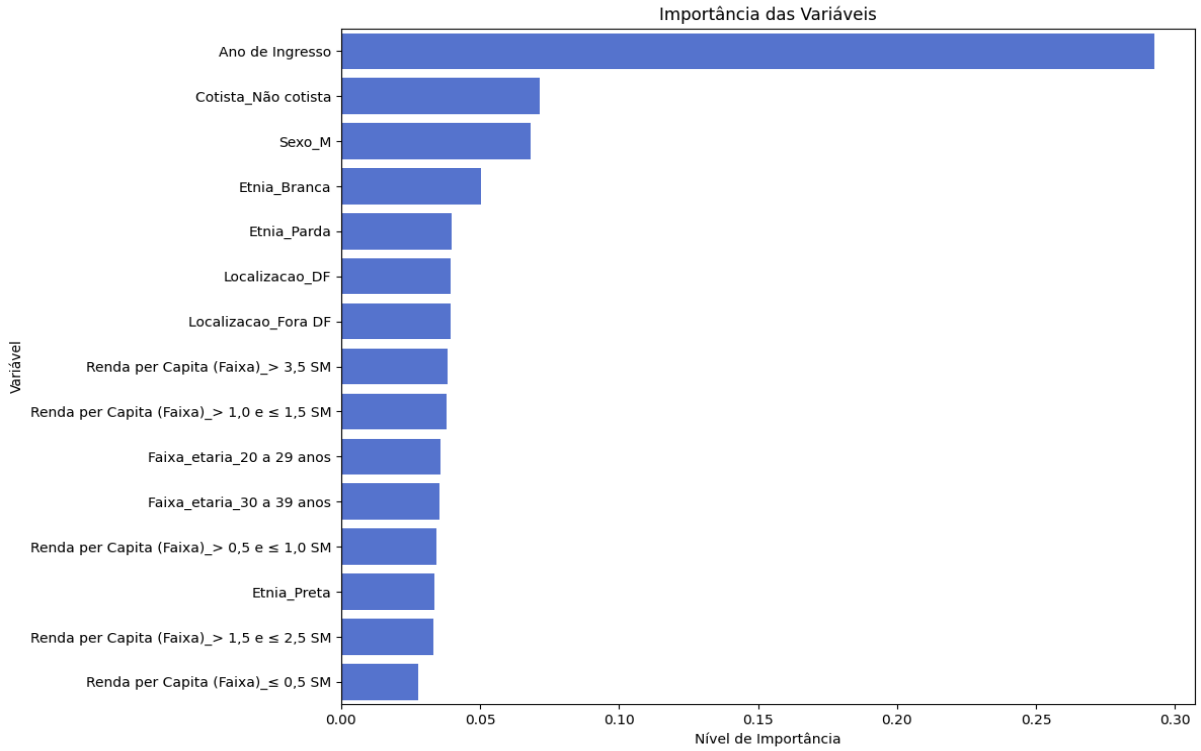


Fonte: Elaborado pelas autoras, (2025).

Já na Figura 12 é destacado as variáveis mais influentes na previsão da evasão acadêmica no curso de TDS. Mostra que o Ano de Ingresso é o fator que exerce maior influência na previsão da evasão, apresentando o maior nível de importância entre todas as variáveis analisadas. Esse achado sugere que aspectos conjunturais relacionados ao período de entrada no curso, como alterações institucionais, mudanças de oferta ou contextos sociais específicos como por exemplo o impacto da pandemia de COVID-19, podem estar diretamente associados ao comportamento de permanência ou abandono dos estudantes. Em seguida, destaca-se a variável Cotista/Não cotista, indicando que políticas de ação afirmativa podem impactar a probabilidade de evasão. A variável Sexo_M também assume posição relevante, sugerindo possíveis diferenças de permanência entre perfis de gênero.

As variáveis relacionadas à etnia (Branca, Parda e Preta) e à faixa etária, especialmente estudantes entre 20 e 29 anos, apresentam níveis intermediários de importância, reforçando o papel de características sociodemográficas no risco de abandono. Por fim, fatores socioeconômicos, como as diferentes faixas de Renda per capita e a Localização geográfica (DF ou fora do DF), aparecem com menor peso relativo, embora ainda contribuem para o desempenho do modelo, indicando que condições financeiras e distância da instituição também influenciam a decisão de permanecer ou evadir.

Figura 12 - Importância das variáveis



Fonte: Elaborado pelas autoras, (2025).

Os resultados obtidos reforçam a importância de estratégias direcionadas para apoiar perfis variados de alunos, sendo essencial considerar fatores como ano de ingresso, status de cotista, sexo, faixa etária e etnia. Em termos de modelagem preditiva, o *AdaBoost* demonstrou o melhor desempenho, identificando com precisão os alunos que de fato evadiram e garantindo o menor número de Falsos Negativos. Essa alta eficiência o torna ideal para a detecção precoce de estudantes em risco. O *Gradient Boosting* também se destacou, provando ser uma alternativa robusta ao manter um bom equilíbrio entre acertos e erros, enquanto o *Random Forest* mostrou-se consistente e confiável na identificação de padrões, sendo uma opção equilibrada para cenários que valorizam a estabilidade operacional.

A escolha final do modelo mais adequado deve, portanto, alinhar-se com as prioridades institucionais. Baseado no objetivo do trabalho, o *AdaBoost* é a escolha superior para o curso Técnico subsequente em Desenvolvimento de Sistemas. Se a prioridade for um bom equilíbrio entre a identificação de evadidos e não evadidos, as outras opções se mostram viáveis. O foco de toda essa análise é, em última instância, a utilização inteligente desses insights para a redução efetiva da evasão acadêmica.

Os resultados evidenciam a superioridade dos algoritmos de *Ensemble* (especialmente o AdaBoost), alinhando-se diretamente ao estudo de Silva e Sousa (2024) no contexto do IFB. Na pesquisa citada, o AdaBoost também liderou com sensibilidade de 0,88 e F1-Score de 0,71, corroborando a robustez desse método para os dados institucionais e superando modelos clássicos como o KNN.

Um diferencial metodológico deste trabalho foi a adoção de uma abordagem híbrida de balanceamento, aplicando Oversampling (SMOTE) no ensino médio e Undersampling no técnico subsequente, este último similar à estratégia utilizada por Silva e Sousa (2024). Essa adaptação permitiu aproveitar melhor os dados disponíveis, mesmo quando eram poucos, gerando modelos que funcionam melhor para a realidade de cada curso.

4.4 Avaliação de Desempenho - Técnico em Informática Integrado ao Ensino Médio

A análise dos modelos sem a aplicação do *GridSearch* permitiu compreender o comportamento inicial de cada algoritmo diante do problema de prever a evasão no Ensino Médio. A partir das matrizes de confusão obtidas, foi possível identificar o padrão de erros e acertos de cada modelo, servindo como ponto de partida para avaliar quais técnicas se beneficiam mais da otimização posterior.

Conforme mostrado na tabela 7, o K-Nearest Neighbors foi o primeiro modelo avaliado e apresentou um desempenho moderado, porém com limitações importantes. Ele acertou 91 estudantes que não evadiram (Verdadeiros Negativos – VN) e apenas 5 estudantes que realmente evadiram (Verdadeiros Positivos – VP). Além disso, registrou 34 Falsos Positivos (FP) e 11 Falsos Negativos (FN). Esses resultados mostram que o KNN teve dificuldade média em identificar corretamente os alunos em risco de evasão, caracterizando um comportamento desfavorável para aplicações práticas, já que muitos alunos que deveriam receber intervenção não seriam detectados.

Tabela 7 – Matrizes de confusão sem Grid Search

	NÃO EVADIU	**EVADIU**
NÃO EVADIU	KNN: 91 LR: 92	KNN:34 LR: 33

EVADIU	RF: 124 SVM:74 GB:119 AB:77	RF: 1 SVM:51 GB:6 AB:48
	KNN: 11 LR: 11 RF:15 SVM:5 GB:15 AB:8	KNN:5 LR: 5 RF:1 SVM:11 GB:1 AB:8

Fonte: Elaborado pelas autoras, (2025).

A *Logistic Regression* apresentou desempenho semelhante ao do KNN. O modelo classificou corretamente 92 alunos que não evadiram (VN) e apenas 5 alunos que efetivamente evadiram (VP). Por outro lado, gerou 33 Falsos Positivos (FP) e 11 Falsos Negativos (FN), evidenciando uma baixa capacidade de separar adequadamente as classes. O modelo não demonstrou efetividade suficiente para ser considerado um modelo confiável para monitoramento de evasão sem ajustes adicionais.

O modelo *Random Forest* apresentou forte capacidade de identificar alunos que permaneceram na escola, alcançando 124 Verdadeiros Negativos, o melhor resultado entre os modelos nesse aspecto. No entanto, sua capacidade de identificar estudantes evadidos foi bastante limitada, com apenas 1 Verdadeiros Positivos e 15 Falsos Negativos. Isso demonstra que, apesar de ser eficiente em reconhecer a classe majoritária, o modelo falha em capturar padrões relacionados à evasão, indicando sensibilidade insuficiente à classe minoritária quando utilizado sem otimização.

O *Support Vector Machine* apresentou o melhor desempenho na identificação dos estudantes que evadiram, acertando 74 casos. No entanto, obteve apenas 11 acertos na classe “Não evadiu” e gerou 51 Falsos Positivos, classificando muitos alunos como se fossem abandonados quando, na verdade, permaneceram. Isso mostra que o SVM, sem ajustes, tende a ser excessivamente sensível à evasão, resultando em um número alto de alarmes falsos, apesar de identificar melhor a classe minoritária.

O *Gradient Boosting* apresentou um desempenho menor do que alguns modelos anteriores, acertando 119 estudantes que permaneceram (VN) e 1 que

evadiram (VP). Entretanto, ainda registrou 15 Falsos Negativos e 6 Falsos Positivos, mostrando dificuldade em distinguir corretamente a classe minoritária.

Por fim, o *AdaBoost* demonstrou resultados mais estáveis entre os modelos avaliados. Ele acertou 77 estudantes que não evadiram (VN) e 8 que evadiram (VP), registrando 48 Falsos Positivos e 8 Falsos Negativos. Embora não tenha apresentado desempenho superior como observado em outros contextos, foi um dos modelos que mostrou comportamento mais equilibrado entre acertos e erros, ainda que a identificação da evasão tenha permanecido limitada.

De forma geral, os resultados sem *GridSearch* evidenciam que todos os modelos tiveram dificuldades significativas na detecção da evasão estudantil no Ensino Médio, especialmente devido ao desbalanceamento entre as classes. Enquanto modelos como *Random Forest*, *Gradient Boosting* e *AdaBoost* demonstraram forte capacidade de identificar estudantes que não evadiram, eles tiveram desempenho fraco na identificação dos casos de evasão. Já modelos como SVM apresentaram maior sensibilidade à classe minoritária, mas às custas de um número elevado de falsos positivos. Esses resultados reforçam a necessidade de ajustes mais estruturados, como o uso do *GridSearch* e técnicas de balanceamento, para melhorar a precisão e a utilidade prática dos modelos na previsão da evasão.

Diante disso, a etapa seguinte do estudo aplicou o *GridSearch* com o objetivo de otimizar os hiperparâmetros de cada algoritmo.

Após a aplicação do *GridSearch*, na tabela 8, foi possível observar melhorias no desempenho de vários modelos, indicando que a otimização de hiperparâmetros exerce um papel significativo no ajuste dos algoritmos à tarefa de previsão da evasão. As novas matrizes de confusão mostram como cada modelo respondeu ao processo de otimização, permitindo comparar os comportamentos antes e depois do ajuste.

O modelo *K-Nearest Neighbors* apresentou avanços importantes após a otimização. Apesar de ainda registrar dificuldades na identificação da evasão, o número de Verdadeiros Positivos aumentou de 5 para 8, e os Falsos Negativos diminuíram para 8. No lado oposto, houve um aumento nos Falsos Positivos (48). Isso indica que o KNN passou a detectar mais alunos em risco, mas ao custo de

classificar equivocadamente mais estudantes como potenciais evadidos, o que pode gerar intervenções desnecessárias.

Tabela 8 – Matrizes de confusão com Grid Search

	NÃO EVADIU	**EVADIU**
NÃO EVADIU	KNN: 77 LR: 92 RF:122 SVM:62 GB:95 AB:33	KNN:48 LR:33 RF:3 SVM:63 GB:30 AB:92
EVADIU	KNN: 8 LR:11 RF:14 SVM:2 GB:13 AB:0	KNN:8 LR:5 RF:2 SVM:14 GB:3 AB:16

Fonte: Elaborado pelas autoras, (2025).

A *Logistic Regression* manteve os resultados, sem variações após a otimização. O modelo registrou 92 Verdadeiros Negativos e 5 Verdadeiros Positivos, valores iguais aos obtidos antes do *GridSearch*. As pequenas mudanças sugerem que o modelo já estava relativamente ajustado e não se beneficiou tanto da otimização quanto outros algoritmos, mantendo um desempenho moderado e ainda limitado na detecção da evasão.

O *Random Forest* foi um dos modelos que mais apresentou estabilidade antes e depois da otimização. O número de acertos na classe “Evadiu” permaneceu baixo, com apenas 2 Verdadeiros Positivos, enquanto o número de Verdadeiros Negativos diminuiu para 122. Esse comportamento reforça que, mesmo com otimização, o Random Forest continua priorizando a classe majoritária, demonstrando dificuldade em aprender padrões relacionados à evasão no Ensino Médio.

O *Support Vector Machine* apresentou melhorias significativas. Os Verdadeiros Positivos aumentaram de 11 para 14, e os Verdadeiros Negativos passaram de 74 para 62. No entanto, o destaque está na redução do número de Falsos Negativos, que caiu para 2, indicando uma melhora na capacidade do modelo de identificar corretamente os alunos que realmente iriam evadir. Apesar disso, o número de Falsos Positivos ainda é elevado (63), o que mostra que o algoritmo continua

sensível e inclinado a classificar alunos como evadidos, mas com maior eficácia do que antes do ajuste.

O *Gradient Boosting* também demonstrou avanços após o *GridSearch*. O modelo passou a identificar 3 alunos que realmente evadiram, contra apenas 1 antes da otimização. Esse aumento expressivo nos Verdadeiros Positivos, acompanhado de uma redução dos Falsos Negativos, evidencia que o *Gradient Boosting* se beneficiou significativamente do ajuste dos hiperparâmetros. Apesar de ainda apresentar erros importantes, o modelo mostrou maior sensibilidade à classe minoritária e melhor equilíbrio geral.

Por fim, o *AdaBoost* apresentou uma das mudanças mais notáveis após o *GridSearch*. O modelo aumentou o número de Verdadeiros Positivos de 8 para 16, além de reduzir os Falsos Negativos para 0. Essa melhora expressiva indica que o algoritmo passou a capturar melhor os padrões associados à evasão, aproximando-se de um desempenho mais adequado ao contexto da previsão educacional. Ainda assim, o número de Falsos Positivos (92) permanece elevado, exigindo cautela na interpretação dos resultados.

De modo geral, a aplicação do *GridSearch* demonstrou impacto positivo na maioria dos modelos, especialmente no KNN, *Logistic Regression* e *AdaBoost,* que apresentaram maior sensibilidade à evasão e melhor capacidade de identificar estudantes em risco. Embora alguns algoritmos ainda tenham mantido dificuldades significativas, como *Gradient Boosting* e, em menor grau, *Random Forest*, os resultados evidenciam que a otimização dos hiperparâmetros é essencial para aumentar o desempenho dos modelos e tornar suas previsões mais confiáveis no contexto educacional.

A comparação entre as métricas antes e depois do *GridSearch* permitiu avaliar o impacto da otimização de hiperparâmetros no desempenho dos modelos. De modo geral, observou-se que alguns algoritmos apresentaram variações de desempenho, mas um ponto de atenção se deve aos baixos valores de precisão (variando entre 0,13 e 0,20). Esse comportamento pode ser um indicativo de que a técnica de Oversampling (SMOTE) não conseguiu gerar amostras sintéticas de qualidade suficiente, introduzindo ruído na base de dados e levando os modelos a cometerem

um número elevado de falsos positivos. A tabela 10 exibe os níveis das métricas de acurácia, precisão, sensibilidade, *F1-score*, e especificidade de cada modelo gerado.

No *K-Nearest Neighbors* , a otimização levou à piora geral das métricas. A acurácia com valores baixos 0.5852, assim como a precisão e a medida-F. Apenas a revocação apresentou melhora, com 0.6464, indicando maior sensibilidade na identificação de alunos que evadiram. Ainda assim, o ganho foi insuficiente para compensar a queda no equilíbrio geral do modelo.

Tabela 9 - Avaliação dos modelos (EMI)

Classificador	Acurácia	Precisão	Revocação	Especificidade	F1-score
kneighbors	0.5852	0.1530	0.6464	0.5788	0.2458
Logistic Regression	0.6553	0.1535	0.4750	0.6781	0.2318
Random Forest	0.8567	0.2550	0.2000	0.9384	0.2194
Support Vector Machine	0.4907	0.1484	0.7536	0.4589	**0.2469**
Gradient Boosting	0.4907	0.1484	0.7536	0.8116	0.2469
Adaboost	0.4446	0.1399	**0.7893**	0.4041	0.2346

Fonte: Elaborado pelas autoras, (2025).

Para a *Logistic Regression* , o *GridSearch* não trouxe qualquer mudança. Todas as métricas permaneceram idênticas, evidenciando que os hiperparâmetros testados não alteraram o comportamento do modelo. Isso sugere que tanto os parâmetros padrão quanto a própria natureza linear do algoritmo limitaram seu desempenho nesse cenário.

O *Random Forest* apresentou uma discreta redução em todas as métricas após a otimização. A acurácia contou com 0.8567, e a medida-F com 0,2194. Mesmo mantendo alta especificidade, a revocação permaneceu baixa, reforçando

que o modelo, mesmo otimizado, continua pouco eficaz na identificação de estudantes que realmente evadem.

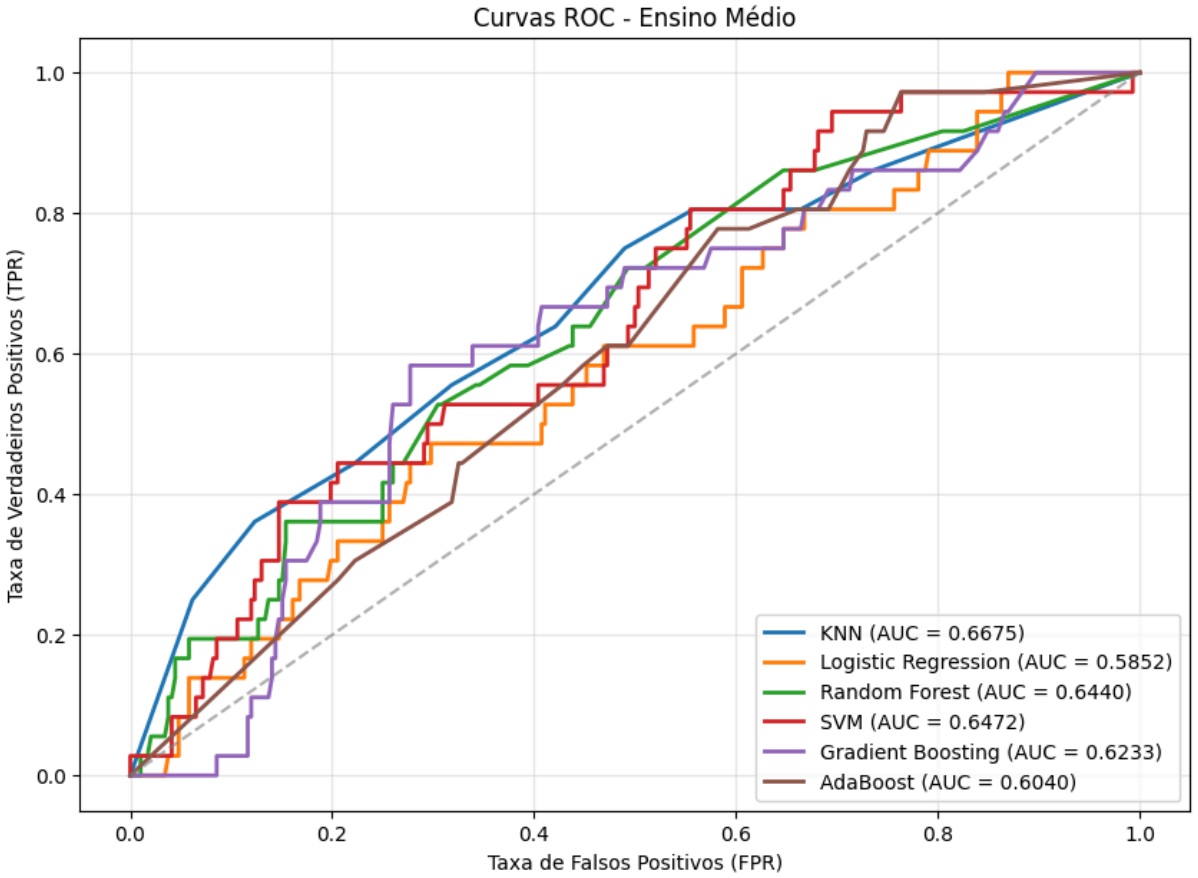
O *Support Vector Machine e o Gradiente Boosting* apresentaram o mesmo desempenho na revocação Também com valores parecidos na acurácia e na precisão, apresentaram um bom desempenho 0.7536. Isso indica que os modelos se tornaram mais sensíveis aos casos de evasão, ainda que à custa de um maior número de falsos positivos.

O *AdaBoost* teve um desempenho na acurácia de 0,4446, enquanto a revocação teve um valor expressivo, chegando a 0.7893. No entanto, essa maior sensibilidade ocorreu acompanhada de forte queda na especificidade e precisão, tornando o modelo mais propenso a erros e menos confiável na prática.

Já na Figura 13 é apresentado os modelos em relação à curva AUC-ROC. A análise das curvas ROC dos modelos otimizados por Grid Search permitiu avaliar a capacidade discriminativa de cada algoritmo na distinção entre alunos evadidos e não evadidos.

A análise das curvas ROC para o ensino médio revela uma melhora no desempenho geral dos modelos após a otimização. O KNN obteve o maior valor de área sob a curva (AUC = 0.6675), indicando a melhor capacidade global de discriminação entre as classes. Em seguida, destacam-se o SVM e o Random Forest, com 0,6472 e 0,6440, respectivamente, demonstrando um desempenho competitivo e superior aos demais. O *AdaBoost* apresentou AUC de 0.6040, seguido de perto pelo *Gradient Boosting* (AUC = 0.6233). Por fim, a *Logistic Regression* obteve o menor valor (AUC = 0.5852). Apesar do KNN apresentar a maior área sob a curva, a seleção final do modelo (AB) priorizou a métrica de revocação, dada a necessidade estratégica de identificar a classe minoritária de evasão com maior sensibilidade.

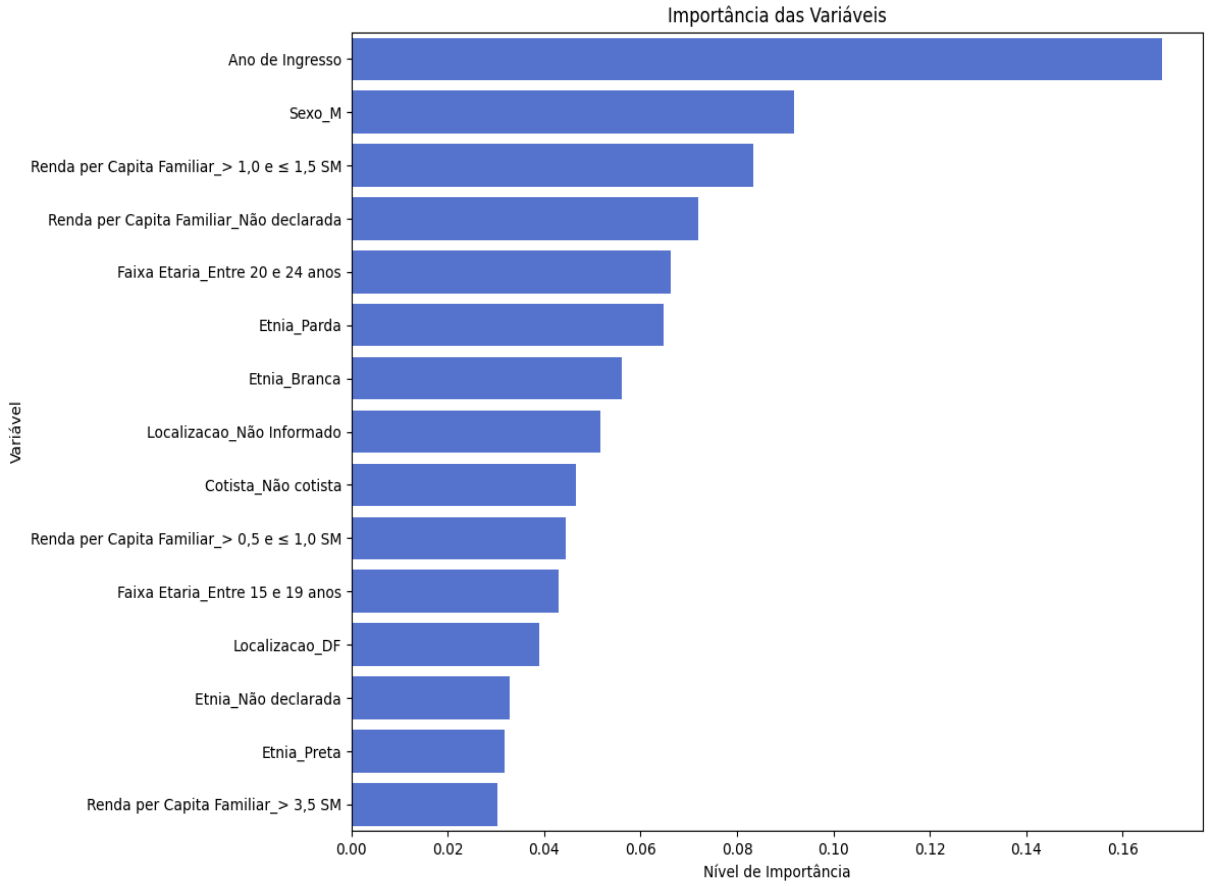
Figura 13 - Curva AUC-ROC



Fonte: Elaborado pelas autoras, (2025).

Já na Figura 14 mostra alguns fatores mais relevantes para a previsão da evasão. Observa-se que o Ano de Ingresso foi o atributo com maior relevância para a previsão da evasão, indicando forte influência do período em que o aluno iniciou o curso no risco de desligamento. Em seguida, destaca-se o Sexo e as variáveis relacionadas à Renda per capita familiar, sugerindo que fatores socioeconômicos desempenham papel importante no fenômeno da evasão. Aspectos demográficos como faixa etária e etnia também contribuíram para a classificação, ainda que em menor intensidade. Por outro lado, atributos como localização e condição de cotista apresentaram impacto menor, indicando que possuem influência mais limitada na tomada de decisão do modelo. Esses resultados reforçam a relevância de características individuais e socioeconômicas na compreensão da evasão escolar.

Figura 14 - Nível de importância das variáveis



Fonte: Elaborado pelas autoras, (2025).

No contexto do Ensino Médio, os resultados obtidos com o algoritmo *AdaBoost* (AB) demonstraram o potencial da aplicação preditiva, alcançando uma sensibilidade (Recall) de 0,78 e F1-Score de 0,23. Esses achados dialogam com o trabalho de Saraiva et al. (2019), que propuseram um modelo de predição para o curso Técnico em Informática do Instituto Federal do Ceará (IFCE). Naquele estudo, os autores alcançaram uma taxa de acerto expressiva de 97,97% ao analisar uma base histórica de oito anos. Tal desempenho superior, pode ser explicado pela inclusão de variáveis de desempenho acadêmico, como notas e frequência, que possuem correlação direta com a permanência do aluno. Em contrapartida, este trabalho focou em variáveis diferentes, cujos padrões de evasão são mais sutis e complexos de detectar.

O ponto chave para esse resultado foi o uso específico do Oversampling (SMOTE). Diante da escassez de registros de evasão (apenas 11%), a geração de

dados sintéticos mostrou-se essencial. Essa estratégia permitiu que o AB aprendesse a identificar os alunos evadidos com maior eficácia, priorizando a detecção da classe minoritária sem o descarte de informações relevantes.

No entanto, pode-se observar que os níveis de acerto do Oversampling foram relativamente baixos, recomendando-se a procura de outras alternativas de algoritmos para o balanceamento de dados como o *SMOTE-NC (Synthetic Minority Oversampling TEchnique - Nominal-Continuous)*, que se trata de uma variação da técnica *SMOTE* utilizada quando os dados possuem variáveis categóricas e numéricas, Islahulhaq (2021).

5 CONCLUSÃO

Este trabalho abordou a análise de perfil de discentes egressos e evadidos dos cursos de EMI e TDS. Foi realizada uma análise e mapeamento do perfil dos alunos de ambos os cursos, além do desenvolvimento dos modelos preditivos reconhecidos pela literatura acadêmica, no qual foram escolhidos com base em suas boas taxas de desempenho para predição de evasão encontrados em trabalhos anteriores, também foram avaliados com as métricas fundamentais para garantir a assertividade dos modelos.

Para realizar o objetivo geral foi utilizada a metodologia CRISP-EDM, onde foi realizado a padronização e normalização utilizando *OneHotEncoder* para as variáveis categóricas e *LabelEnconding* para a variável alvo. Além da utilização de técnicas para o balanceamento dos dados antes do treinamento dos modelos. Os resultados mostraram que o melhor algoritmo de predição no curso Técnico em Informática Integrado ao Ensino Médio foi o Adaboost e para o curso Técnico Subsequente em Análise e Desenvolvimento de Sistemas também permaneceu o Adaboost. Com esses resultados, espera-se que possam ser utilizados para tomadas de decisões de coordenadores e gestores da instituição para avaliar e evitar a evasão escolar ao identificar as variáveis que mais influenciam na evasão tanto no curso de EMI quanto no curso de TDS. Assim, possibilitando a criação de políticas, auxílios e acompanhamento a fim de prevenir que esses alunos que não possuem condições de continuar no curso por situações diversas não abandonem o curso e consigam o certificado de qualificação profissional.

As limitações deste trabalho se deram principalmente com a escassez de estudos sobre evasão relacionadas a cursos técnicos integrados ao ensino médio e técnicos subsequentes, o que limita a análise comparativa com outros trabalhos.

Como sugestão para trabalhos futuros, pode-se realizar uma nova coleta de dados mais recentes. Além disso, a possibilidade de avaliar o desempenho acadêmico e o parecer da motivação que levam esses alunos a evadir, também é sugerido criação de ferramentas de monitoramento que auxiliem gestores e coordenadores nas tomadas de decisão. Espera-se que esse trabalho contribua para estudos de evasão dos cursos técnicos e reforce a importância de pesquisas voltadas para a área de educação profissional e tecnológica.

REFERÊNCIAS

ALBA, Fabio Dall. **Evasão acadêmica em uma instituição de ensino superior privada na região sul do Brasil: do diagnóstico à proposição de um programa de permanência.** Dissertação de Mestrado (Mestrado em Gestão Educacional) – Universidade do Vale do Rio dos Sinos. 2018.

ALMEIDA, Tania Mara Oliveira; FERREIRA, Maria das Graças. Evasão na educação profissional: vencendo os desafios. In: PARANÁ, Secretaria de Estado da Educação. Superintendência de Edu-cação (org.). Os desafios da escola pública paranaense na perspectiva do professor PDE: artigos. Curitiba: **Cadernos PDE**, [S.l.: s.n.], 2014. p. I-84. Disponível em: http://www.diaadiaeducacao.pr.gov.br/portals/cadernospde/pdebusca/producoes_pde/2014/2014_uel_dtec_artigo_tania_mara_de_oliveira_almeida.pdf. Acesso em: 8 jun. 2025.

ANDRADE, Claudio M. V. de; MAGALHÃES, Gestefane Rabbi; ASEVEDO, Raiane; PAES, Julia; OLIVEIRA, Isaias José Ramos; PAGANO, Adriana; REIS, Zilma; GONÇALVES, Marcos A.. Estudo do Impacto de Dados Sintéticos e Paráfrases na Mitigação do Desbalanceamento em Tarefas de Classificação de Textos em Português com Baixa Amostragem. *In*: SIMPÓSIO BRASILEIRO DE BANCO DE DADOS (SBBD), 40. , 2025, Fortaleza/CE. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2025 . p. 837-843. ISSN 2763-8979. DOI: https://doi.org/10.5753/sbbd.2025.247749.

BAKER, R. S. *et al*. Mineração de Dados Educacionais: Oportunidades para o Brasil**. Revista Brasileira de Informática na Educação**., v. 19, 2011.

BARBOSA, Guilherme Matos. *Métodos de aprendizado de máquina aplicados à evasão escolar*. 2022. Dissertação (Mestrado) – Universidade Federal do Paraná. Disponível em: https://acervodigital.ufpr.br/handle/1884/78993. Acesso em: 6 jul. 2025.

BONIFRO, F. D. *et al*. Student dropout prediction. In: SPRINGER. **Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020**, **Proceedings, Part** I 21. [S.l.], 2020. p. 129–140.

BORGES, Rosimeire Aparecida Soares; SILVA, Flavio da; SILVA, Emanuela Francisca Ferreira.A EVASÃO NO CURSO TÉCNICO INTEGRADO AO ENSINO MÉDIO. **Cadernos de Pesquisa,** v. 29, n. 2, p. 13–31, 5 Jul 2022 Disponível em: https://periodicoseletronicos.ufma.br/index.php/cadernosdepesquisa/article/view/15506. Acesso em: 10 jun 2025.

BRASIL. Ministério da Educação. **100 novos IFs**. Disponível em: https://www.gov.br/mec/pt-br/100-novos-ifs. Acesso em: 26 fev. 2026.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha – PNP. Dados estatísticos sobre cursos técnicos do IFB – Campus Brasília. Disponível em: *https://app.powerbi.com/view?r=eyJrIjoiZDhkNGNiYzgtMjQ0My00OGVlLWJjNzYtZWQwYjI2OThhYWM1IiwidCI6IjllNjgyMzU5LWQxMjgtNGVkYi1iYjU4LTgyYjJhMTUzNDBmZiJ9*. Acesso em: 23 nov. 2025.

BREIMAN, L. Random forests. **Machine learning**, Springer, v. 45, p. 5–32, 2001.

CARVALHO, L. M. F. de; REIS, A. V. dos; SIMON, T. T.; CORRÊA, A. D. Comparação de desempenho entre Regressão Logística Binária e Support Vector Machine: ferramentas complementares no diagnóstico médico de sintomas relacionados a distúrbios e qualidade de sono. Caderno Pedagógico, *[S. l.]*, v. 22, n. 6, p. e15465, 2025. DOI: 10.54033/cadpedv22n6-094. Disponível em: https://ojs.studiespublicacoes.com.br/ojs/index.php/cadped/article/view/15465. Acesso em: 6 jul. 2025.

CASTRO, Maria Celi Chaves de; KUENZER, Acácia Zeneida. Ensino médio: uma nova etapa da educação básica? 2. ed. São Paulo: Cortez, 2011.

CHAGAS, P. A.; GOMIDE, J. S.; SANTANA, L. E. A. dos S. Previsão do risco de evasão universitária utilizando aprendizado de máquina: Características e dinâmica da evasão nos cursos de Engenharia da UFRJ-Macaé. **Revista Brasileira de Informática na Educação**, [S. l.], v. 33, p. 1226–1247, 2025. DOI: 10.5753/rbie.2025.5196. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/5196. Acesso em: 24 nov. 2025.

CHUNG, Jae Young, LEE, Sunbok. **Dropout early warning systems for high school students using machine learning**. Children and Youth Services Review, 96:346–353, 2019.

CORRÊA DOS REIS, M.; CORRÊA DOS REIS MAIA, M. Evasão escolar: problema que persevera na educação. **Caminho Aberto: revista de extensão do IFSC**, *[S. l.]*, v. 18, p. 1–19, 2024. DOI: 10.35700/2359-0599.2024.18.3610. Disponível em: https://ojs.ifsc.edu.br/index.php/caminhoaberto/article/view/3610. Acesso em: 18 out. 2025.

DUTRA, J. F.; SOUZA, J. P. L. de; FERNANDES, D. Y. de S. Classificação de estudantes com potencial à evasão: aplicando mineração de dados no contexto de cursos técnicos subsequentes do IFPB. **Revista Principia**, [S. l.], v. 59, n. 3, p. 1009–1027, 2022. DOI: 10.18265/1517-0306a2021id5488. Disponível em: https://periodicos.ifpb.edu.br/index.php/principia/article/view/5488. Acesso em: 26 nov. 2025.

FERREIRA, Anderson; CAVALCANTE, Rodolfo. Análise comparativa entre algoritmos de aprendizagem de máquina em classificação de imagens de radiografia no auxílio ao diagnóstico de pneumonia. *In*: ESCOLA REGIONAL DE COMPUTAÇÃO BAHIA, ALAGOAS E SERGIPE (ERBASE) , 2019, Ilhéus. **Anais** [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019 . p. 245-254.

FERREIRA, L. O. de P.; LARA, H. S.; DANTAS, G. do C.; BRITO, J. N. Detecção de falhas mecânicas através de diferentes técnicas de machine learning. **OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA**, *[S. l.]*, v. 22, n. 9, p. e6780 , 2024. DOI: 10.55905/oelv22n9-137. Disponível em: https://ojs.observatoriolatinoamericano.com/ojs/index.php/olel/article/view/6780. Acesso em: 6 jul. 2025.

FILHO, F. W. B. H.; VINUTO, T. S.; LEAL, B. C. Análise de classificadores para predição de evasão dos campi de uma instituição de ensino federal. In: SBC. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**. [S.l.], 2020. p. 1132–1141.

FILHO, Raimundo Barbosa Silva; ARAÚJO, Ronaldo Marcos de Lima. Evasão e abandono escolar na educação básica no Brasil: fatores, causas e possíveis consequências. **Educação Por Escrito**, *[S. l.]*, v. 8, n. 1, p. 35–48, 2017. DOI: 10.15448/2179-8435.2017.1.24527. Disponível em: https://revistaseletronicas.pucrs.br/porescrito/article/view/24527. Acesso em: 8 jun. 2025.

FREUND, Y., & Schapire, R. E. (1996). "Experiments with a new boosting algorithm." In *Proceedings of the Thirteenth International Conference on Machine Learning* (pp. 148-156). Morgan Kaufmann.

FRIEDMAN, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine." *The Annals of Statistics, 29*(5), 1189-1232.

FUNDAÇÃO ROBERTO MARINHO; ITAÚ EDUCAÇÃO E TRABALHO; FUNDAÇÃO ARYMAX. Juventudes, ensino técnico e o futuro do trabalho: relatório final. São Paulo: Observatório da Educação Profissional e Tecnológica, 2021. Disponível em: https://observatorioept.org.br. Acesso em: 6 jul. 2025.

GUTIÉRREZ, A.L.J. *et al.* **Application of the performance of machine learning techniques as support in the prediction of school dropout.** *Sci Rep* 14, 3957, 2024. Disponível em: https://doi.org/10.1038/s41598-024-53576-1. Acesso em: 6 jul. 2025.

ISLAHULHAQ, Wahyu Wibowo; RATIH, Iis Dewi. Classification of non-performing financing using logistic regression and synthetic minority over-sampling technique-nominal continuous (SMOTE-NC). **International Journal of Advanced Soft Computing and Its Applications**, v. 13, p. 115–128, 2021.

LORENA, A. C.; DE CARVALHO, A. C. P. L. F. Uma Introdução às Support Vector Machines. **Revista de Informática Teórica e Aplicada,** *[S. l.]*, v. 14, n. 2, p. 43–67, 2007. DOI: 10.22456/2175-2745.5690. Disponível em: https://seer.ufrgs.br/index.php/rita/article/view/rita_v14_n2_p43-67. Acesso em: 6 jul. 2025.

LUCHTENBERG, R. *et al.* A. APRENDIZADO DE MÁQUINA EM AMBIENTES HOSPITALARES: UM ESTUDO DE ANÁLISE DE TENDÊNCIAS DE SOBRECARGA EM SISTEMAS DE TECNOLOGIAS DA INFORMAÇÃO E COMUNICAÇÃO. **Revista Contemporânea**, *[S. l.]*, v. 3, n. 9, p. 15866–15893, 2023. DOI: 10.56083/RCV3N9-127. Disponível em: https://ojs.revistacontemporanea.com/ojs/index.php/home/article/view/1766. Acesso em: 13 jul. 2025.

MACEDO, Bruno da Silva; SAPORETTI, Camila Martins. Uso de técnicas de aprendizado de máquina para predição do tempo de graduação dos discentes de Engenharia da Computação na região Sudeste do Brasil. **Revista Brasileira de Computação Aplicada**, [S. l.], v. 16, n. 1, p. 26–37, 2024. DOI: 10.5335/rbca.v16i1.14456. Disponível em: https://ojs.upf.br/index.php/rbca/article/view/14456. Acesso em: 11 nov. 2025.

MACHADO, D. F. M. G.; FERRAZ, D. de O.; SOUSA, A. L. de; JÚNIOR, G. de M. Avaliação de algoritmos de aprendizado de máquina na previsão de evasão escolar: estudo de caso no IFPA campus Altamira. **Cuadernos de Educación y Desarrollo**, [S. l.], v. 16, n. 10, p. e5874 , 2024. DOI: 10.55905/cuadv16n10-067. Disponível em: https://ojs.cuadernoseducacion.com/ojs/index.php/ced/article/view/5874. Acesso em: 18 jun. 2025.

MAGALHÃES, P. V. S.; SOUSA, R. A. da S.; SANTOS, A. C. B.; AGUIAR, F. J. M.; MILFONT, R. T. P. Modelos de Inteligência Artificial para Predição de Evasão escolar: Uma revisão sistemática. Revista Eletrônica de Iniciação Científica em Computação, [S. l.], v. 23, n. 1, p. 48–54, 2025. DOI: 10.5753/reic.2025.5523. Disponível em: https://journals-sol.sbc.org.br/index.php/reic/article/view/5523. Acesso em: 27 fev. 2026.

MORAIS, Flávio L. de; MELO, Ana; MOUTINHO, Mirele; FAGUNDES, Roberta. *Modelos de regressão aplicados na previsão da evasão escolar do ensino básico: uma revisão sistemática da literatura*. In: **SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO (SBIE)**, 32., 2021, Porto Alegre. Anais... Porto Alegre: Sociedade Brasileira de Computação, 2021. p. 168–178. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/18053. Acesso em: 7 jul. 2025.

MOREIRA, Ana Paula de Souza; CARDOSO, Célia Regina O. **Evasão escolar na Educação Profissional e Tecnológica: fatores (in)visíveis para permanência e êxito**. *Revista Cocar*, Belém, v. 18, n. 39, p. 1-20, jan./abr. 2024. Disponível em: https://periodicos.uepa.br/index.php/cocar/article/view/7688/3441. Acesso em: 21 out. 2025.

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**, [S. l.], 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: 26 nov. 2025.

PAIXÃO, G. M. d. M. et al. Machine learning na medicina: Revisão e aplicabilidade. **Arquivos brasileiros de cardiologia**, SciELO Brasil, v. 118, p. 95–102, 2022.

RAMOS, J. *et al*. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. **Anais do XXXI Simpósio Brasileiro de Informática na Educação**, n. 1, p. 1092–1101, 2020. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/12865. Acesso em: 6 jul. 2025.

RUBACK, Lívia; AVILA, Sandra; CANTERO, Lucia. Vieses no Aprendizado de Máquina e suas Implicações Sociais: Um Estudo de Caso no Reconhecimento Facial. *In*: **WORKSHOP SOBRE AS IMPLICAÇÕES DA COMPUTAÇÃO NA SOCIEDADE (WICS)**, 2. , 2021, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2021 . p. 90-101. ISSN 2763-8707. DOI: https://doi.org/10.5753/wics.2021.15967. Acesso em: 6 jul. 2025.

SACRAMENTO, Laura Neta Dias do; ALBUQUERQUE, Monck Charles Nunes de; CYPRIANO, Carlos Alex Cantuário. Estudo sobre Evasão e Permanência no Ensino Técnico de Nível Médio Integrado: Um Mapeamento Sistemática de Literatura. **Revista Labor**, [S. l.], v. 1, n. 26, p. 76–99, 2021. DOI: 10.29148/labor.v1i26.71888. Disponível em: https://periodicos.ufc.br/labor/article/view/71888. Acesso em: 11 nov. 2025.

SARAIVA, Daniel; PEREIRA, Silas; GALLINDO, Érica; BRAGA, Reinaldo; OLIVEIRA, Carina. Uma proposta para predição de risco de evasão de estudantes em um curso técnico em informática. In: WORKSHOP SOBRE EDUCAÇÃO EM COMPUTAÇÃO (WEI), 27., 2019, Belém. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2019. p. 319-333. Disponível em: https://sol.sbc.org.br/index.php/wei/article/view/6639. Acesso em: 23 nov. 2025.

SILVA, F. C. da; CABRAL, T. L. de O.; PACHECO, A. S. V. Dropout or permanence? Predictive models for higher education management. **Education Policy Analysis Archives**, *[S. l.]*, v. 28, p. 149, 2020. DOI: 10.14507/epaa.28.5387. Disponível em: https://epaa.asu.edu/index.php/epaa/article/view/5387. Acesso em: 19 jun. 2025.

SOUSA, Felipe Ferreira de, *et al*. Análise de Classificadores para Predição de Evasão em um Curso Superior de Tecnologia da Informação de um Instituto Federal. **Anais do Computer on the Beach**, v. 16, p. 146–153, 27 maio 2025. Disponível em: https://periodicos.univali.br/index.php/acotb/article/view/21005. Acesso em: 23 nov 2025.

TEODORO, L. de A.; KAPPEL, M. A. A. Aplicação de Técnicas de Aprendizado de Máquina Para Predição de Risco de Evasão Escolar em Instituições Públicas de Ensino Superior no Brasil. **Revista Brasileira de Informática na Educação,** *[S. l.]*, v. 28, p. 838–863, 2020. DOI: 10.5753/rbie.2020.28.0.838. Disponível em: https://journals-sol.sbc.org.br/index.php/rbie/article/view/3691. Acesso em: 13 jul. 2025.

TRIVIÑOS, Antônio Nóbrega da Silva. *Introdução à pesquisa em ciências sociais: a pesquisa qualitativa em educação*. 7. ed. São Paulo: Atlas, 2017.

VASCONCELLOS, Daniella M. *et al*. Análises classificatórias de aprendizado de máquina para identificação de fatores de abandono escolar. *Anais do Simpósio Brasileiro de Informática na Educação (SBIE)*, 2024. Disponível em: https://sol.sbc.org.br/index.php/sbie/article/view/31359. Acesso em: 6 jul. 2025.

VIEIRA, A. C. F.; GALLINDO, E. d. L.; CRUZ, H. A. Plano estratégico para permanência e êxito dos estudantes do ifce. **Fortaleza: IFCE**, 2017