UNIVERSIDADE FEDERAL DO TOCANTINS
CAMPUS UNIVERSITÁRIO DE PALMAS
PROGRAMA DE PÓS-GRADUAÇÃO EM MODELAGEM COMPUTACIONAL DE SISTEMAS

JABSON CAVALCANTE DIAS

TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DA EVASÃO EM CURSOS TÉCNICOS NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA

Palmas, TO
2025

Jabson Cavalcante Dias

TÉCNICAS DE INTELIGÊNCIA ARTIFICIAL PARA PREDIÇÃO DA EVASÃO EM CURSOS TÉCNICOS NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas da Universidade Federal do Tocantins (UFT), para obtenção do título de Doutor em Modelagem Computacional de Sistemas.

Orientador: Prof. Dr. David Nadler Prata

Palmas, TO
2025

Dados Internacionais de Catalogação na Publicação (CIP)
Sistema de Bibliotecas da Universidade Federal do Tocantins

D541t Dias, Jabson Cavalcante.

Técnicas de inteligência artificial para predição da evasão em cursos técnicos na Rede Federal de Educação Profissional, Cientifica e Tecnológica. / Jabson Cavalcante Dias. – Palmas, TO, 2025.

159 f.

Tese (Doutorado) - Universidade Federal do Tocantins – Câmpus Universitário de Palmas - Curso de Pós-Graduação (Doutorado Profissional) em Governança e Transformação Digital - PPGGTD, 2025.

Orientador: David Nadler Prata

1. Inteligência artificial. 2. Evasão escolar. 3. Sistema de predição. 4. Rede Federal de Educação Profissional, Cientifica e Tecnológica. I. Título

CDD 004

Elaborado pelo sistema de geração automática de ficha catalográfica da UFT com os dados fornecidos pelo(a) autor(a).

Jabson Cavalcante Dias

Técnicas de inteligência artificial para predição da evasão em cursos técnicos na Rede Federal de Educação Profissional, Científica e Tecnológica

Tese apresentada ao Programa de Pós-Graduação em Modelagem Computacional de Sistemas. Foi avaliada para obtenção do título de Doutor em Modelagem Computacional de Sistemas e aprovada em sua forma final pelo Orientador e pela Banca Examinadora.

Data de aprovação: 05/12/2025.

Banca Examinadora:

_____

Prof. Dr. David Nadler Prata (Orientador) - UFT

Orientador

_____

Prof. Dr. Alan Kardec Martins Barbiero - UFT

Examinador interno

_____

Prof. Dr. Rogério Nogueira de Sousa - UFT

Examinador interno

_____

Prof. Dr. Antônio da Luz Júnior - IFTO

Examinador externo

_____

Prof. Dr. Fábio Henrique Monteiro Oliveira - IFB

Examinador externo

Dedico o presente trabalho aos meus pais, família e a todos que caminham comigo nessa jornada.

“Chico Xavier costumava ter em cima de sua cama uma placa escrita:

"ISSO TAMBÉM PASSA".

Aí perguntaram para ele o porquê disso.

E ele disse que era para se lembrar que quando estivesse passando por momentos difíceis, poder se lembrar de que eles iriam embora. Que iriam passar. E que ele teria que passar por aquilo por algum motivo.

Mas essa placa também era para lembrá-lo que quando estivesse muito feliz, não deixar tudo para trás e se deixar levar, porque esses momentos também iriam passar e momentos difíceis também viriam de novo. [...]."

Chico Xavier.

AGRADECIMENTOS

Agradeço primeiramente a Deus que me concedeu a oportunidade de realizar um doutorado e me levar onde jamais imaginei estar.

A minha família pela compreensão das minhas ausências, especialmente minha esposa Rosangela e meu filho Fernando por vivenciar comigo angústias, alegrias e realizações.

Agradeço ao Ministério da Educação, por essa grande oportunidade. E aos colegas de trabalho pelo aprendizado e colaboração durante o período do doutorado.

À Universidade Federal de Tocantins pelo suporte nas etapas deste doutorado.

Agradeço ao Prof. Dr. David Nader Prata, que como meu orientador, sempre me apoiou e prontamente esteve ao meu lado para orientar e auxiliar em todos os momentos para o aperfeiçoamento e finalização deste estudo, meu muito obrigado.

Aos demais professores da UFT pela oportunidade de compreender a importância da modelagem computacional com uso de técnicas de inteligência artificial para sua aplicação na gestão pública.

Aos colegas do doutorado, pelos momentos de compartilhamento ao longo dessa trajetória.

A todos que de alguma forma contribuíram para o alcance desse objetivo.

RESUMO

A evasão escolar é amplamente reconhecida como um dos maiores desafios educacionais em todos os países. Apesar da relevância do tema, existem poucos estudos que identificam os fatores de uma perspectiva quantitativa. Isso ocorre principalmente pela falta de dados longitudinais, essenciais para uma avaliação adequada dos fatores que levam à evasão ao longo da trajetória do estudante. Compreender as causas da evasão e suas características pode fornecer valiosos subsídios para a elaboração de estratégias e políticas para reduzir a evasão em cursos técnicos ofertados pela Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil. O objetivo deste projeto é analisar, com a utilização de técnicas de Inteligência Artificial, os fatores que contribuem para a evasão de estudantes em cursos técnicos, a partir da análise de dados extraídos do Sistema Nacional de Informações da Educação Profissional e Tecnológica e divulgados pela Plataforma Nilo Peçanha. A pesquisa é de natureza quantitativa com a utilização das técnicas de pesquisa bibliográfica e documental. A análise e interpretação dos resultados fornecerá subsídios para a elaboração de insumos capazes de atuar nessa importante questão que é tão impactante para a educação e formação técnica de estudantes, a evasão. Para isso, esse trabalho está dividido em três artigos. O primeiro, apresenta as técnicas de aprendizado de máquina utilizadas para predizer a evasão em cursos técnicos na Rede Federal. O segundo artigo, apresenta a análise e interpretação dos resultados da evasão com uso de SHapley Additive Explanations (SHAP). Por último, o terceiro artigo apresenta a plataforma PrevIA, um simulador desenvolvido com inteligência artificial para fornecer a probabilidade de evasão do estudante com base na inserção de novos dados relacionados às suas características pessoais, sociodemográficas e do curso pretendido. Os Resultados da modelagem computacional mostram um desempenho satisfatório do modelo, com um índice de recall de 69% de assertividade e uma área sob a curva ROC (AUC) de 78%, demonstrando a capacidade da abordagem em apoiar os processos decisórios que levam a evasão, e consequentemente, a formular melhores políticas de assistência estudantil para reduzir os efeitos desse fenômeno que causa prejuízos sociais, acadêmicos e financeiros para a educação profissional no país.

Palavras-chave: evasão escolar; inteligência artificial; aprendizado de máquina; Educação Profissional e Tecnológica.

ABSTRACT

School dropout is widely recognized as one of the greatest educational challenges in all countries. Despite the relevance of the topic, there are few studies that identify the factors from a quantitative perspective. This is mainly due to the lack of longitudinal data, which is essential for an adequate assessment of the factors that lead to dropout throughout a student's academic career. Understanding the causes of dropout and its characteristics can provide valuable insights for the development of strategies and policies to reduce dropout in technical courses offered by the Federal Network of Professional, Scientific, and Technological Education in Brazil. The objective of this project is to analyze, using data extracted from the National System of Information on Professional and Technological Education and disclosed by the Nilo Peçanha Platform, the factors that lead students in technical courses to abandon their studies, through the use of Artificial Intelligence techniques. The research is quantitative in nature, using bibliographic and documentary research techniques. The analysis and interpretation of the results will provide input for the development of measures capable of addressing this significant issue, which has such a major impact on students' education and technical training: dropout rates. This work is divided into three articles. The first presents the machine learning techniques used to predict dropout rates in technical courses in the Federal Network. The second article presents the analysis and interpretation of dropout results using Shapley Additive Explanations (SHAP). Finally, the third article presents the PrevIA platform, a simulator developed with artificial intelligence to provide the probability of student dropout based on the insertion of new data related to their personal and sociodemographic characteristics and the intended course. The results of the computational modeling show satisfactory performance of the model, with a recall rate of 69% accuracy and an area under the ROC curve (AUC) of 78%, demonstrating the approach's ability to support decision-making processes that lead to dropout and, consequently, to formulate better student assistance policies to reduce the effects of this phenomenon, which causes social, academic, and financial losses for professional education in the country.

Keywords: school dropout; artificial intelligence; machine learning; Federal Network; vocational education and training.

LISTA DE ILUSTRAÇÕES

Figura 1 - Taxa de eficiência acadêmica em cursos técnicos na Rede Federal EPCT em 2023 20

Artigo 1:
Figura 1 - Taxa de eficiência acadêmica em cursos técnicos na RFEPCT (2020-2023) 31
Figura 2 - Etapas de elaboração do projeto baseado no CRISP-DM 36
Figura 4 - Matriz de correlação de Pearson entre variáveis categóricas codificadas e a situação de evasão 43
Figura 5 - Desempenho do treinamento dos modelos com cross validation 53
Figura 6 - Desempenho do treinamento dos modelos com cross validation (com hiperparâmetros) 54
Figura 7 - Desempenho do treinamento dos modelos com cross validation - com SMOTE ... 54
Figura 8 - Matriz de confusão do modelo CatBoost para predição de evasão escolar 59
Figura 9 - Curva ROC do modelo CatBoost para predição de evasão escolar 61
Figura 10 - Distribuição das probabilidades previstas para alunos evadidos e não evadidos .. 61
Figura 11 - Ordenação dos *scores* de probabilidade - Taxa de evasão por decil 62

Artigo 2:
Figura 1 - Etapas de elaboração do projeto baseado no CRISP-DM 77
Figura 2 - Percentual de evadidos registrado na base de dados após tratamento dos dados 80
Figura 3 - Desempenho do treinamento dos modelos com cross validation (com parâmetros) 85
Figura 4 - Curva ROC do modelo CatBoost para predição de evasão escolar 89
Figura 5 - Mapa de calor dos valores SHAP para o modelo testado 91
Figura 6 - Importância média das variáveis com base nos valores SHAP absolutos 92
Figura 7 - Sumário com distribuição dos valores SHAP por variável 93
Figura 8 - Influência individual dos atributos no resultado do modelo preditivo com SHAP [posição:10] 94
Figura 9 - Influência dos atributos no resultado do modelo preditivo com gráfico de força SHAP [posição:10] 96
Figura 10 - Influência dos atributos no resultado do modelo preditivo com gráfico de barra SHAP [posição:10] 96
Figura 11 - Influência individual dos atributos no resultado do modelo preditivo com SHAP [posição:15445] 97
Figura 12 - Influência dos atributos no resultado do modelo preditivo com gráfico de força SHAP [posição:15445] 98
Figura 13 - Influência dos atributos no resultado do modelo preditivo com gráfico de barra SHAP [posição:15445] 99
Figura 14 - Impactos da idade e carga horária mínima no modelo preditivo com gráfico de dispersão SHAP - separado por Carga Horária Mínima 100

Artigo 3:
Figura 1 - Taxa de eficiência em cursos técnicos na RFEPCT (2020-2023) 114
Figura 2 - Ciclo de vida de projeto com CRISP-DM 118
Figura 3 - Etapas de elaboração do projeto baseado no CRISP-DM 119
Figura 4 - Percentual de evadidos registrado na base de dados utilizada 123
Figura 5 - Desempenho do treinamento dos modelos com cross validation e com ajustes de hiperparâmetros 129

Figura 6 - Tela inicial da plataforma PrevIA..... 132
Figura 7 - Tela com o formulário para simulação da probabilidade de evasão – Plataforma PrevIA..... 133
Figura 8 - Resultado gerado pelo simulador de probabilidade de evasão – Plataforma PrevIA 134
Figura 9 - Tela com indicadores de evasão – Plataforma PrevIA 136
Figura 10 - *Insights* gerados pelo modelo GPT-4o-mini da OpenAI – Plataforma PrevIA... 137
Figura 11 - Mapa dinâmico de calor com a proporção de evadidos por Estados – Plataforma PrevIA..... 137
Figura 12 - Orçamento federal para assistência estudantil na Rede Federal de EPCT período de 2020 a 2023..... 139
Figura 13 - Tela da plataforma PrevIA com o percentual de probabilidade de aluno que não evadiu no conjunto de dados de teste..... 141
Figura 14 - Tela da plataforma PrevIA com o percentual de probabilidade de aluno que evadiu no conjunto de dados de teste..... 142

LISTA DE TABELAS

Artigo 1:
Tabela 1 - Dicionário de dados utilizados no estudo 38
Tabela 2 - Granularidade das variáveis do conjunto de dados 39
Tabela 3 - Desempenho comparativo de modelos preditivos com base em acurácia, precisão, recall, F1-score e ROC-AUC 50
Tabela 4 - Desempenho dos modelos de aprendizado de máquina com ajustes de hiperparâmetros 50
Tabela 5 - Desempenho comparativo de modelos preditivos aplicando SMOTE 51
Tabela 6 - Métricas de avaliação do modelo CatBoost no conjunto de teste 57
Tabela 7 - Métricas de desempenho global do modelo CatBoost (conjunto de teste) 58

Artigo 2:
Tabela 1 - Descrição das variáveis do conjunto de dados 78
Tabela 2 - Desempenho dos modelos de aprendizado de máquina aplicando hiperparâmetros 83
Tabela 3 - Métricas de avaliação do modelo CatBoost no conjunto de teste 87
Tabela 4 - Métricas de desempenho global do modelo CatBoost 89

Artigo 3:
Tabela 1 - Descrição das variáveis da base de dados 120
Tabela 2 - Desempenho dos modelos de aprendizado de máquina aplicando hiperparâmetros 127
Tabela 3 - Métricas de avaliação do modelo CatBoost 130
Tabela 4 - Categorias de risco de evasão escolar na plataforma PrevIA 135

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
AM	Aprendizado de Máquina
AUC	Area Under the Curve
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNCT	Catálogo Nacional de Cursos Técnicos
CRISP	Cross-Industry Standard Process
DM	Data Mining
EaD	Educação a Distância
EDM	Mineração de Dados Educacionais
EJA	Educação de Jovens e Adultos
EPCT	Educação Profissional, Científica e Tecnológica
EPT	Educação Profissional e Tecnológica
FIC	Formação Inicial e Continuada
FN	False Negative
FP	False Positive
GPT	Generative Pre-trained Transformer
IA	Inteligência Artificial
IBGE	Instituto Brasileiro de Geografia e Estatística
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LIME	Local Interpretable Model-agnostic Explanations
MEC	Ministério da Educação
PNAES	Plano Nacional de Assistência Estudantil
PNE	Plano Nacional de Educação
PR	Precisão-Recall
PROEJA	Programa Nacional de Integração da Educação Profissional com a Educação Básica na Modalidade EJA
RFEPCT	Rede Federal de Educação Profissional, Científica e Tecnológica
ROC	Receiver Operating Characteristic
RFP	Renda Familiar Per Capita
RSL	Revisão Sistemática da Literatura
SETEC	Secretaria de Educação Profissional e Tecnológica
SHAP	SHapley Additive exPlanations
SIOP	Sistema Integrado de Orçamento e Planejamento
SISTEC	Sistema Nacional de Informações da Educação Profissional e Tecnológica
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TCU	Tribunal de Contas da União
TN	True Negative
TP	True Positive
UE	Unidade de Ensino
UF	Unidade da Federação
UTFPR	Universidade Tecnológica Federal do Paraná
S/I	Sem Informação
XAI	Inteligência Artificial Explicável

LISTA DE SÍMBOLOS

$E[f(x)]$	Valor esperado da predição do modelo *f* sobre todo o conjunto de dados (baseline).
$f(x)$	Predição do modelo para a instância específica *x*.
x	Instância de dados (vetor de features) sendo analisada.
$\sum$	Soma dos termos subsequentes (operador de somatório).
i	Índice que representa a *i*-ésima variável no modelo.
M	Número total de variáveis (features) no modelo.
ϕi	Valor SHAP (SHapley Additive exPlanations) da *i*-ésima variável, representando sua contribuição marginal para a predição $f(x)$.

SUMÁRIO

1 INTRODUÇÃO 18

1.1 PROBLEMA DE PESQUISA 19
1.1.1 Hipótese 21
1.1.2 Delimitação de Escopo 21
1.1.3 Justificativa 22

2 OBJETIVOS 23

2.1 OBJETIVO GERAL 23
2.2 OBJETIVOS ESPECÍFICOS 23

3 METODOLOGIA 24

3.1 METODOLOGIA DA PESQUISA 24
3.2 PROCEDIMENTOS METODOLÓGICOS 25
3.3 ESTRUTURA DA TESE 25

4 ARTIGO 1 27

1 INTRODUÇÃO 28

2 REVISÃO DA LITERATURA 29

2.1 EVASÃO EM CURSOS TÉCNICOS 30
2.2 NÚMEROS DA EVASÃO EM CURSOS TÉCNICOS NO BRASIL 30
2.3 USO DE INTELIGÊNCIA ARTIFICIAL NA EVASÃO ESCOLAR 32
2.4 MODELOS DE APRENDIZADO DE MÁQUINA 33

3 METODOLOGIA 35

3.1 ENTENDIMENTO DOS DADOS 36
3.1.1 Coleta dos dados 37
3.1.2 Dicionário de Dados 37
3.1.3 Valores presentes no conjunto de dados 38
3.2 ANÁLISE DOS DADOS: REALIZAÇÃO DA LIMPEZA, TRANSFORMAÇÃO, ENGENHARIA DE ATRIBUTOS E ANÁLISE EXPLORATÓRIA DOS DADOS. 39
3.2.1 Construção de novos dados 40
3.2.2 Pré-processamento dos dados 41
3.2.3 Substituição de valores nas colunas 41
3.2.4 Retirada de colunas por questões computacionais 41
3.2.5 Ordenação e substituição de valores com uso de *Ordinal Encoder* 41

3.2.6 Transformação de variáveis categóricas em numéricas com *OneHotEncoder* 42
3.2.7 Exclusão de colunas (n-1) para evitar multicolinearidade 42
3.3 CORRELAÇÕES DAS VARIÁVEIS DO CONJUNTO DE DADOS 43

4 RESULTADOS E DISCUSSÃO 45

4.1 DADOS DA EVASÃO NA REDE FEDERAL EPCT 45
4.2 MODELAGEM COMPUTACIONAL 48
4.3 AVALIAÇÃO DOS MODELOS 49
4.4 TESTE DOS MODELOS DE APRENDIZADO DE MÁQUINA 49
4.5 DESEMPENHO DOS MODELOS 52
4.6 TESTE E AVALIAÇÃO DO MELHOR MODELO 55

5 CONCLUSÃO 63

5 ARTIGO 2 70

1 INTRODUÇÃO 71

2 REVISÃO DA LITERATURA 73

2.1 SHAP (SHAPLEY ADDITIVE EXPLAINATIONS) 75

3 MATERIAIS E MÉTODOS 76

3.1 COLETA DOS DADOS 77
3.1.1 Descrição dos dados 78
3.2 PRÉ-PROCESSAMENTO, LIMPEZA, TRANSFORMAÇÃO, ENGENHARIA DE ATRIBUTOS E ANÁLISE EXPLORATÓRIA DOS DADOS. 79
3.3 ANÁLISE COM SHAP 81

4 RESULTADOS E DISCUSSÃO 82

4.1 MODELAGEM DOS MODELOS DE APRENDIZADO DE MÁQUINA 82
4.2 ANÁLISE DOS MODELOS DE APRENDIZADO DE MÁQUINA 83
4.3 DESEMPENHO DOS MODELOS 85
4.4 DESEMPENHO DO ALGORITMO CATBOOSTCLASSIFIER 87
4.5 INTERPRETANDO OS RESULTADOS DO CATBOOST COM SHAP 90

5 CONCLUSÃO 101

6 ARTIGO 3 109

1 INTRODUÇÃO 110

2 REVISÃO DA LITERATURA 112

2.1 EVASÃO ESCOLAR 112
2.2 REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA 113
2.3 EVASÃO NA REDE FEDERAL DE EPCT 114
2.4 USO DE INTELIGÊNCIA ARTIFICIAL NA EVASÃO ESCOLAR 115
2.4.1 Algoritmos de Gradient Bosting 116
2.5 PLATAFORMA WEB PARA PREDIÇÃO DA EVASÃO ESCOLAR .. 116

3 MATERIAL E MÉTODOS 117

3.1 CONJUNTO DE DADOS E PRÉ-PROCESSAMENTO DE DADOS 119
3.1.1 Dicionário de Dados 120
3.2 ENGENHARIA DE ATRIBUTOS: PRÉ-PROCESSAMENTO, LIMPEZA, TRANSFORMAÇÃO E ANÁLISE EXPLORATÓRIA DOS DADOS. 121
3.3 ETAPA DA MODELAGEM 124
3.4 ETAPA DE AVALIAÇÃO 124
3.5 ETAPA DE IMPLEMENTAÇÃO 125

4 RESULTADOS 125

4.1 MODELAGEM COMPUTACIONAL 125
4.2 AVALIAÇÃO DOS MODELOS 126
4.3 COMPARAÇÃO DOS MODELOS DE APRENDIZADO DE MÁQUINA 127
4.4 DESEMPENHO DOS MODELOS 128
4.5 CATBOOSTCLASSIFIER 129
4.6 IMPLEMENTAÇÃO DA PLATAFORMA PREVIA 130

5 DISCUSSÃO 138

5.1 ESTRATÉGIAS PARA MITIGAR A EVASÃO NA REDE FEDERAL 138
5.2 A CONTRIBUIÇÃO DA PLATAFORMA PREVIA PARA MITIGAR A EVASÃO EM CURSOS TÉCNICOS 141

6 CONCLUSÃO, LIMITAÇÕES E TRABALHOS FUTUROS 144

7 CONSIDERAÇÕES FINAIS DA TESE 155

REFERÊNCIAS 158

apêndice a – Projeto disponibilizado no Github 160

1 INTRODUÇÃO

O abandono dos estudos representa um problema de grande impacto econômico e social no Brasil. Estima-se que o país perca cerca de R$ 220 bilhões por ano, valor correspondente a cerca de 3% do PIB nacional, devido a fatores associados ao abandono escolar, como menor tempo em atividades produtivas, remunerações inferiores e expectativa de vida reduzida (BARROS et al., 2021).

A evasão escolar pode ser compreendida como a situação em que estudantes, após iniciarem seus cursos, abandonam ou se desligam antes da conclusão. Assim, é caracterizada pelo rompimento do percurso formativo do estudante, manifestando-se no desligamento institucional ou na desistência do curso antes da obtenção do diploma. Essa condição, por sua vez, acarreta na estagnação do crescimento profissional e da mobilidade social, além de gerar efeitos psicológicos adversos, como a diminuição da autoestima e do senso de propósito (PEDDITZI, 2021).

No âmbito da educação profissional e tecnológica (EPT), esse fenômeno é reconhecido como um problema de grande relevância em escala global. Diversos estudos internacionais têm demonstrado crescente preocupação com esse fenômeno, revelando que países de diferentes contextos socioeconômicos buscam compreender os fatores associados ao abandono escolar, com o objetivo de formular políticas públicas eficazes e estratégias de prevenção (YI et al., 2015). Nesse contexto, o Brasil não é uma exceção. A realidade brasileira reflete os desafios enfrentados em outras nações no que diz respeito à permanência dos estudantes na EPT.

A literatura especializada aponta certo consenso sobre os fatores que afetam a conclusão em cursos técnicos, como aspectos socioeconômicos, institucionais e individuais. No entanto, ainda existem divergências quanto aos fatores específicos que exercem maior influência na decisão do estudante de abandonar o curso (YI et al., 2015). Predominam as abordagens centradas no comportamento do estudante, com ênfase na identificação de motivações individuais para o abandono. No entanto, há uma lacuna importante relacionada à avaliação de variáveis contextuais que pode afetar a decisão do aluno de permanecer ou não no curso (BÖHN; DEUTSCHER, 2022).

A evasão na EPT representa um obstáculo significativo ao desenvolvimento socioeconômico de um país, especialmente em um contexto em que a demanda por qualificação profissional se intensifica progressivamente. A interrupção precoce dos estudos técnicos ou de programas de aprendizagem profissional não compromete apenas a formação dos indivíduos,

mas também suas chances de inserção ou reinserção no mercado de trabalho, além de dificultar o acesso a novas oportunidades educacionais (BESSEY; BACKES-GELLNER, 2015).

Dados recentes de auditoria realizada pelo Tribunal de Contas da União (TCU), em 2024, revelaram que os cursos técnicos ofertados pela Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) apresentaram uma taxa média de evasão de 41%, evidenciando a gravidade do problema no âmbito da educação profissional pública federal (BRASIL, 2024).

O uso de técnicas de aprendizado de máquina tem se consolidado como uma ferramenta promissora para a identificação de padrões e a previsão de comportamentos a partir da análise de dados. No campo educacional, essas técnicas, especialmente com uso de algoritmos de classificação, estão sendo utilizadas para prever a probabilidade de evasão escolar, permitindo uma atuação preventiva por parte das instituições de ensino. Ao antecipar casos de abandono, é possível implementar estratégias mais eficazes de retenção e apoio ao estudante (MATZ et al., 2023; RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020; BAKER; SIEMENS, 2014).

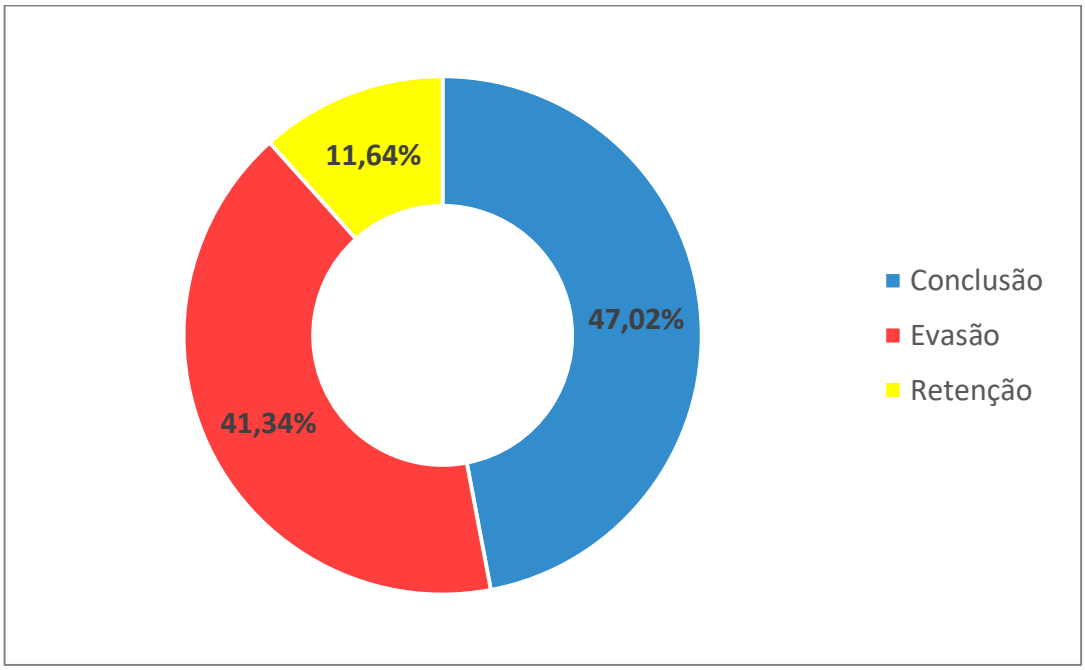
Sob essa perspectiva, é possível incorporar técnicas de inteligência artificial na construção de modelos preditivos voltados à identificação da evasão em cursos técnicos ofertados pela RFEPCT.

Diante desse cenário, este trabalho propõe uma contribuição à área da educação profissional e tecnológica por meio do desenvolvimento de uma modelagem computacional baseada em algoritmos de aprendizado de máquina. O objetivo é estimar a probabilidade de evasão em cursos técnicos, oferecendo uma ferramenta analítica que pode subsidiar ações institucionais preventivas e formulação de políticas públicas voltadas à permanência e ao êxito estudantil.

1.1 Problema de pesquisa

Os elevados índices de evasão em cursos técnicos representam um desafio persistente para os sistemas de ensino. Essa problemática ganha contornos ainda mais preocupantes, considerando os significativos investimentos realizados por meio do orçamento federal para garantir a manutenção e a expansão da RFEPCT. Essa rede tem como missão central ampliar o acesso à educação técnica de qualidade em todo o território nacional, o que torna a alta taxa de evasão um fator crítico que compromete a efetividade das políticas públicas e a alocação eficiente de recursos (BRASIL, 2025).

Figura 1 - Taxa de eficiência acadêmica em cursos técnicos na Rede Federal EPCT em 2023



Fonte: elaborado pelo autor a partir de dados da Plataforma Nilo Peçanha, 2025.

Por outro lado, os modelos computacionais baseados em técnicas de aprendizado de máquina têm se tornado cada vez mais sofisticados e capazes de oferecer alto grau de precisão na análise de dados estruturados.

A aplicação de modelagem computacional, permite o desenvolvimento de modelos preditivos capazes de estimar a probabilidade de evasão em cursos técnicos, utilizando dados públicos disponibilizados pela plataforma oficial da Rede Federal de EPCT, mantida pelo Ministério da Educação (MEC).

Considerando esse cenário, aliado ao domínio ainda limitado de profissionais técnicos e gestores que tenham conhecimento de técnicas estatísticas e de predição para identificar os motivos que levam à evasão em cursos técnicos, surgem algumas questões centrais que podem ser direcionadas para predizer a evasão em cursos técnicos ofertados pela RFEPCT:

Há um padrão recorrente entre os estudantes que evadem dos cursos técnicos na RFEPCT?

Quais variáveis são mais relevantes para indicar os fatores associados à evasão escolar em cursos técnicos?

É possível utilizar técnicas de inteligência artificial para detectar padrões de evasão de forma confiável?

De que maneira a inteligência artificial pode, a partir dos dados disponíveis, contribuir para a previsão da evasão de estudantes na educação profissional técnica?

Diante dessas indagações, este projeto tem como objetivo central responder à seguinte pergunta de pesquisa:

O uso de técnicas de inteligência artificial pode contribuir para identificar a probabilidade de evasão de estudantes em cursos técnicos ofertados pela Rede Federal de EPCT?

1.1.1 Hipótese

Embora o uso de técnicas de inteligência artificial para análise preditiva não seja uma inovação recente, ainda se faz necessária uma análise criteriosa das variáveis disponibilizadas pelo MEC, referentes à Rede Federal de EPCT, a fim de avaliar sua aplicabilidade na identificação dos fatores associados à evasão discente.

O fato é que desde a criação da Rede Federal em 2008 não há evidência de que tenha sido realizado um estudo de abrangência nacional com o objetivo de predizer a evasão em cursos técnicos a partir dos dados públicos disponíveis.

A implementação de modelos computacionais baseados em aprendizado de máquina apresenta-se como uma estratégia promissora para identificar os principais fatores que contribuem para o abandono dos cursos técnicos na Rede Federal de EPCT (MACHADO; FERREIRA; COSTA, 2021).

Considerando o contexto das taxas elevadas de evasão escolar nos cursos técnicos, este estudo parte da seguinte hipótese de pesquisa:

A aplicação de técnicas de aprendizado de máquina, com base nos dados registrados no Sistema Nacional de Informações da Educação Profissional e Tecnológica (Sistec) e disponibilizados pela Plataforma Nilo Peçanha (PNP), possibilita realizar uma análise preditiva eficaz da evasão escolar em cursos técnicos ofertados pela Rede Federal de EPCT.

1.1.2 Delimitação de Escopo

Este estudo utilizará a base de microdados de matrícula, denominada "Eficiência Acadêmica", referente ao ano de 2023, disponibilizada publicamente na Plataforma Nilo Peçanha. Serão considerados apenas os cursos do tipo técnico, totalizando aproximadamente 145 mil registros de matrículas provenientes de todas as instituições que compõem a Rede Federal de EPCT em âmbito nacional.

Os dados serão submetidos a processos de análise, tratamento, limpeza e seleção, de modo a prepará-los para o treinamento de modelos de aprendizado de máquina. Em seguida, serão identificadas as principais métricas de desempenho e a aplicação dos dados de teste, com o objetivo de avaliar, explicar e selecionar o modelo preditivo mais eficaz.

Por fim, os resultados obtidos serão descritos e disponibilizados por meio de uma plataforma online, que permitirá tanto a simulação da probabilidade de evasão, com base na inserção de novos dados individuais de estudantes, quanto o acesso aos principais indicadores da evasão escolar na Rede Federal de EPCT, possibilitando a visualização de chance percentual de evasão em cursos técnicos selecionados.

1.1.3 Justificativa

Considerando a importância do ensino técnico profissionalizante como instrumento de promoção da educação integral e de ampliação da empregabilidade, torna-se imprescindível investigar os fatores que contribuem para a evasão escolar nesse segmento específico. A evasão escolar é amplamente reconhecida como um dos principais desafios enfrentados pelo sistema educacional brasileiro, sendo que os índices nacionais permanecem elevados em comparação com aqueles observados em países desenvolvidos (BRASIL, 2023).

A realização de análises sobre os dados educacionais da Rede Federal, com o propósito de promover o sucesso na formação técnica por meio da aplicação de técnicas de inteligência artificial, representa uma proposta inovadora e ainda pouco explorada no âmbito do serviço público. Tal abordagem pode contribuir significativamente para o aumento da eficiência e para a eficácia dos processos institucionais, promovendo o uso estratégico dos recursos tecnológicos já disponíveis como suporte à tomada de decisões mais assertivas no aprimoramento de políticas públicas voltadas ao enfrentamento da evasão escolar na EPT.

A administração pública, nesse sentido, deve buscar soluções que promovam maior agilidade, transparência e efetividade em seus serviços e sistemas, objetivando o êxito dos estudantes que buscam uma via de desenvolvimento pessoal e inserção qualificada no mercado de trabalho por meio da educação profissional pública.

O uso de técnicas de inteligência artificial tem se mostrado promissor na construção de modelos preditivos que otimizam a experiência do usuário e ampliam a capacidade de ação dos gestores. No caso específico da evasão em cursos técnicos, essas ferramentas podem direcionar intervenções e políticas específicas voltadas à retenção e ao sucesso acadêmico dos discentes.

2 OBJETIVOS

O objetivo geral e os específicos do projeto serão apresentados nas seções seguintes.

2.1 Objetivo Geral

O presente estudo tem como objetivo geral ampliar a compreensão sobre os fatores associados à evasão escolar em cursos técnicos na Rede Federal de EPCT, com o propósito de auxiliar na formulação de políticas públicas voltadas à promoção da eficiência acadêmica. Para isso, propõe-se o desenvolvimento de uma modelagem computacional baseada em técnicas de inteligência artificial, e assim, implementar uma ferramenta web capaz de simular a probabilidade de evasão em cursos técnicos, possibilitando ações preventivas mais eficazes por parte das instituições de ensino.

2.2 Objetivos Específicos

Para o alcance do objetivo geral proposto, os seguintes objetivos específicos foram estabelecidos:

- Revisar a literatura científica e a legislação relacionada à temática da evasão escolar na Rede Federal de EPCT;
- Investigar o uso de técnicas de aprendizado de máquina como ferramenta preditiva de evasão escolar em cursos técnicos;
- Aplicar algoritmos de aprendizado de máquina com o intuito de desenvolver modelo preditivo para evasão escolar (artigo 1);
- Interpretar e explicar os resultados do modelo de aprendizado de máquina escolhido (artigo 2);
- Apresentar uma aplicação web baseada em inteligência artificial que permite simular a probabilidade de um estudante evadir de cursos técnicos (artigo 3).

3 METODOLOGIA

Nesta seção, será apresentada a metodologia utilizada na pesquisa, bem como uma visão geral dos procedimentos utilizados para o desenvolvimento da tese.

3.1 Metodologia da Pesquisa

A pesquisa realizada é de natureza aplicada, tendo como objetivo gerar conhecimentos que possam ser utilizados na resolução de problemas concretos e específicos do contexto educacional, em especial na Rede Federal. A abordagem adotada é quantitativa, caracterizando-se pelo uso de técnicas estatísticas para mensuração e análise dos dados, possibilitando compreender relações entre variáveis, causas e efeitos relacionados à evasão escolar.

De acordo com Creswell (2010), a pesquisa quantitativa permite testar teorias de forma objetiva, examinando a relação entre variáveis mensuráveis por meio de instrumentos padronizados. Os dados obtidos são analisados com o auxílio de procedimentos estatísticos, o que viabiliza a formulação de inferências, explicações alternativas, generalizações e replicações dos resultados.

Além disso, utilizou-se o procedimento de pesquisa bibliográfica, com base em livros, artigos científicos, teses, dissertações e anais de congressos. Essa abordagem visa construir um referencial teórico robusto e confiável sobre a temática estudada (YIN, 2015). Segundo Cervo, Silva e Bervian (2007), a pesquisa bibliográfica busca explicar problemas a partir de referências teóricas já publicadas e pode ser conduzida de forma independente ou complementar a outras modalidades de pesquisa, como a descritiva ou a experimental.

Quanto às ferramentas e recursos computacionais utilizados para o desenvolvimento do projeto, destacam-se: o ambiente de desenvolvimento Visual Studio Code (VSCode), com suporte à linguagem Python 3.11.4 e ao Jupyter Notebook. Foram empregadas bibliotecas reconhecidas na ciência de dados, como *Pandas 2.3.1*, *NumPy 2.2.6*, *Matplotlib 3.10.3*, *Plotly 6.2.0*, *Scikit-learn 1.6.1*, *Streamlit 1.47.0*, *OpenAI 0.28.0*, *GeoPandas 1.1.1*, *Pickle 3.13.5*, *Optuna 4.5.0*, *SHAP 0.48.0*, entre outras.

Para a estruturação das etapas da modelagem computacional, adotou-se a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*). Essa metodologia é utilizada em projetos de mineração de dados e aprendizado de máquina, por oferecer uma estrutura iterativa e sistematizada para o desenvolvimento de soluções baseadas em dados (CHAPMAN et al., 2000). Sua aplicação permitiu uma organização eficiente das etapas, desde a compreensão do negócio até a implementação final do modelo preditivo.

3.2 Procedimentos Metodológicos

O desenvolvimento deste projeto teve início com uma ampla pesquisa bibliográfica, com o intuito de identificar o estado da arte acerca da evasão escolar na EPT, bem como analisar o grau de aprofundamento com que o tema vem sendo explorado nos últimos anos. A partir dessa fundamentação teórica, procedeu-se à análise quantitativa dos dados disponíveis, com foco nos índices de evasão registrados nos cursos técnicos ofertados pela Rede Federal de EPCT.

A pesquisa bibliográfica possibilitou a compreensão das principais abordagens metodológicas e conceituais adotadas pelos pesquisadores no estudo da evasão escolar. Além disso, foram revelados os impactos significativos desse fenômeno em diversas dimensões sociais e econômicas, evidenciando a complexidade do problema. Destaca-se, nesse contexto, a crescente produção científica voltada à aplicação de técnicas de inteligência artificial na predição da evasão escolar, o que demonstra o potencial dessas tecnologias como instrumentos de apoio à gestão educacional.

A obtenção e análise dos dados referentes à evasão escolar nessa rede permitiram o mapeamento de características predominantes do fenômeno, favorecendo a identificação de padrões e variáveis mais associadas à evasão. A abordagem quantitativa adotada forneceu subsídios relevantes para o entendimento dos fatores determinantes da evasão no contexto da EPT.

Por fim, a aplicação de algoritmos de aprendizado de máquina possibilitou a construção e validação de um modelo preditivo voltado à antecipação de possíveis casos de evasão escolar, atendendo ao objetivo central deste estudo. As etapas metodológicas seguidas ao longo da pesquisa, bem como os resultados obtidos, serão detalhadas em cada artigo acadêmico elaborado.

3.3 Estrutura da tese

Além deste capítulo de introdução e um capítulo final com as considerações finais da tese, este trabalho está organizado em três artigos.

No primeiro artigo são explorados os algoritmos de aprendizado de máquina aplicados à predição da evasão em cursos técnicos da Rede Federal de Educação Profissional, Científica

e Tecnológica (RFEPCT). O estudo analisa os principais métodos utilizados e seus fundamentos teóricos, com ênfase nos algoritmos de *boosting*, apresentando uma comparação de desempenho entre diferentes modelos. Além disso, são apontadas as limitações das abordagens existentes e são indicadas perspectivas para pesquisas futuras no campo da predição educacional.

Com base nos resultados do primeiro estudo, o segundo artigo aprofunda a análise da evasão ao investigar os fatores que influenciam esse fenômeno, utilizando técnicas de interpretação de modelos no contexto da Inteligência Artificial Explicável (XAI), com uso do método SHAP (SHapley Additive exPlanations). O artigo discute a relevância de variáveis institucionais, acadêmicas e sociais, evidenciando como cada uma impacta as taxas de permanência e evasão. A partir dessa análise, são propostas reflexões sobre como tais resultados podem orientar políticas e estratégias de mitigação da evasão na Rede Federal de EPCT.

Por fim, o terceiro artigo apresenta a plataforma PrevIA, uma ferramenta que simula a probabilidade de evasão em cursos técnicos na Rede Federal de EPCT construída a partir do modelo desenvolvido e testado nos estudos anteriores. A plataforma integra diferentes dimensões analíticas e fornece índices e relatórios preditivos que podem apoiar a gestão educacional. Além disso, o artigo discute desafios técnicos e institucionais de sua implementação, bem como as potenciais contribuições da PrevIA para o fortalecimento das estratégias de acompanhamento e prevenção da evasão na Rede Federal de EPCT.

4 ARTIGO 1

APLICAÇÃO DE ALGORITMOS DE APRENDIZADO DE MÁQUINA PARA PREDIÇÃO DA EVASÃO EM CURSOS TÉCNICOS NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA DO BRASIL

Revista(s) alvo:

Journal of Artificial Intelligence Research (JAIR) - (https://www.jair.org/index.php/jair/about)

Empirical Research in Vocational Education and Training - (https://link.springer.com/journal/40461)

RESUMO

A evasão escolar é um dos principais desafios educacionais em escala global e afeta diretamente a qualidade da formação profissional e a inserção dos estudantes no mercado de trabalho. No contexto brasileiro, os cursos técnicos ofertados pela Rede Federal de Educação Profissional, Científica e Tecnológica apresentam índices elevados de evasão, o que compromete a efetividade das políticas públicas e o aproveitamento dos investimentos realizados. Este estudo tem como objetivo desenvolver um modelo computacional capaz de prever a evasão em cursos técnicos dessa rede, utilizando técnicas de aprendizado de máquina. A pesquisa utiliza análise bibliográfica e documental com a aplicação de algoritmos de inteligência artificial, especialmente o algoritmo CatBoost, voltado à classificação preditiva. O modelo foi treinado a partir de variáveis com dados pessoais, demográficos e acadêmicos dos estudantes. Os resultados demonstram um desempenho aceitável, alcançando índice de recall de 69% e área sob a curva ROC (AUC) de 78%. Esses indicadores evidenciam a capacidade do modelo em antecipar casos potenciais de evasão e auxiliar nos processos decisórios institucionais. A principal contribuição do estudo consiste na proposição de uma modelagem computacional que possa auxiliar gestores e formuladores de políticas públicas a identificar perfis de risco e a implementar ações preventivas. Ao propor subsídios para reduzir a evasão em cursos técnicos, a pesquisa contribui para fortalecer a eficiência acadêmica dessa rede, minimizar danos sociais e financeiros e ampliar as oportunidades de formação profissional no Brasil.

Palavras-chave: evasão escolar; aprendizado de máquina; CatBoost; cursos técnicos; Rede Federal.

ABSTRACT

School dropout is one of the main educational challenges on a global scale and directly affects the quality of professional training and the integration of students into the labor market. In the Brazilian context, technical courses offered by the Federal Network of Professional, Scientific, and Technological Education have high dropout rates, which compromises the effectiveness of public policies and the use of investments made. This study aims to develop a computational model that can predict dropout rates in technical courses in the Federal Network, using machine learning techniques. The research uses bibliographic and documentary analysis with the application of artificial intelligence algorithms, especially the CatBoost algorithm, focused on predictive classification. The model was trained using variables with students' personal, demographic, and academic data. The results show acceptable performance, achieving a recall rate of 69% and an area under the ROC curve (AUC) of 78%. These indicators demonstrate the model's ability to anticipate potential cases of dropout and assist in institutional decision-making processes. The main contribution of the study is the proposal of a computational model that can help managers and public policy makers identify risk profiles and implement preventive actions. By proposing subsidies to reduce dropout rates in technical courses, the research contributes to strengthening the academic efficiency of the Federal Network, minimizing social and financial damage, and expanding professional training opportunities in Brazil.

Keywords: school dropout; machine learning; CatBoost; technical courses; Federal Network.

1 INTRODUÇÃO

A evasão escolar caracteriza-se pelo rompimento do vínculo do estudante com um curso de formação, ocorrendo de forma antecipada à sua conclusão, independentemente da duração total originalmente definida.

A evasão nos cursos de educação profissional e tecnológica (EPT) constitui-se como um grave entrave ao progresso socioeconômico nacional, sobretudo em um cenário de crescente demanda por mão de obra qualificada. O abandono antes da formação técnica ou de programas de aprendizagem gera prejuízos que afetam diretamente as possibilidades de empregabilidade e a capacidade de retorno ao mercado de trabalho dos cidadãos, bem como limita seu acesso às trajetórias educacionais futuras (BESSEY; BACKES-GELLNER, 2015; HOLTMMANN; SOLGA, 2023).

Na EPT, a evasão é um problema de relevância mundial, conforme atestam diversas pesquisas internacionais. Os estudos apontam para uma preocupação crescente em compreender os motivos do abandono escolar, visando à criação de estratégias e políticas públicas de

prevenção (YI et al., 2015). Nesse contexto, a realidade do Brasil não é exceção, compartilhando os mesmos obstáculos de outras nações para garantir a continuidade dos estudantes na educação profissional e tecnológica.

A análise das causas da evasão escolar pode incluir uma variedade de elementos, que vão desde características individuais até fatores familiares, sociais, aspectos do sistema de ensino e a relevância curricular dos cursos (YI et al., 2015). A literatura ainda é dominada por perspectivas que focam no comportamento e nas motivações individuais do estudante, deixando uma lacuna significativa na investigação de variáveis contextuais que influenciam a decisão de permanência (BÖHN; DEUTSCHER, 2022).

A gravidade do problema da evasão na EPT ofertada pela Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) foi revelada por uma auditoria do Tribunal de Contas da União (TCU), em 2024, que apresentou uma taxa média de evasão de 41% nos cursos técnicos (BRASIL, 2024a).

As técnicas de aprendizado de máquina, notadamente os algoritmos de classificação, consolidam-se como um método promissor para a predição de comportamentos por meio da análise de dados. Aplicadas ao campo educacional, permitem predizer o risco de evasão escolar, o que viabiliza às instituições de ensino uma atuação proativa na implementação de medidas de retenção e apoio ao estudante (RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020).

Sob essa perspectiva, surge a necessidade de incorporar técnicas de Inteligência Artificial (IA) na construção de modelos preditivos voltados à identificação da evasão em cursos técnicos ofertados pela RFEPCT.

Diante desse cenário, este artigo propõe uma contribuição à área da educação profissional e tecnológica por meio do desenvolvimento de uma modelagem computacional baseada em algoritmos de aprendizado de máquina. Serão aplicadas diferentes técnicas de classificação com o objetivo de identificar aquela que apresenta melhor desempenho na predição da evasão em cursos técnicos. A ferramenta resultante poderá subsidiar ações institucionais preventivas e orientar a formulação de políticas públicas voltadas à permanência e ao êxito estudantil.

2 REVISÃO DA LITERATURA

2.1 Evasão em cursos técnicos

O abandono de um curso técnico antes de seu término, conhecido como evasão escolar, é um conceito claro, porém suas causas apresentam inúmeros fatores. A reflexão sobre o tema envolve desde características individuais dos estudantes e suas condições familiares e sociais até elementos estruturais da educação e a relevância da utilidade dos componentes curriculares (BRASIL, 2023a).

Estudos já realizados sobre a evasão escolar confirmam, ainda que parcialmente, os fatores que contribuem para o afastamento dos estudantes da EPT. Aqueles que não concluem uma formação profissionalizante frequentemente atribuem sua desistência à ausência de apoio social, apontando esse fator como central para a decisão de abandonar os estudos (MEEUWISSE; SEVERIENS; BORN, 2010).

De acordo com pesquisa realizada na Hungria, a evasão escolar é influenciada por um conjunto de fatores que abrangem o ambiente escolar e o familiar. Entre esses determinantes, estão as experiências anteriores do aluno na escola, suas condições de moradia, a situação socioeconômica e o contexto do mercado de trabalho local (CSEHNÉ PAPP; HÉDER-RIMA; DAJNOKI, 2021).

A influência do mercado de trabalho na decisão de evadir varia conforme o contexto nacional. Pesquisas mostram que, no Canadá, condições econômicas adversas como baixos salários e alto desemprego, correlacionam-se com menor evasão, enquanto uma idade mais avançada dos estudantes aumenta o abandono. Em contraste, na Suíça, a abundância de empregos para não qualificados eleva a evasão. Já na Itália, não foi detectada uma relação significativa entre o desemprego regional e o fenômeno da evasão escolar (BESSEY; BACKES-GELLNER, 2015).

Böhn e Deutscher (2022), em uma revisão abrangente da literatura, sistematizaram 666 causas potenciais para a evasão na educação profissional e tecnológica a partir de 70 estudos. Suas análises revelaram que a maioria das pesquisas se concentram no âmbito das motivações e comportamentos individuais dos estudantes. No entanto, os autores identificaram uma lacuna crítica na produção acadêmica sobre o tema e apontaram que poucos foram os estudos que buscaram avaliar as condições do ambiente e sua influência na evasão (BÖHN; DEUTSCHER, 2022), sinalizando assim a necessidade de maior investigação sobre fatores contextuais.

2.2 Números da evasão em cursos técnicos no Brasil

Os dados quantitativos referentes às matrículas e às taxas de evasão nos cursos técnicos da RFEPCT são sistematicamente consolidados e divulgados através da plataforma Nilo Peçanha (PNP), conforme estabelecido pelo Ministério da Educação (MEC) (BRASIL, 2023). Esses indicadores são publicados anualmente, após o encerramento do período letivo, constituindo-se como importante ferramenta de monitoramento para gestores educacionais.

Segundo a metodologia oficial estabelecida pela PNP, a taxa de evasão escolar é definida como:

> "O percentual de matrículas que perderam vínculo com a instituição de ensino no ano de referência, sem terem concluído o curso, em relação ao total de matrículas registradas no mesmo período" (BRASIL, 2019, p. 27).

O cálculo é realizado mediante a seguinte fórmula matemática:

Evadidos (%) = (Evadidos / Total de Matrículas) x 100.

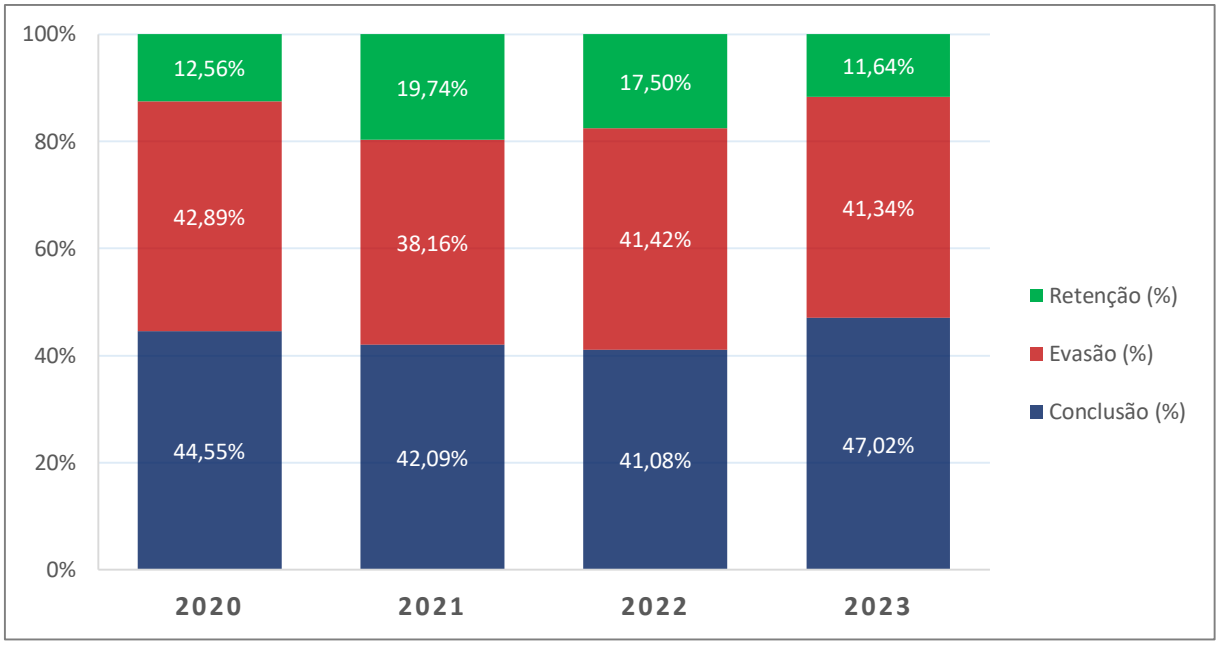
A evasão é calculada por ciclo de matrículas:

> “Um ciclo de matrícula envolve a oferta de um curso com uma carga horária definida, com mesma data de início e mesma previsão de término, visando englobar um conjunto de matrículas de alunos para obtenção de uma mesma certificação ou diploma. A análise do indicador “por ciclo” é realizada considerando a situação de matrícula dos alunos com fim de ciclo previsto para o ano anterior ao de referência. ” (BRASIL, 2019, p. 27).

Assim, o método de análise da PNP realiza o levantamento das situações de matrículas um ano após o término do ciclo. Portanto, as matrículas do ano de 2023, por exemplo, são referentes a ciclos que terminam em 2022.

A figura 1 apresenta a evolução das taxas de evasão em turmas concluídas na RFEPCT no período de 2020 a 2023.

Figura 1 - Taxa de eficiência acadêmica em cursos técnicos na RFEPCT (2020-2023)



Fonte: elaborado pelo autor a partir de dados da Plataforma Nilo Peçanha, 2025.

2.3 Uso de Inteligência Artificial na evasão escolar

As técnicas de aprendizado de máquina mais comuns empregadas, nomeadamente são: árvores de decisão (MURTHY, 1998), redes neurais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MITCHELL, 1997), naïve bayes (DOMINGOS; PAZZANI, 1997), random forest (BREIMAN, 2001), K-Nearest Neighbors (MITCHELL, 1997), regressão logística (LONG, 1997) e máquinas de vetores de suporte (BURGES, 1998).

O aprendizado de máquina é um campo de estudo voltado à análise, desenvolvimento e aplicação de algoritmos capazes de construir modelos preditivos ou decisórios de forma indutiva, a partir de dados. Segundo Mitchell (1997), um programa de computador aprende a partir de uma experiência em uma classe de tarefas quando seu desempenho nessas tarefas, avaliado por uma métrica adequada, melhora ao longo do tempo em decorrência da experiência acumulada.

Os métodos de aprendizado de máquina são classificados em três abordagens principais: aprendizado supervisionado, não supervisionado e aprendizado por reforço. No aprendizado supervisionado, o modelo é treinado a partir de exemplos rotulados, ou seja, pares entrada-saída previamente conhecidos, com o objetivo de aprender uma função de mapeamento que se generalize bem para novos dados. Já o aprendizado não supervisionado busca identificar padrões, estruturas ou agrupamentos nos dados sem a presença de rótulos ou respostas

explícitas, sendo utilizado em tarefas como redução de dimensionalidade e análise de agrupamentos. Adicionalmente, o aprendizado por reforço baseia-se na interação entre um agente e o ambiente, em que o aprendizado ocorre por meio de recompensas ou penalidades atribuídas às ações tomadas, com o objetivo de maximizar uma função de retorno ao longo do tempo (RUSSELL; NORVIG, 2010).

As principais tarefas abordadas no âmbito do aprendizado de máquina, incluem a classificação, a regressão e o agrupamento (*clustering*). A classificação consiste na predição da categoria ou classe à qual uma determinada observação pertence, é utilizada em problemas como diagnóstico médico, detecção de fraude e identificação de evasão escolar. A regressão, por sua vez, tem como objetivo a estimativa de valores numéricos contínuos com base em um conjunto de variáveis preditoras, sendo útil em aplicações como previsão de preços, demanda de mercado ou temperatura. O agrupamento, por outro lado, busca identificar e organizar automaticamente conjuntos de dados em *clusters*, ou grupos, formados por observações similares entre si, sem a necessidade de rótulos pré-definidos, sendo frequentemente empregado em segmentação de clientes e análise exploratória de dados (HASTIE; TIBSHIRANI; FRIEDMAN, 2009).

A aplicação de técnicas de aprendizado de máquina para a previsão da evasão escolar tem sido explorada na literatura científica, demonstrando avanços metodológicos significativos e a obtenção de resultados relevantes (BAKER; SIEMENS, 2014; MATZ et al., 2023; RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020).

2.4 Modelos de aprendizado de máquina

A aplicação de classificadores de aprendizado de máquina tem se mostrado eficaz em tarefas de classificação binária, especialmente em contextos preditivos como a identificação da evasão escolar (KRÜGER; BRITTO; BARDDAL, 2023). Os algoritmos a seguir são reconhecidos por sua robustez, versatilidade e desempenho em diferentes domínios de aplicação.

Support Vector Machine (SVM)

O Support Vector Machine (SVM) é um algoritmo de classificação supervisionada que utiliza os dados de treinamento para definir sua função de decisão, sendo comumente empregado em problemas de aprendizado de máquina binário (RUSSELL; NORVIG, 2010). Sua principal funcionalidade consiste na construção de um hiperplano em um espaço N-

dimensional, sendo N o número de atributos do conjunto de dados, com o objetivo de separar, da forma mais clara possível, as diferentes classes presentes nos dados. Esse hiperplano é definido de modo a maximizar a margem entre os pontos das duas classes, promovendo maior capacidade de generalização do modelo.

Linear Support Vector Classifier (Linear SVC)

O Linear SVC é uma variação do algoritmo *Support Vector Machine* que utiliza uma função de kernel linear. É eficaz em espaços de alta dimensionalidade e busca encontrar o hiperplano ótimo que maximiza a margem entre as classes. Sua eficiência computacional o torna adequado para tarefas de classificação binária com muitos atributos, especialmente quando a separação entre as classes pode ser aproximada por uma função linear (RIFKIN; CLAO, 2003).

K-Nearest Neighbors (KNN)

O KNN é um algoritmo de classificação supervisionada baseado em instâncias, cuja função de decisão é determinada pelos dados de treinamento, sendo, portanto, considerado um método não paramétrico. Sua premissa básica consiste em representar cada instância como um ponto em um espaço N-dimensional, sendo N o número de atributos das observações. A classificação de uma nova instância é determinada com base na classe majoritária entre os k vizinhos mais próximos, definidos por uma métrica de distância, geralmente a distância Euclidiana. De acordo com Mitchell (1997), o KNN parte do pressuposto de que instâncias próximas no espaço de atributos tendem a compartilhar a mesma classe, o que permite inferir padrões sem a necessidade de um modelo de treinamento explícito.

Árvores de Decisão

Classificadores baseados em árvores de decisão são métodos supervisionados que operam por meio da divisão recursiva do conjunto de dados em subconjuntos cada vez mais homogêneos, com base em valores dos atributos. Essa estrutura hierárquica resulta em nós de decisão e folhas, sendo estas associadas às classes finais. O processo de aprendizado consiste na identificação dos atributos que mais contribuem para a separação das classes, formando regras de decisão interpretáveis (MURTHY, 1998).

Random Forest

O Random Forest é um algoritmo de aprendizado supervisionado baseado em conjuntos de árvores de decisão. Ele utiliza o método de *bagging* (*bootstrap aggregating*) para construir múltiplas árvores com subconjuntos aleatórios dos dados e dos atributos. A predição final é feita por votação (classificação) dos resultados de cada árvore. Essa estratégia reduz a variância

do modelo e melhora sua capacidade de generalização, sendo especialmente eficaz em problemas com muitas variáveis e possíveis interações entre elas (BREIMAN, 2001).

Regressão Logística

A Regressão Logística é um modelo estatístico utilizado para tarefas de classificação binária. Baseia-se na função logística (sigmóide) para modelar a probabilidade de ocorrência de uma classe como função linear dos atributos de entrada. É valorizada por sua simplicidade, interpretabilidade e desempenho competitivo em cenários com dados linearmente separáveis (HOSMER; LEMESHOW, 2000).

XGBoost (Extreme Gradient Boosting)

O XGBoost é uma técnica de aprendizado baseada em *boosting*, que constrói modelos sequenciais de árvores de decisão com o objetivo de corrigir os erros das árvores anteriores. Utiliza uma função de perda diferenciável e técnicas de regularização para evitar o sobreajuste, sendo altamente eficiente em termos de tempo e desempenho preditivo. Seu uso tem se destacado em competições de ciência de dados e aplicações práticas que exigem alta acurácia (CHEN; GUESTRIN, 2016).

LightGBM (Light Gradient Boosting Machine)

O LightGBM é uma implementação otimizada do algoritmo de *gradient boosting*, desenvolvida para melhorar a velocidade e o consumo de memória em grandes volumes de dados. Diferencia-se por utilizar a estratégia de crescimento de árvore baseada em folhas (*leafwise*), ao invés do método tradicional *levelwise*, o que permite obter maior acurácia em menor tempo de treinamento. É particularmente eficaz para conjuntos de dados com alta dimensionalidade (KE et al., 2017).

CatBoost (Categorical Boosting)

O CatBoost é um algoritmo de aprendizado de máquina baseado em *gradient boosting*, com foco especial no tratamento eficiente de variáveis categóricas. Diferencia-se de outros algoritmos de *boosting* por incorporar técnicas avançadas como *Ordered Boosting* e *Target Statistics*, que evitam o *target leakage* durante o treinamento. Além disso, o CatBoost realiza a conversão automática de atributos categóricos, reduzindo a necessidade de pré-processamento. O modelo é robusto, apresenta alta acurácia e tem se mostrado eficaz mesmo em conjuntos de dados heterogêneos e desbalanceados (DOROGUSH et al., 2018).

3 METODOLOGIA

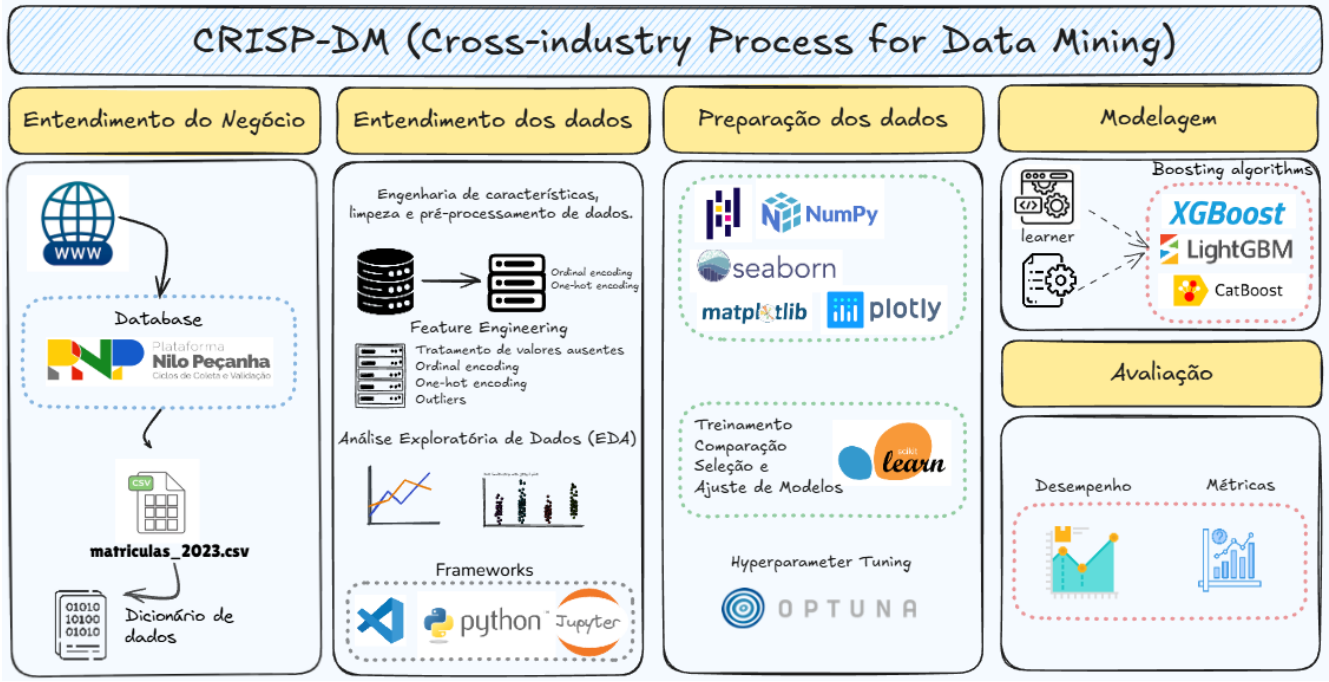
A abordagem metodológica foi estruturada em cinco etapas principais: entendimento do negócio; entendimento dos dados por meio da coleta e pré-processamento; preparação dos dados com foco em engenharia de atributos e análise exploratória; construção dos modelos preditivos; e avaliação de desempenho.

Para orientar o desenvolvimento da parte computacional do estudo, foi utilizada as cinco primeiras etapas da metodologia CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (MARTINEZ, 2019). O CRISP-DM aborda partes de um problema, definindo um modelo de processo que fornece uma estrutura para a execução de projetos independentes da tecnologia utilizada.

Trata-se de um processo que orienta as atividades de mineração de dados de forma sistemática, favorecendo a organização e o controle das etapas envolvidas, desde a compreensão do problema até a avaliação e implementação dos modelos preditivos (IBM, 2023).

A figura 2 apresenta uma visão geral do fluxo de execução dessas etapas, destacando as atividades realizadas e os artefatos utilizados em cada fase do processo.

Figura 2 - Etapas de elaboração do projeto baseado no CRISP-DM



Fonte: elaborado pelo autor, 2025.

3.1 Entendimento dos dados

A fase de entendimento dos dados teve como propósito realizar uma análise exploratória inicial para compreender o fenômeno da evasão. A detecção de padrões e inconsistências neste estágio é fundamental, pois serve como alicerce para as demais fases do projeto, assegurando que a modelagem seja baseada em dados confiáveis e consistentes.

3.1.1 Coleta dos dados

O conjunto de dados utilizado neste estudo foi obtido por meio do Portal de Dados Abertos do Governo Federal (BRASIL, 2023). A origem dos dados é a Plataforma Nilo Peçanha, disponível no portal do Ministério da Educação (BRASIL, 2024a).

A base contempla informações referentes às matrículas de alunos da RFEPCT no ano de 2023, totalizando 145.831 registros. As informações incluem dados demográficos, institucionais e acadêmicos relacionados aos cursos técnicos. Convém destacar que os dados já se encontram rotulados com base na variável referente à categoria da situação de matrícula.

A análise considerou apenas os dados de matrículas em cursos do tipo técnico. Essa delimitação fundamenta-se na legislação que rege a RFEPCT, a qual estabelece que, no mínimo, 50% da oferta educacional das instituições dessa rede deve ser destinada à educação técnica de nível médio (BRASIL, 2008). Além disso, a exigência de uma carga horária mínima de 800 horas para esses cursos permite um acompanhamento anual da trajetória acadêmica dos estudantes.

Além da base principal, também foi incorporada ao projeto uma base de dados referente às regiões metropolitanas do Brasil, disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (IBGE, 2023).

O objetivo dessa integração foi identificar as unidades de ensino localizadas em regiões metropolitanas, a fim de investigar se a localização geográfica exerce alguma influência significativa sobre a probabilidade de evasão escolar. A hipótese é que fatores territoriais, como infraestrutura urbana, mobilidade e acesso a oportunidades, exerçam influência sobre a permanência dos estudantes nos cursos técnicos.

3.1.2 Dicionário de Dados

Com o intuito de garantir a clareza e a reprodutibilidade da análise, apresenta-se a seguir o dicionário de dados. Ele fornece a descrição estruturada de todas as variáveis que compõem o conjunto de dados utilizado na investigação do fenômeno da evasão.

Tabela 1 - Dicionário de dados utilizados no estudo

Variável	Descrição	Tipo
Categoria situação (target)	Situação do aluno em relação à matrícula: *em curso*, *concluído* ou *evadido*.	Categórica nominal
Cor/raça	Cor ou raça autodeclarada do aluno: *branca*, *preta*, *parda*, *amarela* ou *indígena*.	Categórica nominal
Idade	Idade do aluno no momento da matrícula ou da coleta dos dados, em anos.	Numérica discreta
Sexo	Gênero do aluno: *masculino* ou *feminino*.	Categórica nominal
Renda familiar	Faixa de renda familiar do aluno, classificada em diferentes níveis.	Categórica ordinal
Modalidade de ensino	Modalidade em que o aluno estuda: *presencial* ou *educação a distância*.	Categórica nominal
Tipo de oferta	Forma de oferta do curso: *subsequente*, *concomitante* ou *integrado*.	Categórica nominal
Turno	Turno de estudo: *matutino*, *vespertino*, *noturno* ou *integral*.	Categórica nominal
Nome de curso	Nome do curso técnico no qual o aluno está matriculado.	Categórica nominal
Eixo tecnológico	Área de conhecimento ou eixo tecnológico ao qual o curso pertence.	Categórica nominal
Carga horária mínima	Carga horária mínima exigida para conclusão do curso, expressa em horas.	Numérica discreta
UF (Unidade da Federação)	Estado da instituição de ensino (ex.: *SP*, *RJ*, *BA*).	Categórica nominal
Município	Município onde está localizada a unidade de ensino.	Categórica nominal
Região	Região geográfica do Brasil: *Norte*, *Nordeste*, *Centro-Oeste*, *Sudeste* ou *Sul*.	Categórica nominal
Instituição	Nome da instituição da RFEPCT à qual a unidade pertence.	Categórica nominal
Unidade de ensino (UE)	Nome da unidade de ensino onde o curso técnico é ofertado.	Categórica nominal
Região metropolitana da UE (Unidade de Ensino)	Indica se a unidade de ensino está localizada em uma região metropolitana (sim ou não).	Categórica nominal

Fonte: elaborado pelo autor, 2025.

3.1.3 Valores presentes no conjunto de dados

A seguir, apresenta-se os valores únicos de cada variável. Esta análise visa caracterizar a diversidade e a granularidade dos dados, sendo fundamental para identificar inconsistências, orientar a codificação de categorias e preparar a base para a modelagem preditiva.

Tabela 2 - Granularidade das variáveis do conjunto de dados

Variável	Descrição
CATEGORIA_SITUACAO (Alvo)	Classifica a situação da matrícula do aluno como *Em curso*, *Concluído* ou *Evadido*.
COR_RACA	Autodeclaração de cor ou raça do aluno: *Branca*, *Preta*, *Parda*, *Amarela* ou *Indígena ou Não informado*;
IDADE	Faixa etária dos alunos varia entre 4 e 84 anos, indicando a necessidade de análise de *outliers* para identificação de valores extremos.
SEXO	Gênero do aluno: *Masculino*, *Feminino* ou *Não informado*.
RENDA_FAMILIAR_PER_CAPITA	Faixas salariais ordenadas: $0 < RFP \leq 0,5$; $0,5 < RFP \leq 1,0$; $1,0 < RFP \leq 1,5$; $1,5 < RFP \leq 2,5$; $2,5 < RFP \leq 3,5$; $RFP > 3,5$; *Não declarada*.
MODALIDADE_DE_ENSINO	Modalidade educacional: *Educação Presencial* ou *Educação à Distância*.
TIPO_DE_OFERTA	Forma de ingresso no curso técnico: *Subsequente*, *Proeja - Subsequente*, *Concomitante*, *Proeja – Concomitante*, *Integrado ou Proeja - Integrado*.
TURNO	Turno de realização das aulas: *Matutino*, *Vespertino*, *Noturno* ou *Integral*.
NOME_DO_CURSO	Registro de 140 diferentes cursos técnicos ofertados.
EIXO_TECNOLOGICO	Agrupamento dos cursos em 13 eixos tecnológicos, conforme o Catálogo Nacional de Cursos Técnicos (CNCT).
CARGA_HORARIA_MINIMA	Três categorias de carga horária mínima exigida: 800, 1.000 e 1.200 horas, conforme o CNCT.
UF	Representação das 27 Unidades Federativas do Brasil.
MUNICIPIO	Registro de matrículas em 550 municípios distintos.
REGIAO	Corresponde às cinco grandes regiões geográficas do Brasil: *Norte*, *Nordeste*, *Centro-Oeste*, *Sudeste* e *Sul*.
INSTITUICAO	Conjunto de 62 instituições da RFEPCT com matrícula no conjunto de dados de 2023.
UNIDADE_DE_ENSINO	Matrículas em 618 unidades de ensino distribuídas em todo o território nacional.
REGIAO_METROPOLITANA_UE	Indica se a unidade de ensino está situada em região metropolitana, com os valores *Sim* ou *Não*.

Fonte: elaborado pelo autor, 2025.

3.2 Análise dos dados: realização da limpeza, transformação, engenharia de atributos e análise exploratória dos dados.

A exploração das principais características do conjunto de dados teve o intuito obter uma compreensão abrangente da estrutura das informações e identificar possíveis pontos críticos antes da etapa de modelagem. A análise exploratória foi essencial para garantir a qualidade dos dados, orientar a escolha de técnicas de pré-processamento e definir estratégias adequadas para a construção de modelos preditivos. Os seguintes aspectos foram abordados:

- Análise de valores nulos: identificação das variáveis com dados ausentes, bem como a avaliação da proporção desses valores em relação ao total de registros (PYLE, 1999).
- Contagem de valores únicos por variável: análise de valores distintos por coluna, especialmente para variáveis categóricas. Essa análise foi relevante para identificar a cardinalidade das variáveis e avaliar a necessidade de codificações adequadas (ex.*: one hot encoding*, *label encoding*), principalmente para algoritmos sensíveis a esse aspecto.
- Visualização da quantidade de dados por variável: representação gráfica do volume de observações associadas a cada variável, auxiliando na identificação de possíveis inconsistências, desequilíbrios ou erros de preenchimento.
- Análise de *outliers*: identificação de valores como o intervalo interquartil (IQR) e o desvio padrão, identificando assim os extremos em variáveis numéricas. A presença de *outliers* pode impactar negativamente o desempenho de algoritmos de aprendizado de máquina, sendo fundamental tratá-los de forma criteriosa (HAN; KAMBER; PEI, 2022).
- Distribuição dos dados: avaliação do comportamento estatístico das variáveis numéricas (ex.: média, mediana, desvio padrão, assimetria) e categóricas (ex.: frequência absoluta e relativa), buscou-se compreender a dispersão, tendências centrais e possíveis assimetrias nos dados.
- Apresentação da categoria da variável alvo: análise da distribuição da categoria da variável de interesse em termos de distribuição entre as classes (evadido, em curso e concluído), com ênfase no balanceamento entre as categorias.
- Análise das variáveis em relação à variável alvo: exploração da relação entre as variáveis explicativas e a variável de saída, com o objetivo de identificar padrões relevantes para a predição.

3.2.1 Construção de novos dados

Houve integração com base externa, como os dados de regiões metropolitanas fornecidos pelo IBGE, permitindo o enriquecimento da análise sob a perspectiva geográfica. A variável alvo CATEGORIA_SITUACAO foi convertida em um formato binário (evadido = 1;

não evadido = 0), de forma a viabilizar sua utilização em algoritmos de classificação supervisionada (PANG et al., 2021).

3.2.2 Pré-processamento dos dados

As variáveis categóricas foram codificadas por meio das técnicas *OneHotEncoder* ou *Ordinal Encoder*, conforme a natureza ordinal ou nominal da variável. O conjunto de dados foi dividido em treino e teste com o uso da técnica *Stratified KFold*, mantendo a proporção da variável alvo nas amostras, o que é essencial para evitar viés na avaliação dos modelos (BRANCO; TORGOSO; RIBEIRO, 2016).

3.2.3 Substituição de valores nas colunas

A padronização dos valores na coluna "Cor/Raça" foi uma etapa crucial para assegurar a consistência categórica da variável, um requisito fundamental para análises estatísticas confiáveis. Essa intervenção é particularmente relevante em bases educacionais, onde variáveis sensíveis, como a autodeclaração étnico-racial, frequentemente apresentam registros inconsistentes ou incompletos, o que pode comprometer a qualidade da modelagem (BRASIL, 2023).

3.2.4 Retirada de colunas por questões computacionais

A redução da dimensionalidade foi aplicada para aumentar a eficiência computacional, reduzir o tempo de treinamento e mitigar o risco de *overfitting* (DOMINGUES et al., 2022; HAN; KAMBER; PEI, 2022). Para tanto, as colunas "município" e "unidade_de_ensino" foram excluídas devido à sua alta cardinalidade, que imporia um custo computacional excessivo durante a codificação e o armazenamento. Adicionalmente, a variável categórica original "renda_familiar" foi suprimida após a criação de sua versão codificada em valores numéricos ordinais. Essa decisão visou eliminar redundâncias e assegurar maior clareza no conjunto final de atributos preparado para a modelagem.

3.2.5 Ordenação e substituição de valores com uso de *Ordinal Encoder*

O tratamento de variáveis categóricas ordinais é uma etapa fundamental na preparação de dados para modelos de aprendizado de máquina. Para a variável "Renda Familiar", cujas faixas possuem uma hierarquia natural bem definida (da menor para a maior), aplicou-se o *OrdinalEncoder* da biblioteca *Scikit-learn* (PEDREGOSA et al., 2011). Esta técnica é a mais

adequada para esse contexto, pois preserva a relação de ordenamento inerente aos dados, convertendo as categorias em uma sequência de inteiros que reflete sua ordem relativa. Dessa forma, garante-se que os algoritmos de aprendizado possam interpretar corretamente a natureza progressiva da informação, otimizando o processo de modelagem (HAN; KAMBER; PEI, 2022).

3.2.6 Transformação de variáveis categóricas em numéricas com *OneHotEncoder*

Para garantir a compatibilidade com algoritmos de aprendizado de máquina que exigem entradas numéricas, variáveis categóricas nominais sem uma relação de ordem inerente, foram codificadas utilizando a técnica *One-Hot Encoding*. Por meio do *OneHotEncoder* da biblioteca *Scikit-learn* (PEDREGOSA et al., 2011), cada categoria única foi transformada em uma nova coluna binária (valor 0 ou 1), indicando a presença ou ausência daquela categoria específica para cada observação.

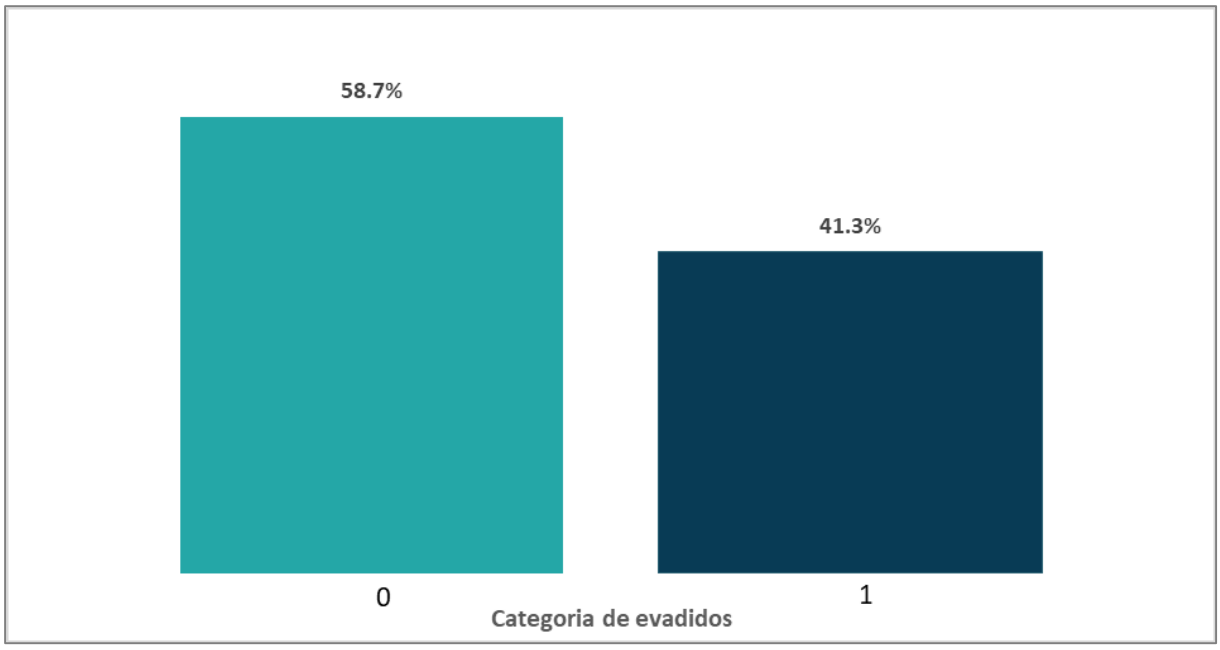
3.2.7 Exclusão de colunas (n-1) para evitar multicolinearidade

A multicolinearidade entre variáveis explicativas pode prejudicar a estabilidade e a interpretação de modelos de aprendizado de máquina, sobretudo aquelas sensíveis a correlações entre atributos. Para mitigar esse problema em variáveis codificadas via *One-Hot Encoding*, adotou-se a estratégia de remoção de uma categoria de referência n-1 (PILLAI; RJU MOHAN, 2024).

Ao realizar a codificação de variáveis categóricas por meio do *OneHotEncoder*, foi gerado um conjunto de colunas binárias correspondente ao número de categorias distintas da variável original. Como as colunas binárias geradas por essa codificação são linearmente dependentes, pois são mutuamente excludentes, a retenção de todas as categorias introduziria multicolinearidade perfeita (JAMES et al., 2021). Dessa forma, para cada variável categórica submetida ao processo, uma coluna foi eliminada, mantendo-se apenas n-1 colunas para representar as categorias originais.

Após a elaboração da engenharia de atributos, o conjunto de dados apresentou leve desbalanceamento conforme figura 3:

Figura 3.2 - Percentual de evadidos registrado na base de dados após pré-processamento



Fonte: elaborado pelo autor com biblioteca Python, 2025.

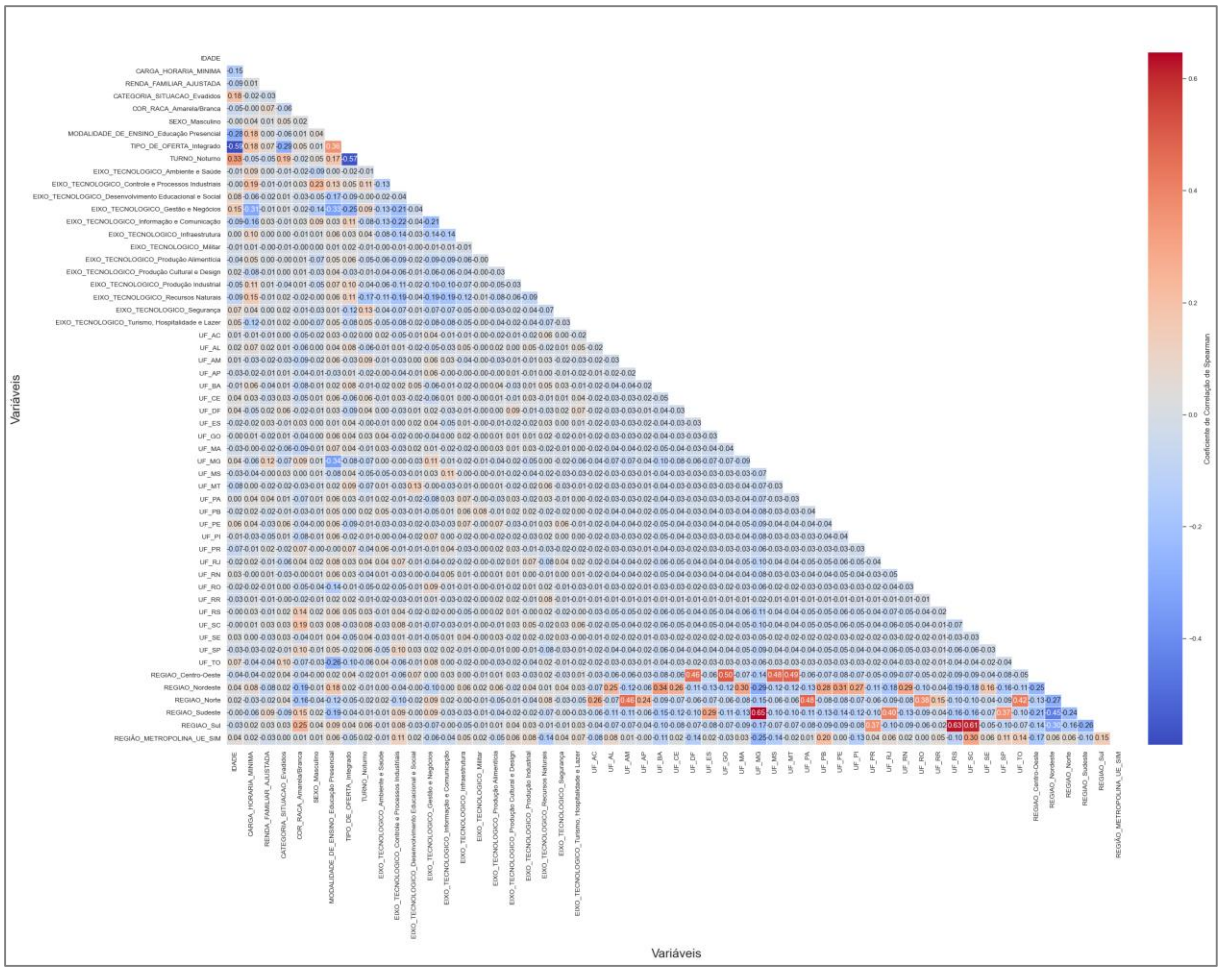
A análise da taxa de evasão apresentada no gráfico revela que 41,3% dos estudantes evadiram dos cursos técnicos na RFEPCT, enquanto 58,7% não evadiram.

3.3 Correlações das variáveis do conjunto de dados

O coeficiente de correlação de Spearman foi o método utilizado, pois permite avaliar a relação entre duas variáveis quando se dispõe de dados ordinais ou de classificação. O objetivo é analisar o nível e a direção da relação monotônica entre as variáveis, de forma que coeficientes altos sugerem que as ordens relativas dos indivíduos em uma variável são parecidas com as ordens na outra variável (SPEARMAN, 2010; DELLINGER, 2017).

Neste estudo, a utilização da correlação de Spearman permitiu a análise combinada de variáveis numéricas e categóricas, que foram previamente convertidas em representações binárias (0 e 1). Isso possibilitou a identificação de padrões de associação significativos entre os atributos analisados. As variáveis nome_de_curso, municipio, instituicao e unidade_de_ensino foram excluídas da elaboração da matriz devido à sua alta cardinalidade, o que poderia resultar em uma apresentação visual pouco agradável.

Figura 4 - Matriz de correlação de Spearman entre variáveis categóricas codificadas e a situação de evasão



Fonte: elaborado pelo autor com biblioteca Python, 2025.

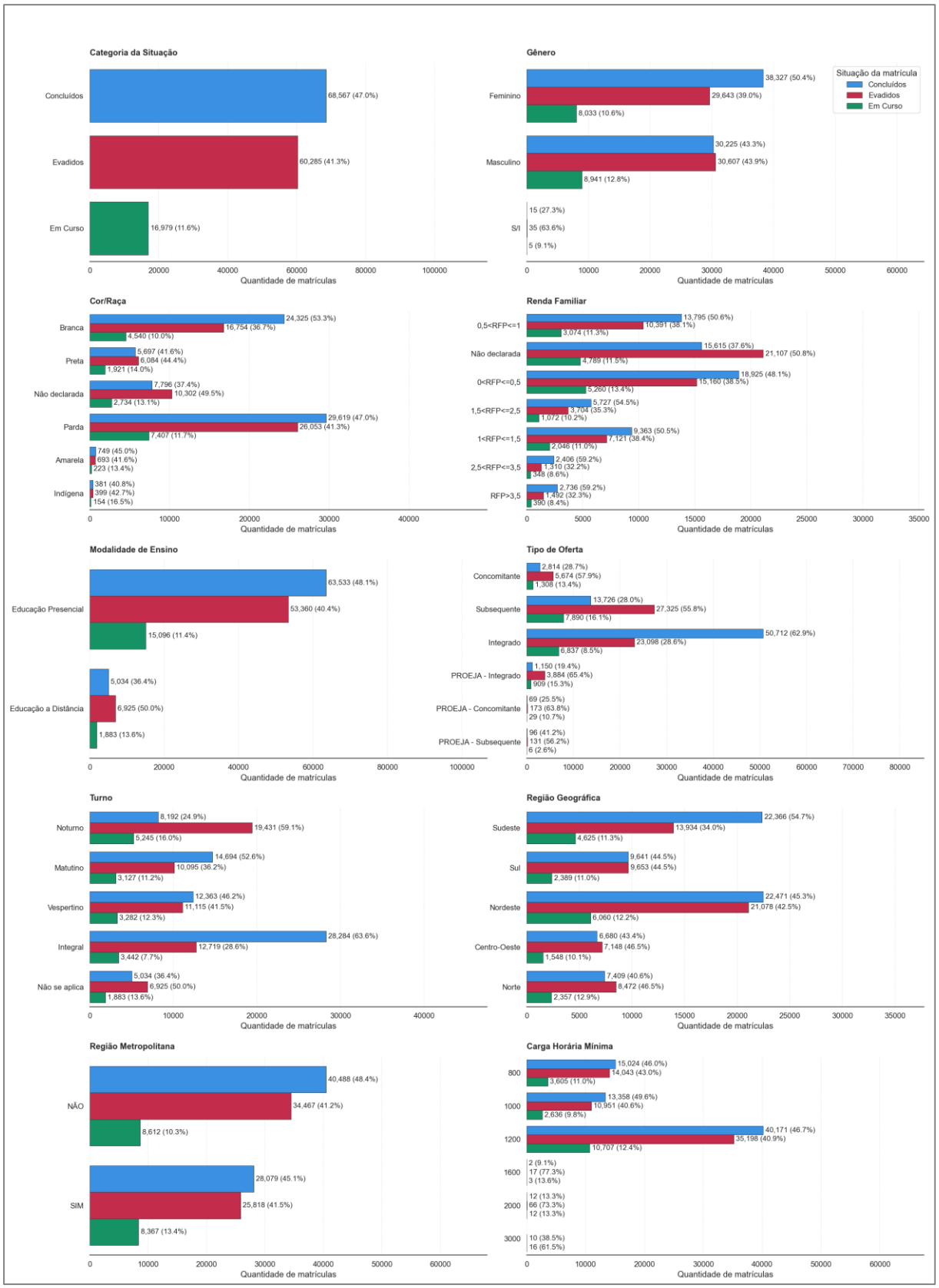
A figura 4 apresenta a matriz de correlação de Spearman, construída a partir dos coeficientes de correlação e assumindo que os valores representados são p-valores associados aos testes de signficância, e não coeficientes de correlação, conforme a formatação de tabela. De modo geral, a matriz indica que as relações entre as variáveis são fracas, com a maior parte dos coeficientes próximos de zero. As correlações mais evidentes concentram-se em pares conceitualmente relacionados, como idade e turno noturno (correlação positiva moderada), sugerindo que alunos mais velhos tendem a se matricular em cursos noturnos, e tipo de oferta integrado e modalidade presencial, que apresentam correlação negativa consistente, refletindo diferenças estruturais entre essas formas de oferta. Além disso, há correlações negativas entre alguns eixos tecnológicos, o que é esperado, pois a codificação binária (*one-hot encoding*) faz com que a presença em um eixo exclua a participação em outro. Essas associações, embora estatisticamente significativas, não indicam causalidade, mas coerência com a organização acadêmica dos cursos.

4 RESULTADOS E DISCUSSÃO

4.1 Dados da evasão na Rede Federal EPCT

A compreesnsão dos dados de evasão na RFEPCT por meio de uma análise descritiva e comparativa foi um passo essencial para entender os padrões de permanência e abandono escolar. A figura 5 exibe gráficos descritivos relacionados a dez variáveis do modelo de dados. As variáveis nome_de_curso, municipio, uf, instituicao e unidade_de_ensino foram excluídas da elaboração devido à sua alta cardinalidade, o que poderia resultar em uma apresentação visual ilegível. Esses gráficos foram elaborados com base na classificação das matrículas em três condições acadêmicas: concluídas, evadidas e em curso. A visão proporciona a comparação da distribuição dos alunos com base em diversos atributos sociodemográficos, institucionais e acadêmicos, permitindo a detecção de padrões, assimetrias e comportamentos diferenciados vinculados a cada tipo de matrícula.

Figura 5 - Distribuição das matrículas por situação acadêmica e características sociodemográficas e institucionais



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A análise dos dados de evasão da Rede Federal de EPCTevidencia que se trata de um fenômeno considerável, representando 41,3% do total de matrículas, percentual próxima ao de

alunos que concluíram os cursos (47,0%). Do ponto de vista sociodemográfico, a maior taxa de evasão é observada entre os alunos do sexo masculino (43,9%), proporção ligeiramente superior à dos que concluem os cursos (43,3%), em consonância com estudos que associam gênero masculino a maior probabilidade de abandono escolar (GREIG, 2019; BÖHN; DEUTSCHER, 2022). Em relação à cor/raça, os grupos mais vulneráveis são os estudantes pretos (44,4%) e indígenas (42,7%), enquanto que alunos brancos apresentam os menores índices de evasão, com 36,7%, o que revela a persistência de desigualdades raciais no que se refere à permanência escolar. No que diz respeito à renda familiar do estudante, desconsiderando os casos sem informação, os alunos que recebem entre 0 e 0,5 salário mínimo (38,5%) e entre 1 e 1,5 salário mínimo (38,4%) são os que mais abandonam em seus grupos, ao passo que aqueles com renda entre 2,5 e 3,5 salários mínimos (32,2%) e acima de 3,5 salários mínimos (32,3%) apresentam menores taxa de evasão, em linha com evidências de que condições socioeconômicas mais desfavoráveis aumentam o risco de evasão (RUMBERGER et al., 1990).

Em relação às características institucionais e acadêmicas, a modalidade de Educação a Distância apresenta a maior taxa de evasão (50,0%). Esse índice é significativamente superior ao observado no grupo da Educação Presencial, que registra uma taxa de 40,4% (WOODLEY; SIMPSON, 2014). Quanto ao tipo de oferta, os cursos concomitantes registram taxa de evasão de 57,9%. Enquanto que no âmbito do PROEJA, os cursos integrados concentram o maior número absoluto de evadidos (65,4%), possivelmente em função do perfil heterogêneo e das condições de vulnerabilidade do público atendido. Em contraste, os cursos integrados apresentam as menores taxas de evasão (28,6%), configurando-se como a oferta com maior capacidade de retenção estudantil (BRIDGELAND; DILIULIO; BALFANZ, 2009).

A análise do turno da oferta revela que a evasão é mais acentuada nos cursos técnicos noturnos, que concentra 59,1% dos alunos que abandonaram os estudos, sendo, assim, o turno com maior risco de evasão escolar. Esse padrão sugere que fatores externos podem ter impacto, como a necessidade de concliliar trabalho e estudo, algo frequente entre os estudantes desse turno (GONÇALVES; PASSOS; PASSOS, 2005). Em sentido oposto, os cursos em tempo integral exibem as menores taxas de evasão (28,6%), reforçando a ideia de que uma permanência mais prolongada na instituição e um vínculo mais sólido com a escola contribuem para a permanência dos alunos.

Em termos regionais, as maiores taxas de evasão são observadas nas regiões Norte e Centro-Oeste (46,5%), enquanto a região Sudeste apresenta a menor proporção de evadidos (34,0%). Isso destaca diferenças regionais consideráveis (BÖHN; DEUTSCHER, 2022). Considerando a localização das instituições, observa-se que aquelas situadas fora das áreas

metropolitanas apresentam uma taxa de evasão menor (41,2%) em comparação com as localizadas em áreas metropolitanas (41,5%). O que indica que o abandono não é um problema restrito ao ambiente urbano-metropolitano. Ademais, isso indica a importância de estratégias institucionais de permanência que considerem as especificidades territoriais e socioeconômicas dos estudantes (ALVAREZ; ALVES; MATOS, 2021).

Por fim, em relação à carga horária mínima, os cursos de 800 horas registram as maiores taxas de evasão, aproximdamente 43,0%, enquanto cargas horárias mais elevadas (1.600 horas ou mais) apresentam percentuais elevados, embora a interpretação desses resultados demande cautela em função do baixo número de matrículas nessas ofertas.

Os resultados reforçam que a evasão na RFEPCT é fortemente condicionada por fatores socioeconômicos, regionais e pelo desenho pedagógico dos cursos, exigindo ações específicas, articuladas a políticas de apoio acadêmico e social e a estratégias de organização curricular mais sensíveis às condições de vida dos estudantes

4.2 Modelagem computacional

A modelagem envolveu a aplicação de algoritmos de aprendizado de máquina com o propósito de prever a evasão escolar com base no conjunto de variáveis explicativas. Inicialmente, os dados foram divididos em dois subconjuntos: 80% para treinamento e 20% para teste, garantindo que os dados de teste não fossem utilizados no processo de treinamento, que é uma boa prática para evitar vazamento de dados (*data leakage*) e permitir uma avaliação mais realista do desempenho dos modelos (GÉRON, 2019).

O parâmetro *random_state* da função *train_test_split()* foi utilizado para controlar a aleatoriedade na divisão dos conjuntos de treinamento e teste, assegurando que a amostragem fosse reproduzível em execuções futuras do código. A aplicação do parâmetro teve como objetivo assegurar a consistência dos experimentos e possibilitar a comparação objetiva entre diferentes configurações de modelos (HANA; LOFSTEAD, 2022).

Durante o pré-processamento, foi aplicada a função *fit_transform()* apenas sobre os dados de treino, enquanto os dados de teste foram submetidos unicamente ao método *transform()*, assegurando que o modelo não fosse influenciado por informações do conjunto de teste durante o aprendizado.

O processo de modelagem teve como foco a identificação do algoritmo com melhor desempenho para avançar às etapas seguintes, como seleção de atributos, ajuste de hiperparâmetros e avaliação final. Para isso, foram considerados tanto a média dos ROC-AUC

scores obtidos nas validações cruzadas quanto o equilíbrio entre viés e variância, visando selecionar um modelo generalizável e com bom desempenho preditivo.

4.3 Avaliação dos modelos

Foram treinados e comparados diversos algoritmos de aprendizado de máquina. Para garantir uma avaliação robusta e reduzir o viés decorrente do desbalanceamento do alvo (evasão), foi empregada a técnica de validação cruzada estratificada *(stratified k-fold cross-validation*), tornando o processo de validação mais confiável em contexto de desbalanceamento (KUHN; JOHNSON, 2013).

A principal métrica adotada para avaliação dos modelos foi o ROC-AUC score, uma vez que a acurácia pode ser enganosa em cenários com classes desbalanceadas. O ROC-AUC permite avaliar a capacidade do modelo em distinguir entre as classes, independentemente de sua distribuição.

Durante o processo de modelagem, também foram aplicadas técnicas de seleção de características e ajustes de hiperparâmetros, visando aprimorar o desempenho dos modelos e reduzir a complexidade computacional. A etapa de ajuste foi realizada com abordagens sistemáticas de busca pelos melhores hiperparâmetros, otimizando os resultados de acordo com a métrica de interesse.

Após a seleção do modelo com melhor desempenho, foi realizada a avaliação final utilizando o conjunto de teste, permitindo observar o comportamento do modelo em um cenário realista de produção.

4.4 Teste dos modelos de aprendizado de máquina

Após o treinamento dos modelos de aprendizado de máquina, procedeu-se à fase de teste com o objetivo de avaliar o desempenho geral de cada algoritmo na predição da evasão escolar. Para essa etapa, foi utilizado um conjunto de teste separado, composto exclusivamente por dados não utilizados durante o treinamento.

Foram testados nove algoritmos: CatBoost, XGBoost, LightGBM, Support Vector Machine (SVM), Random Forest, Linear SVC, K-Nearest Neighbors (KNN), Regressão Logística e Árvore de Decisão. Cada modelo foi previamente treinado utilizando validação cruzada estratificada no conjunto de treinamento.

Os modelos foram avaliados com base em seu desempenho médio nas dobras da validação cruzada, e posteriormente testados no conjunto de teste para verificação de sua capacidade preditiva fora da amostra de treinamento. Os resultados obtidos possibilitaram a comparação objetiva entre os algoritmos, subsidiando a escolha do modelo mais adequado para uso em produção.

Inicialmente, foi elaborada a pontuação de desempenho considerando as métricas de acurácia, precisão, recall, F1-score e ROC-AUC dos nove classificadores.

Tabela 3 - Desempenho comparativo de modelos preditivos com base em acurácia, precisão, recall, F1-score e ROC-AUC

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**ROC-AUC**
CatBoost	0,72	0,69	0,59	0,64	0,78
XGBoost	0,71	0,67	0,58	0,63	0,69
LightGBM	0,71	0,67	0,58	0,62	0,69
Linear SVC	0,68	0,64	0,52	0,57	0,66
SVM	0,68	0,64	0,52	0,57	0,66
Random Forest	0,68	0,62	0,57	0,59	0,66
KNN	0,68	0,62	0,56	0,59	0,66
Logistic Regression	0,68	0,64	0,49	0,56	0,65
Decision Tree	0,66	0,61	0,52	0,56	0,64

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela demonstra que o modelo CatBoost apresentou o melhor desempenho geral, destacando-se principalmente na métrica de ROC-AUC, com valor de 0,78, seguida de acurácia (0,72) e precisão (0,69). Em comparação, o XGBoost e o LightGBM mantêm desempenhos competitivos, com métricas próximas, especialmente na acurácia (ambos com 0,71) e no ROC-AUC (0,69).

Os algoritmos baseados em *boosting* mostraram superioridade em relação às demais abordagens, sendo mais indicados para tarefas que exigem melhor equilíbrio entre as métricas de avaliação. Essa análise está em consonância com a literatura especializada, que destaca os métodos de *ensemble* como estratégias eficazes na melhoria do desempenho preditivo em contextos de classificação supervisionada.

A seguir, apresenta-se o desempenho dos modelos com a aplicação de ajustes de hiperparâmetros.

Tabela 4 - Desempenho dos modelos de aprendizado de máquina com ajustes de hiperparâmetros

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**ROC-AUC**
CatBoost	0,72	0,65	0,69	0,67	0,78
XGBoost	0,71	0,63	0,69	0,66	0,71

LightGBM	0,70	0,63	0,69	0,66	0,70
Linear SVC	0,67	0,60	0,62	0,61	0,66
SVM	0,68	0,64	0,52	0,57	0,66
Random Forest	0,67	0,61	0,59	0,60	0,66
KNN	0,66	0,60	0,55	0,57	0,65
Logistic Regression	0,66	0,59	0,63	0,61	0,66
Decision Tree	0,65	0,57	0,60	0,59	0,54

Fonte: elaborado pelo autor com biblioteca Python, 2025.

Após a aplicação de técnicas de otimização de hiperparâmetros, os modelos de aprendizado de máquina foram reavaliados. A tabela apresenta o desempenho comparativo dos algoritmos, destacando os modelos CatBoost, XGBoost e LightGBM que apresentaram os melhores resultados, especialmente na métrica ROC-AUC. O CatBoost, em particular, obteve o valor mais elevado (0,78), demonstrando superior capacidade preditiva. Com destaque para o aumento do Recall (0,69). Os resultados indicam que a otimização de hiperparâmetros contribuiu significativamente para o aprimoramento dos modelos, aumentando sua performance preditiva em um cenário de desbalanceamento de classes.

A seguir, o desempenho dos modelos com a aplicação de SMOTE (*Synthetic Minority Over-sampling Technique*) é apresentado.

Tabela 5 - Desempenho comparativo de modelos preditivos aplicando SMOTE

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**ROC-AUC**
CatBoost	0,72	0,64	0,69	0,66	0,78
XGBoost	0,71	0,63	0,69	0,66	0,71
LightGBM	0,70	0,63	0,69	0,66	0,70
Linear SVC	0,67	0,62	0,62	0,61	0,66
SVM	0,68	0,64	0,52	0,57	0,66
Random Forest	0,67	0,60	0,59	0,60	0,66
KNN	0,66	0,60	0,55	0,57	0,65
Logistic Regression	0,66	0,59	0,63	0,61	0,66
Decision Tree	0,65	0,57	0,60	0,59	0,64

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela 5 apresenta a comparação entre os principais algoritmos de aprendizado de máquina quanto às métricas de avaliação após a aplicação da técnica de balanceamento de classes SMOTE. Observa-se que, mesmo após a aplicação do balanceamento, os modelos CatBoost, XGBoost e LightGBM mantêm-se como os de melhor desempenho. O CatBoost se destaca com ROC-AUC de 0,78 e acurácia de 0,72. Os modelos XGBoost e LightGBM apresentam valores próximos para acurácia (0,71 e 0,70, respectivamente), precisão (0,63), e

Recall (0,69), sendo consistentes em múltiplas métricas, o que demonstra estabilidade após o balanceamento das classes.

A análise comparativa entre diferentes algoritmos de classificação demonstraram que os modelos CatBoost, XGBoost e LightGBM apresentaram os melhores desempenhos em múltiplas métricas, mesmo após a aplicação da técnica de balanceamento SMOTE. No entanto, para problemas como a evasão escolar, que requer atenção na escolha das métricas de avaliação mais apropriadas, a acurácia não deve ser utilizada como principal critério de avaliação, visto que pode ocultar erros significativos, sobretudo em cenários com classes desbalanceadas (SOKOLOVA; LAPALME, 2009).

As métricas recall e F1-score são mais adequadas ao problema em questão. O recall, ou sensibilidade, mede a capacidade do modelo em identificar corretamente os estudantes que efetivamente evadiram, classe de maior interesse neste estudo. Sua importância reside no fato de que a não identificação de um aluno com risco de evasão pode ter consequências educacionais e sociais relevantes (ARTHANA; MAYSANJAYA; PRADNYANA; DANTES, 2024). Quanto à métrica F1-score, que representa a média harmônica entre precisão e recall, é relevante quando se deseja equilibrar a taxa de falsos positivos e a capacidade de identificar alunos evadidos (BAKARIWIE; ASAMOAH; DUWIEJUAH, 2025).

A análise demonstra que o modelo CatBoost obteve o melhor resultado em recall (0,69) e F1-score (0,67) no cenário com hiperparâmetros, indicando sua superioridade na identificação da classe minoritária, sem comprometer o equilíbrio geral.

De forma complementar, a métrica ROC-AUC demonstrou-se uma boa ferramenta de avaliação geral da capacidade discriminatória dos modelos, especialmente no caso do CatBoost, que alcançou o valor mais elevado (0,78). No entanto, seu uso é mais recomendado como suporte à análise, não como critério único para decisão.

Assim, no contexto da previsão de evasão escolar, a escolha dos modelos deve priorizar aqueles que apresentam os melhores valores de recall e F1-score, pois estas métricas estão diretamente relacionadas à missão de identificar os estudantes com maior risco de evasão. O uso da ROC-AUC como métrica secundária auxilia na avaliação da robustez do modelo, enquanto a acurácia deve ser interpretada com cautela e não ser utilizada isoladamente como métrica de desempenho principal.

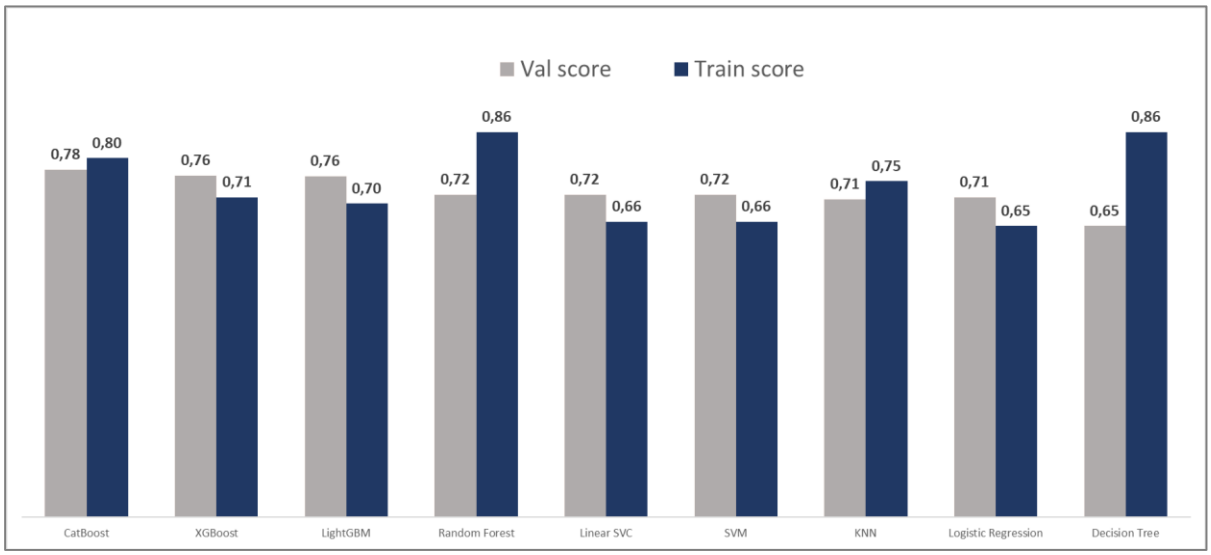
4.5 Desempenho dos modelos

A avaliação do desempenho dos modelos foi realizada com base nos resultados obtidos por meio da validação cruzada. Para essa análise, foram considerados os valores médios dos *scores* de validação (*val score*) e de treinamento (*train score*). O valor médio de validação representa o desempenho geral do modelo ao longo dos diferentes *folds* da validação cruzada, sendo um indicador importante da capacidade de generalização.

O escore de treinamento, por sua vez, permitiu avaliar a ocorrência de *overfitting*. Quando o valor de *train score* é significativamente superior ao valor de *val score*, pode-se inferir que o modelo está se ajustando excessivamente aos dados de treino, apresentando, portanto, baixa capacidade de generalização. Adicionalmente, mesmo que a média do AUC seja elevada, um desvio padrão alto indica variabilidade nos resultados entre os *folds*, o que pode comprometer a robustez do modelo.

A figura a seguir apresenta o desempenho dos classificadores avaliados com base nessas métricas.

Figura 3 - Desempenho do treinamento dos modelos com *cross validation*



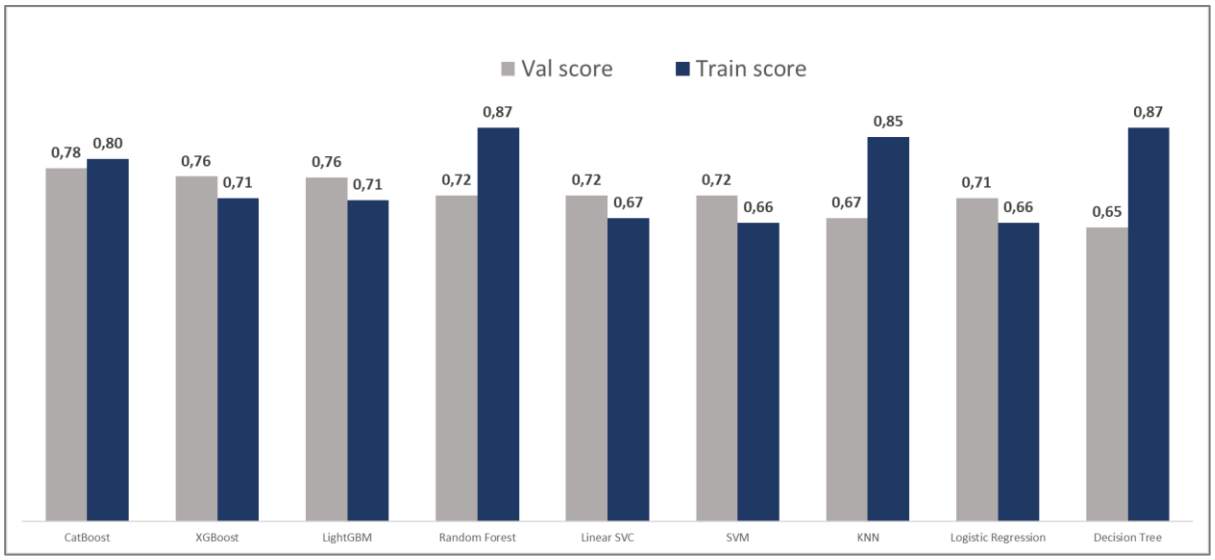
Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 6 apresenta o desempenho inicial de diferentes algoritmos de aprendizado de máquina aplicados à previsão de evasão escolar, utilizando os *scores* obtidos nos conjuntos de treinamento (*train score*) e validação (*val score*), considerando valores médios obtidos por meio de validação cruzada (*cross validation*). Não foram aplicadas técnicas de otimização de hiperparâmetros, nem estratégias de balanceamento das classes.

Observa-se que os modelos CatBoost, XGBoost e LightGBM obtiveram valores equilibrados entre treino e validação, evidenciando maior capacidade de generalização mesmo

sem ajustes avançados. O CatBoost, por exemplo, apresentou 0,80 de acurácia no treino e 0,78 na validação, sendo um dos mais estáveis.

Figura 4 - Desempenho do treinamento dos modelos com *cross validation* (com hiperparâmetros)

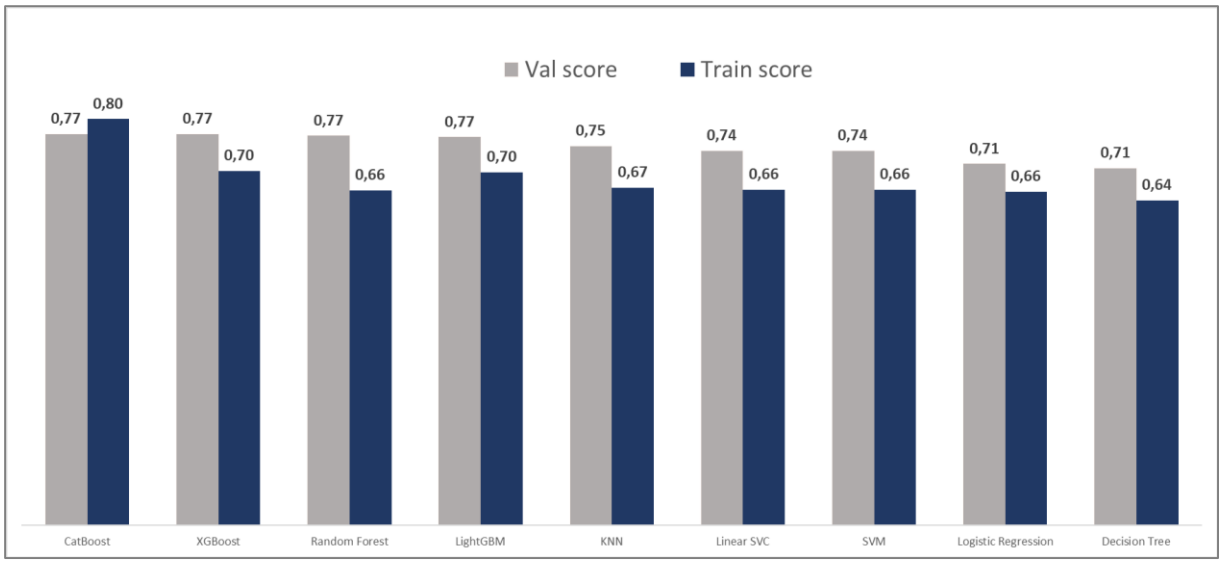


Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 7 apresenta os *scores* médios de desempenho obtidos por diferentes modelos de aprendizado de máquina após a aplicação de validação cruzada (*cross validation*) e ajuste de hiperparâmetros. As métricas utilizadas referem-se aos conjuntos de treinamento *(train score*) e validação (*val score*), com o objetivo de avaliar a capacidade de generalização dos algoritmos na tarefa de predição da evasão escolar.

Observa-se que os modelos CatBoost, XGBoost e LightGBM apresentaram bom equilíbrio entre os *scores* de treino e validação, destacando-se o CatBoost, com valores de 0,80 (treino) e 0,78 (validação). Este resultado sugere que os algoritmos baseados em técnicas de *boosting* demonstraram elevada capacidade de generalização, mesmo diante de um problema de classificação binária com leve desbalanceamento de classes.

Figura 5 - Desempenho do treinamento dos modelos com *cross validation* - com SMOTE



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 8 apresenta o desempenho comparativo dos modelos de aprendizado de máquina aplicados, sendo que, foi utilizada validação cruzada (*cross validation*) aliada à técnica de balanceamento por *oversampling* (SMOTE), mas sem aplicação de otimização de hiperparâmetros.

De modo geral, os modelos CatBoost, XGBoost, Random Forest e LightGBM destacaram-se com boas performances de validação (acima de 0,76), sendo o CatBoost o mais consistente, com 0,80 no treino e 0,77 na validação, o que demonstra bom equilíbrio e capacidade preditiva.

4.6 Teste e avaliação do melhor modelo

Após a análise comparativa entre diferentes algoritmos de classificação supervisionada, o modelo CatBoost foi selecionado para realizar a predição da evasão escolar. A escolha foi fundamentada nos resultados obtidos ao longo do processo de avaliação, no qual o desempenho dos modelos foi mensurado por meio de métricas relevantes para o contexto do problema, tais como precisão, recall, F1-score e ROC-AUC.

O CatBoost destacou-se por apresentar valores superiores nessas métricas em relação aos demais algoritmos testados, demonstrando maior capacidade de identificar corretamente os casos de evasão (classe positiva) e de manter um equilíbrio adequado entre precisão e sensibilidade. Dessa forma, considerando a natureza do problema que exige atenção especial à

detecção de estudantes em risco de evasão, o desempenho consistente do CatBoost justifica sua adoção como o modelo preditivo final neste estudo.

Um diferencial relevante do CatBoost é sua capacidade nativa de lidar com variáveis categóricas, o que o torna particularmente vantajoso em contextos com grande presença desse tipo de dado, como é o caso deste estudo. Com o uso do CatBoost, não foi necessário aplicar técnicas tradicionais de pré-processamento como *Label Encoding* ou *One-Hot Encoding*, uma vez que o algoritmo realiza o tratamento interno dessas variáveis de forma eficiente. Isso reduziu a complexidade do pipeline de preparação dos dados, minimizando o risco de perda de informação e evitando a alta dimensionalidade que pode ser gerada por codificações como o *One-Hot* (CATBOOST, 2025).

Além disso, ao lidar internamente com as variáveis categóricas, o CatBoost preserva a semântica e as relações intrínsecas entre as categorias, o que contribui para uma modelagem mais fiel ao comportamento real dos dados. Tais características, aliadas ao desempenho consistente observado nas validações, justificam plenamente a adoção do CatBoost como modelo final neste trabalho (PROKHORENKOVA et al., 2017).

A avaliação do modelo final foi realizada com base no conjunto de teste (*X_test*), que simula um cenário real de produção, composto por dados não utilizados durante o processo de treinamento. Essa abordagem garante uma estimativa mais fidedigna do desempenho preditivo do modelo em situações reais.

Dado que o problema é de classificação binária com classe desbalanceada, a simples utilização da acurácia como métrica de desempenho não é suficiente. Por esse motivo, foram consideradas métricas mais robustas, como:

ROC-AUC (Área sob a Curva ROC): para avaliar a capacidade discriminativa do modelo entre as classes;

F1-score: média harmônica entre precisão e recall, especialmente útil em cenários de desbalanceamento;

Precisão e Recall: medem, respectivamente, a proporção de verdadeiros positivos entre as previsões positivas e entre os casos efetivamente positivos;

PR-AUC (Área sob a Curva Precisão-Recall): adequada para bases de dados desbalanceadas;

Índice de Gini: diretamente derivado da métrica ROC-AUC, mede o poder de discriminação do modelo;

Brier score: avalia a calibração das probabilidades preditas; e

Acurácia: incluída como métrica complementar.

Como o objetivo foi classificar e ordenar os alunos com base na probabilidade de evasão, não foi necessária a calibração adicional das probabilidades. As saídas do modelo já foram consideradas adequadas para esse propósito, funcionando como um sistema de ranqueamento para identificar os estudantes com maior probabilidade de evadir.

A etapa de avaliação do modelo constituiu uma das fases que requer mais atenção no ciclo de desenvolvimento de soluções de aprendizado de máquina, uma vez que permite estimar sua capacidade de generalização para dados não observados (DOMINGOS, 2012).

Para garantir uma estimativa robusta do desempenho preditivo, o modelo foi avaliado por meio da técnica de validação cruzada estratificada, a qual assegura a preservação da proporção entre classes em cada subdivisão dos dados, reduzindo o risco de viés na avaliação (JAMES et al., 2021). A métrica de desempenho adotada foi a Área sob a Curva ROC (AUC), pois reflete a capacidade do modelo em distinguir entre classes independentemente do ponto de corte adotado (FAWCETT, 2006).

O modelo foi configurado com os parâmetros *auto_class_weights*='Balanced', para compensar o leve desbalanceamento do alvo (evadido), e *eval_metric*='AUC', priorizando a capacidade discriminativa entre os alunos que irão ou não evadir.

Os resultados obtidos no conjunto de dados de teste, que simula a entrada de novos dados em ambiente produtivo, foram analisados e interpretados com o objetivo de fornecer *insights* sobre os padrões associados à evasão escolar. Essa análise final permitiu verificar a aplicabilidade do modelo e seu potencial de impacto no contexto educacional analisado.

Tabela 6 - Métricas de avaliação do modelo CatBoost no conjunto de teste

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
0 (Não evadido)	0,77	0,74	0,75	17.102
1 (Evadido)	0,65	0,69	0,67	12.049
Acurácia geral			**0,72**	29.151
Média macro	0,71	0,71	0,71	
Média ponderada	0,72	0,72	0,72	

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela 6 apresenta as principais métricas de desempenho do modelo CatBoost aplicadas ao conjunto de teste. O modelo obteve uma acurácia geral de 72%, indicando bom desempenho na classificação correta dos exemplos. No entanto, a análise isolada das classes

revela informações mais relevantes, especialmente considerando o leve desbalanceamento da variável alvo (evadido).

Para a classe 0 (não evadido), o modelo alcançou uma precisão de 77% e um recall de 74%, resultando em um F1-score de 0,75, demonstrando maior segurança ao identificar corretamente os estudantes que permaneceram. Já para a classe 1 (evadido), os resultados foram 65% de precisão, 69% de recall e F1-score de 0,67, evidenciando um desempenho razoável na identificação de alunos evadidos.

As médias macro e ponderada foram, respectivamente, 0,71 e 0,72 nas três métricas principais (precisão, recall e F1-score), o que demonstra que o modelo mantém um desempenho equilibrado entre as classes, mesmo com o desbalanceamento presente.

Os resultados reforçam que o modelo é adequado para a tarefa de classificação binária de evasão escolar, contribuindo para uma tomada de decisão orientada por dados. Como destaca Fawcett (2006), métricas como o F1-score e a área sob a curva ROC são especialmente úteis em contextos com classes desbalanceadas, complementando a interpretação da acurácia.

Tabela 7 - Métricas de desempenho global do modelo CatBoost (conjunto de teste)

Métrica	**Valor**
Acurácia	0,7183
Precisão	0,7213
Recall	0,6893
F1-score	0,6692
ROC-AUC	0,7785
Índice de Gini	0,5569
PR-AUC	0,7139
Brier Score	0,1908

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela 7 resume as principais métricas de avaliação do modelo CatBoost ajustado, destacando sua capacidade preditiva para o problema de evasão escolar em cursos técnicos. A acurácia de 72% indica que o modelo acertou a previsão em cerca de 72% dos casos, o que é compatível com os resultados observados em tarefas de classificação com dados reais e moderadamente desbalanceados.

A precisão (0,72) e o recall (0,69) indicam que o modelo possui um bom equilíbrio entre a identificação correta de casos positivos (evadido) e a contenção de falsos positivos. A métrica de precisão foi ponderada com uso do parâmetro *average*='weighted', que é uma abordagem que busca fornecer uma média mais representativa do desempenho do modelo quando há desequilíbrio entre as classes. Em problemas de classificação binária ou multiclasse, como o

tratado neste estudo, as classes podem estar desbalanceadas, ou seja, uma classe pode conter significativamente mais exemplos do que outra. Ao usar *average*='weighted', a precisão de cada classe é multiplicada pela sua proporção de amostras no conjunto de teste. A precisão ponderada fornece uma visão mais realista e equilibrada do desempenho do modelo ao considerar esse desbalanceamento, sendo, portanto, mais apropriada para a avaliação global do classificador.

O F1-score de 0,67, que combina ambas as métricas, reforça essa estabilidade, mesmo diante de possíveis assimetrias na distribuição das classes.

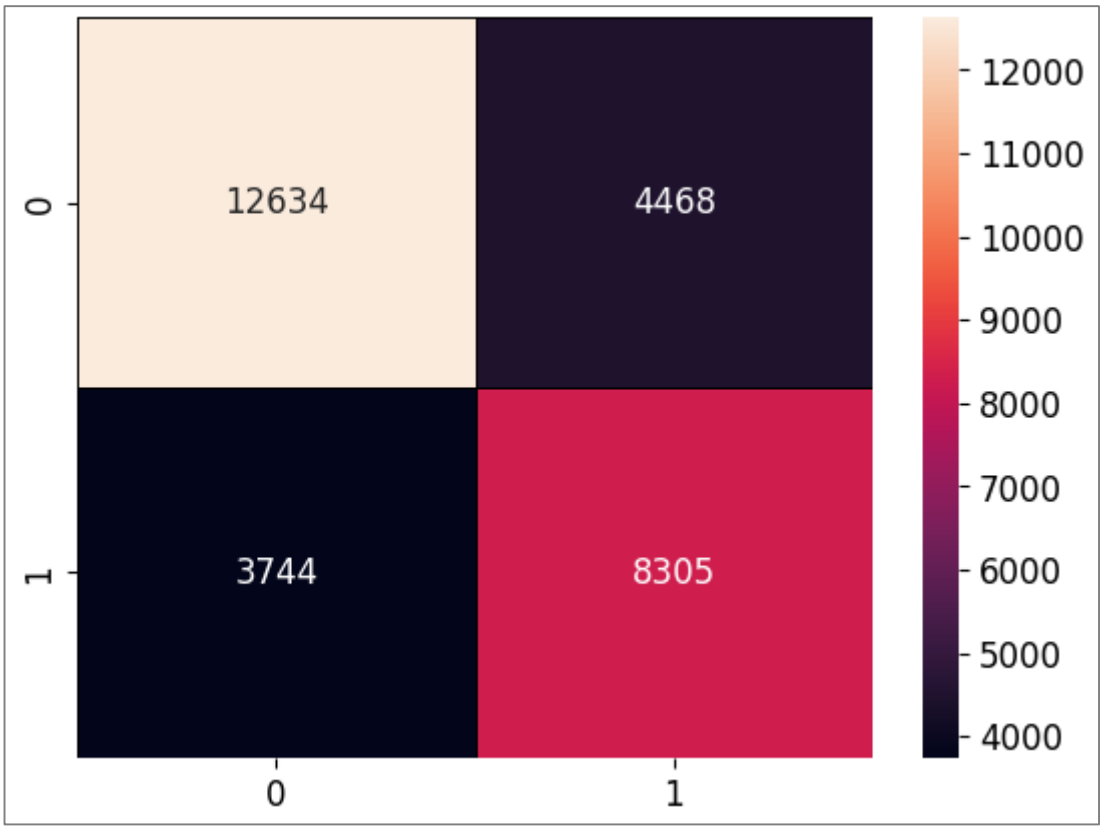
A métrica ROC-AUC (0,78) revela uma boa capacidade discriminativa do modelo entre as classes, conforme recomendação de Fawcett (2006).

O Índice de Gini, equivalente a 0,56, reforça a performance do modelo, já que esse índice é uma transformação da métrica ROC-AUC (Gini = 2 x ROC-AUC - 1). A PR-AUC (Área sob a Curva de Precisão-Recall) foi de 0,7139, indicando um bom desempenho mesmo em situações de desbalanceamento da classe positiva. Por fim, o Brier Score (0,19), que mede a calibração das probabilidades preditas, apresenta um valor satisfatório para modelos de classificação probabilística.

Os resultados confirmam que o modelo atende aos requisitos do projeto, sendo capaz de classificar os estudantes com desempenho consistente e confiável, além de apresentar métricas robustas para análises posteriores e tomada de decisão baseada em evidências.

A análise da matriz de confusão possibilitou compreender não apenas a acurácia global do modelo, mas também seus erros específicos, particularmente em contextos em que o desbalanceamento de classes pode mascarar o verdadeiro desempenho. Isso é especialmente relevante em contextos educacionais, onde prever corretamente a evasão pode guiar ações preventivas (HANEY et al., 2018).

Figura 6 - Matriz de confusão do modelo CatBoost para predição de evasão escolar



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A matriz de confusão apresentada resume o desempenho do modelo CatBoost na tarefa de classificação binária, distinguindo entre alunos que evadiram (1) e não evadiram (0). Os resultados são interpretados da seguinte forma:

Verdadeiros negativos: 12.634 alunos que não evadiram e foram corretamente classificados como tal.

Falsos positivos: 4.468 alunos que não evadiram, mas foram incorretamente classificados como evadidos.

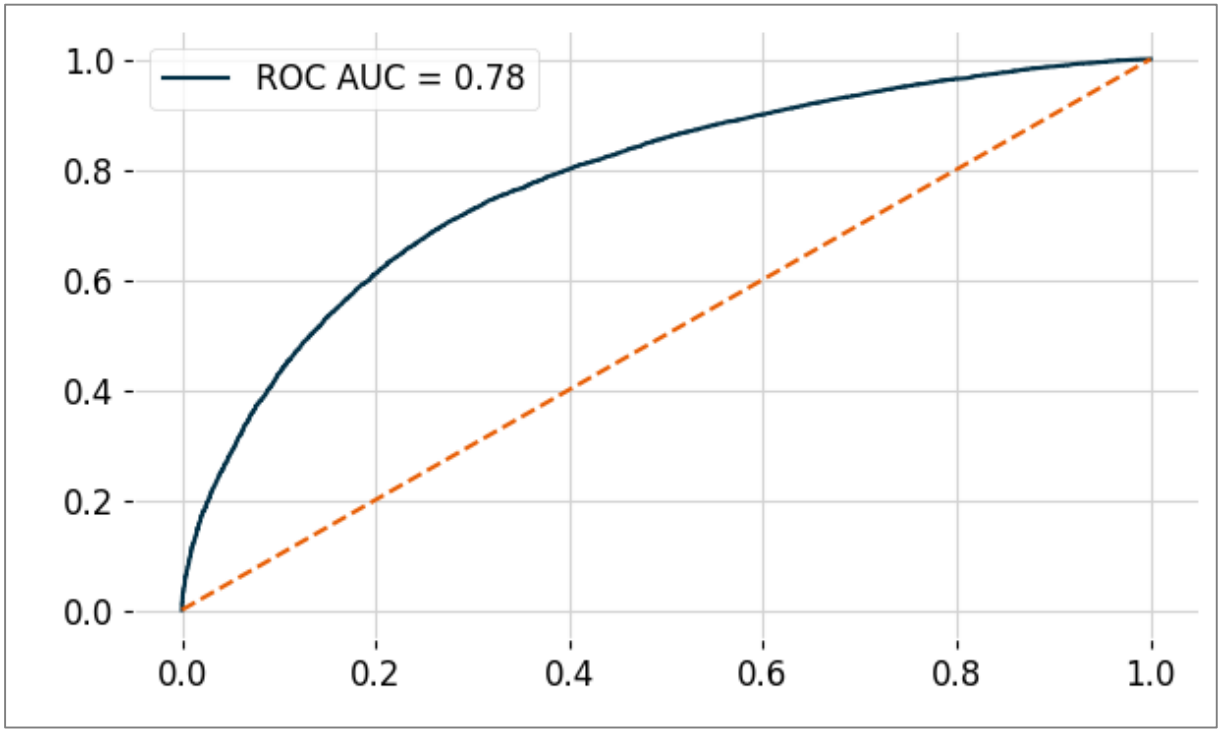
Falsos negativos: 3.744 alunos que evadiram, mas foram classificados como não evadidos.

Verdadeiros positivos: 8.305 alunos corretamente identificados como evadidos.

É possível observar um desempenho satisfatório do modelo, com maior acerto na classe majoritária (não evadido), mas ainda com boa capacidade de identificar corretamente os alunos que evadiram, demonstrando a capacidade discriminativa do classificador.

A Curva ROC (*Receiver Operating Characteristic*) foi utilizada para avaliar o desempenho do modelo de classificação CatBoost na distinção entre estudantes evadidos e não evadidos. A curva representa a taxa de verdadeiros positivos em função da taxa de falsos positivos para diferentes limiares de classificação.

Figura 7 - Curva ROC do modelo CatBoost para predição de evasão escolar

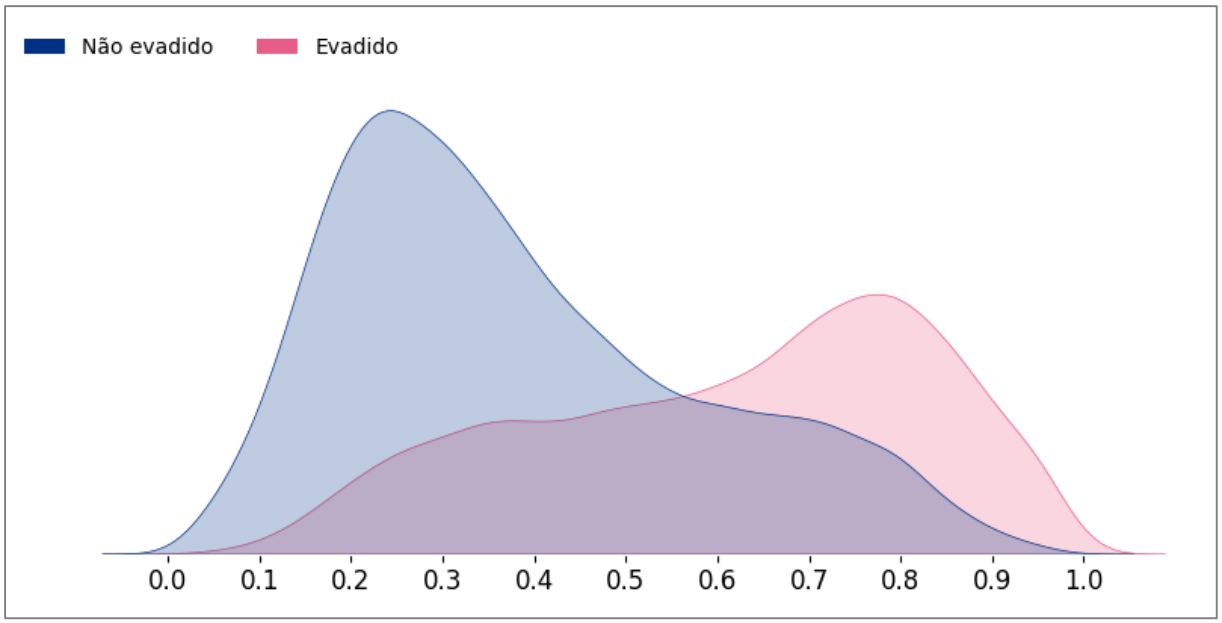


Fonte: elaborado pelo autor com biblioteca Python, 2025.

O valor da área sob a curva (AUC = 0,78) indica uma boa capacidade discriminativa do modelo, visto que valores próximos de 1 sugerem excelente desempenho, enquanto valores próximos de 0,5 indicam desempenho semelhante ao acaso (FAWCETT, 2006). Assim, o modelo demonstra habilidade considerável em separar corretamente as duas classes, mesmo diante de um leve desbalanceamento na distribuição da variável alvo (SAITO; REHMSMEIER, 2015).

A análise da distribuição das probabilidades de evasão previstas pelo modelo CatBoost, separando os casos reais de alunos evadidos e não evadidos é útil para avaliar a capacidade discriminativa do modelo, ou seja, sua habilidade de atribuir diferentes probabilidades a cada classe (MACHADO; LIMA, 2021). O gráfico de densidade kernel (KDE) mostra a distribuição das probabilidades previstas de evasão para dois grupos: alunos que evadiram (em rosa) e alunos que não evadiram (em azul).

Figura 8 - Distribuição das probabilidades previstas para alunos evadidos e não evadidos



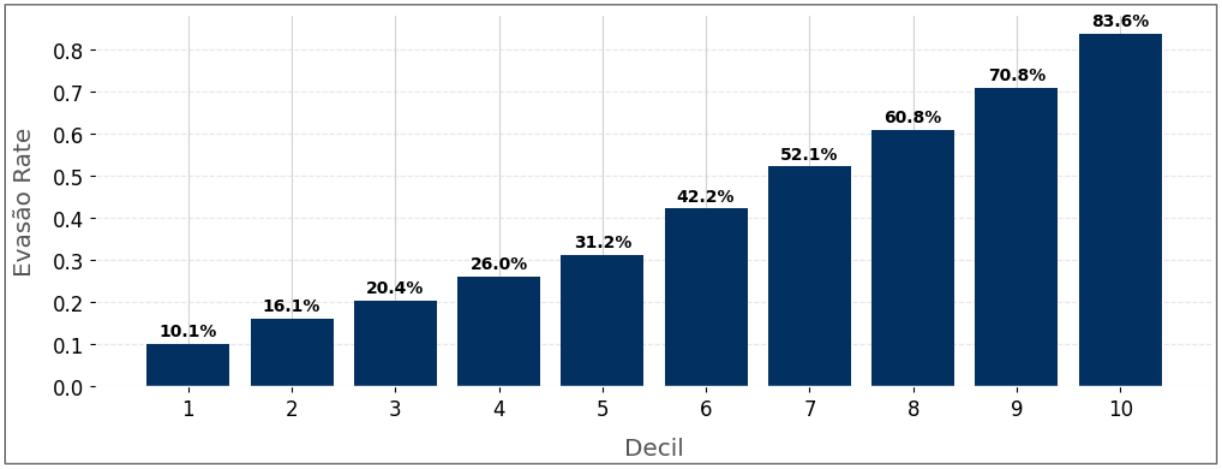
Fonte: elaborado pelo autor com biblioteca Python, 2025.

Observa-se que os alunos identificados como não evadidos concentram-se em probabilidades mais baixas (pico aproximadamente de 0,6), enquanto os alunos evadidos apresentam maior densidade em probabilidades superiores a 0,6. A sobreposição entre as curvas, no entanto, revela a existência de uma faixa intermediária onde as classes se confundem, o que pode dificultar a tomada de decisão em limiares próximos de 0,5.

O gráfico mostra que o modelo consegue diferenciar razoavelmente bem entre evadidos e não evadidos, mas não perfeitamente.

A utilização do método de decil permite uma visualização clara e prática da distribuição do risco, facilitando a priorização de intervenções em grupos com maior probabilidade de evasão (MENEZES; LIMA, 2020).

Figura 9 - Ordenação dos *scores* de probabilidade - Taxa de evasão por decil



Fonte: elaborado pelo autor com biblioteca Python, 2025.

O gráfico apresentado ilustra a taxa de evasão distribuída por decil, ordenada de acordo com *scores* de probabilidade, evidenciando um aumento progressivo da taxa de evasão do primeiro ao décimo decil. O primeiro decil apresenta uma taxa de evasão de 10,1%, enquanto o décimo decil alcança 83,6%, indicando que indivíduos com *scores* mais altos têm maior probabilidade de evasão.

Este comportamento é consistente com estudos que apontam para a eficácia da segmentação em decisões baseadas em modelos preditivos para identificar grupos com diferentes riscos de evasão (SANTOS et al., 2019).

Além disso, a tendência crescente da taxa de evasão reforça a validade do modelo de predição utilizado, pois confirma a capacidade de discriminar grupos conforme o risco estimado.

5 CONCLUSÃO

O estudo descreveu as etapas de análise e preparação dos dados, correspondentes a uma das fases essenciais da realização da modelagem computacional para realizar aprendizado de máquina. Entre essas etapas, destacam-se o entendimento do negócio e a normalização dos dados, as quais são fundamentais para garantir a qualidade e a consistência das informações utilizadas. Tais procedimentos exigiram uma execução criteriosa, sobretudo diante do grande volume de dados envolvidos. Em função dessa complexidade, a metodologia proposta neste estudo buscou a aplicabilidade em um projeto real na área da educação profissional e tecnológica.

Ao concluir este estudo, a principal contribuição consiste na implementação de uma metodologia baseada em técnicas de aprendizado de máquina para predição da evasão em cursos técnicos da RFEPCT. O objetivo principal foi aplicar diversos algoritmos para predição da probabilidade de um estudante evadir, com base em informações sociodemográficas e características do curso realizado.

Para tanto, foram adotadas abordagens supervisionadas de aprendizado de máquina. Os resultados indicam que os algoritmos baseados em *boosting* (CatBoost, XGBoost e LightGBM) apresentaram desempenho superior, mesmo após o reequilíbrio da base de dados por meio do SMOTE. Esses modelos apresentam maior capacidade de generalização e adaptabilidade frente ao balanceamento artificial de classes, sendo, portanto, os mais indicados para problemas com desbalanceamento na variável alvo.

Destaca-se o CatBoost que apresentou melhor desempenho preditivo, demonstrando bons resultados em precisão, sensibilidade e especificamente quanto ao ROC-AUC, que demonstrou a capacidade discriminatória do modelo em diferenciar alunos evadidos e não evadidos.

A evasão em cursos técnicos é um processo social que demanda ações complexas de prevenção e acompanhamento, por meio da implementação de políticas públicas que incentivem a permanência escolar.

Dessa forma, esta modelagem computacional visa contribuir em termos teóricos, por subsidiar políticas institucionais mais eficazes, promovendo uma gestão baseada em dados e possibilitando a identificação precoce de estudantes em risco de evasão. Do ponto de vista prático, apresenta uma metodologia que permite, a partir de variáveis categóricas e contínuas, identificar o risco de evasão escolar, o que viabiliza a intervenção antecipada por parte da gestão educacional.

Cabe ressaltar que a base de dados utilizada se refere exclusivamente às matrículas em situação de finalização no ano de 2023 e apresenta um número limitado de variáveis disponibilizadas para o acompanhamento e a gestão da RFEPCT.

Dessa forma, trabalhos futuros podem ampliar a robustez dos modelos preditivos e aprofundar a compreensão dos fatores relacionados à evasão escolar em cursos técnicos. Há possibilidades de continuidade da pesquisa como: a utilização de série histórica de matrículas para verificar se os valores preditos permanecem; a exploração de outras técnicas e algoritmos de aprendizado de máquina; a definição de outros parâmetros para otimização dos modelos testados; e a realização dos testes realizados considerando outras variáveis para verificar se melhora o poder preditivo do modelo CatBoost.

Declaração de IA Generativa e tecnologias assistidas por IA em processo de escrita

Durante a preparação deste trabalho, o autor utilizou o ChatGPT-4 e o DeepSeek Latest Version para melhorar a legibilidade e a linguagem. Após o uso desta ferramenta, o autor revisou e editou o conteúdo conforme necessário e assume total responsabilidade pelo conteúdo da publicação.

REFERÊNCIAS

ALVAREZ, K. R.; ALVES, S. C.; MATOS, R. P. School dropout in technical courses integrated to the Federal Network high school: Survey of motivational factors and

intervention proposals. Research, Society and Development, v. 10, n. 6, p. e12510615630, 2021. DOI: 10.33448/rsd-v10i6.15630.

ARTHANA, I. Ketut Resika; MAYSANJAYA, I. Made Dendi; PRADNYANA, Gede Aditra; DANTES, Gede Rasben. Optimizing Dropout Prediction in University Using Oversampling Techniques for Imbalanced Datasets. International Journal of Information and Education Technology, v. 14, n. 8, p. 1052–1060, ago. 2024. Disponível em: https://www.ijiet.org/vol14/IJIET-V14N8-2133.pdf. Acesso em: 28 maio 2025.

BAKARIWIE, Amiru; ASAMOAH, Dominic; DUWIEJUAH, Abudu Ballu. Prevention of student attrition: a data-backed approach to school counselling using Delphi technique and multiple classification algorithms. Discover Education, v. 4, n. 1, p. 1–13, 31 jul. 2025. Disponível em: https://doi.org/10.1007/s44217-025-00494-7. Acesso em: 28 maio 2025.

BAKER, R. S.; SIEMENS, P. S. Educational data mining and learning analytics. In: SAWYER, R. K. (ed.). The Cambridge Handbook of the Learning Sciences. 2. ed. Cambridge: Cambridge University Press, 2014. p. 253–274. DOI: https://doi.org/10.1017/cbo9781139519526.016.

BESSEY, D.; BACKES-GELLNER, U. Regional unemployment and educational attainment in vocational training. Economics of Education Review, v. 44, p. 1–18, 2015. Disponível em: https://doi.org/10.1016/j.econedurev.2014.10.003. Acesso em: 29 maio 2025.

BÖHN, S.; DEUTSCHER, V. Determinants of dropout in vocational education and training: a systematic review. Empirical Research in Vocational Education and Training, v. 14, n. 3, p. 1-25, 2022.

BURGES, Christopher J. C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, v. 2, n. 2, p. 121–167, 1998.

BRANCO, Paula; TORGOSO, Luís; RIBEIRO, Rita P. A Survey of Predictive Modeling on Imbalanced Domains. ACM Computing Surveys, v. 49, n. 2, p. 1–50, 2016. DOI: https://doi.org/10.1145/2907070.

BRASIL. Ministério da Educação. Guia de Referência Metodológica. Brasília, DF: MEC; dadosabertos.mec.gov.br, 2019. Disponível em: https://dadosabertos.mec.gov.br/images/pdf/grm-2020-isbn-revisado.pdf. Acesso em: 1 de ago. 2025.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha – Microdados da Eficiência Acadêmica – 2023. Brasília: MEC, 2023. Disponível em: https://dadosabertos.mec.gov.br/pnp/item/261-2023-microdados-eficiencia-academica. Acesso em: 31 maio 2025.

BRASIL. Presidência da República. Lei nº 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica. Diário Oficial da União: seção 1, Brasília, DF, ano 145, n. 251, p. 1, 30 dez. 2008. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm. Acesso em: 31 maio 2025.

BRASIL. Centro de Gestão e Estudos Estratégicos (CGEE). A educação profissional e tecnológica no Brasil: análise e subsídios para políticas públicas. Brasília, DF: CGEE, 2023a. Disponível em: https://www.cgee.org.br. Acesso em: 27 maio 2025.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha (PNP). Brasília, DF: MEC, 2024. Disponível em: https://www.gov.br/mec/pt-br/pnp. Acesso em: 14 jul. 2024.

BRASIL Tribunal de Contas da União. Auditoria operacional na Rede Federal de Educação Profissional, Científica e Tecnológica: Relatório de Auditoria. Brasília: TCU, 2024a.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

BRIDGELAND, J. M.; DILIULIO, J. J.; BALFANZ, R. On the front lines of schools: perspectives of teachers and principals on the high school dropout problem. Washington, DC: Civic Enterprises; Peter D. Hart Research Associates; AT&T Foundation; America's Promise Alliance, 2009. 60 p.

CATBOOST documentation. Categorical features | CatBoost. Disponível em: https://catboost.ai/docs/en/features/categorical-features. Acesso em: 10 jun. 2025.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

CSEHNÉ PAPP, I.; HÉDER-RIMA, M.; DAJNOKI, K. Causes of student dropout in vocational training from the perspective of teachers. Practice and Theory in Systems of Education, v. 16, n. 2, p. 3–15, 2021. Disponível em: https://doi.org/10.1556/2063.16.2021.07. Acesso em: 29 maio 2025.

DELLINGER, J. Correlation, Spearman. In: ALLEN, M. (org.). The SAGE encyclopedia of communication research methods. v. 4. Thousand Oaks: SAGE Publications, 2017. p. 274–275. DOI: 10.4135/9781483381411.n101.

DOMINGOS, P. A few useful things to know about machine learning. Communications of the ACM, v. 55, n. 10, p. 78–87, 2012.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, v. 29, n. 2–3, p. 103–130, 1997.

DOMINGUES, R.; BUENO, T. M.; LENGLER, L.; SANTOS, R. M.; FERREIRA, A. C.; OLIVEIRA, M. R. Data preprocessing techniques for machine learning. Journal of Big Data, v. 9, n. 1, p. 1–26, 2022.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULYAEV, Andrey. CatBoost: gradient boosting with categorical features support. In: Proceedings of the Workshop on ML Systems at NIPS 2018. Montréal: NeurIPS, 2018. Disponível em: https://arxiv.org/abs/1810.11363. Acesso em: 20 de maio de 2025.

FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861–874, 2006.

GÉRON, Aurélien. Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: concepts, tools, and techniques to build intelligent systems. 2. ed. Sebastopol: O'Reilly Media, 2019.

GONÇALVES, L. R.; PASSOS, S. R. M. M. S. dos; PASSOS, Á. M. dos. Novos rumos para o ensino médio noturno: como e por que fazer? Ensaio: Avaliação e Políticas Públicas em Educação, Rio de Janeiro, v. 13, n. 48, p. 345-360, jul./set. 2005.

GREIG, M. J. Factors affecting modern apprenticeship completion in Scotland. Labor: Personnel Economics eJournal, 2019.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 4th ed. Amsterdam: Elsevier, 2022.

HANA, Ahmed; LOFSTEAD, Jay. Managing randomness to enable reproducible machine learning. In: ACM SYMPOSIUM ON PRINCIPLES OF DISTRIBUTED COMPUTING, 41., 2022, [S.l.]. Proceedings [...]. New York: ACM, 2022. Disponível em: <https://dl.acm.org/doi/10.1145/3526062.3536353\>. Acesso em: 4 ago. 2025.

HANEY, W.; RUSSELL, M.; GEBERT, J. Dropping Out: Why Students Drop Out of High School and What Can Be Done About It. Harvard University, 2018. Disponível em: https://dash.harvard.edu/. Acesso em: 20 maio 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The elements of statistical learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HOLTMMANN, A. C.; SOLGA, H. Dropping or stopping out of apprenticeships: the role of performance- and integration-related risk factors. Zeitschrift für Erziehungswissenschaft, v. 26, p. 469–494, 2023. DOI: https://doi.org/10.1007/s11618-023-01151-1.

HOSMER, D. W.; LEMESHOW, S. Applied logistic regression. 2. ed. New York: Wiley, 2000.

IBGE – Instituto Brasileiro de Geografia e Estatística. Recortes metropolitanos e aglomerações urbanas. Rio de Janeiro: IBGE, [2023]. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/18354-recortes-metropolitanos-e-aglomeracoes-urbanas.html. Acesso em: 31 maio 2025.

IBM. Introdução ao CRISP-DM. IBM Documentation, 2023. Disponível em: https://www.ibm.com/docs/pt-br/spss-modeler/18.4.0?topic=guide-introduction-crisp-dm. Acesso em: 15 jun. 2025.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: With Applications in R. 2. ed. New York: Springer, 2021.

KE, G.; MENG, Q.; FINLEY, T.; WANG, T.; CHEN, W.; MA, W.; YE, Q.; LIU, T.-Y. LightGBM: a highly efficient gradient boosting decision tree. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 31., 2017, Long Beach. Proceedings... Long Beach: [s.n.], 2017.

KUHN, M.; JOHNSON, K. Applied Predictive Modeling. New York: Springer, 2013.

KUHN, M.; JOHNSON, K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Boca Raton: CRC Press, 2019.

KRÜGER, João Gabriel Corrêa; BRITTO, Alceu; BARDDAL, Jean Paul. An explainable machine learning approach for student dropout prediction. 2023. Disponível em: https://doi.org/10.2139/ssrn.4253068. Acesso em: 27 maio 2025.

LONG, J. Scott. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage Publications, 1997.

MACHADO, R. A.; LIMA, E. A. A. Introdução à Ciência de Dados: fundamentos e aplicações. Rio de Janeiro: Elsevier, 2021.

MARTINEZ, W. L. Computational statistics handbook with MATLAB. 3. ed. Boca Raton: Chapman and Hall/CRC, 2019.

MATZ, Sandra C. et al. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, [S. l.], v. 13, n. 1, 5705, 2023. Disponível em: https://doi.org/10.1038/s41598-023-32593-6.

MEEUWISSE, M.; SEVERIENS, S. E.; BORN, M. P. Goals and perceptions of the learning environment of successful and unsuccessful ethnic minority students in university education. Learning and Individual Differences, v. 20, n. 5, p. 587–591, 2010a. Disponível em: https://doi.org/10.1016/j.lindif.2010.07.002. Acesso em: 29 maio 2025.

MENEZES, João; LIMA, Maria. Estratégias de mitigação da evasão escolar: uma abordagem preditiva. Revista Brasileira de Educação, v. 25, n. 78, p. 102-120, 2020.

MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997.

MURTHY, S. K. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery, v. 2, n. 4, p. 345–389, 1998.

PANG, G.; SHEN, C.; CAO, L.; VAN DEN HENGEL, A. Deep learning for anomaly detection: a review. ACM Computing Surveys, v. 54, n. 2, p. 1–38, 2021. Disponível em: https://doi.org/10.1145/3439950. Acesso em: 27 maio 2025.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; et al. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 27 maio 2025.

PILLAI, N. Vijayamohanan; RJU MOHAN, A. Perfect Multicollinearity and Dummy Variable Trap: Explaining the Unexplained. MPRA Paper, 2024. Disponível em: ttps://mpra.ub.uni-muenchen.de/120376/. Acesso em: 29 de maio 2025.

PROKHORENKOVA, Liudmila; GUSEV, Gleb; VOROBEV, Aleksandr; DOROGUSH, Anna Veronika; GULIN, Andrey. CatBoost: unbiased boosting with categorical features. 2017. Disponível em: https://doi.org/10.48550/arXiv.1706.09516. Acesso em: 27 maio 2025.

PYLE, Dorian. Data preparation for data mining. San Francisco: Morgan Kaufmann, 1999.

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

RIFKIN, R.; CLAO, A. Regularized linear classification with AdaBoost. In: Advances in Neural Information Processing Systems, v. 15. Cambridge: MIT Press, 2003.

RUMBERGER, R. W. et al. Family influences on dropout behavior in one California high school. Sociology of Education, v. 63, n. 4, p. 283-299, 1990. DOI: 10.2307/2112876.

RUSSELL, Stuart J.; NORVIG, Peter. Inteligência artificial. 3. ed. São Paulo: Elsevier, 2010.

SAITO, T.; REHMSMEIER, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PloS one, v. 10, n. 3, p. e0118432, 2015.

SANTOS, Pedro et al. Modelos preditivos para evasão: aplicação em ambientes educacionais. Ciência da Computação, v. 30, n. 3, p. 45-58, 2019.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427–437, 2009.

SPEARMAN, C. The proof and measurement of association between two things. International Journal of Epidemiology, Oxford, v. 39, n. 5, p. 1137–1150, 2010.

YI, H.; ALLEN, J.; SOLGA, H. Firm-Provided Training and the Career Prospects of Young Workers: Training Volumes, Signaling Effects, and Labor Market Outcomes. European Sociological Review, v. 31, n. 5, p. 525–540, 2015. Disponível em: https://doi.org/10.1093/esr/jcv062. Acesso em: 29 maio 2025.

WOODLEY, A.; SIMPSON, O. Student dropout: the elephant in the room. In: ZAWACKI-RICHTER, O.; ANDERSON, T. (ed.). On-line distance education: towards a research agenda. Edmonton: AU Press, 2014. p. 459-484.

5 ARTIGO 2

INTELIGÊNCIA ARTIFICIAL EXPLICÁVEL (XAI) E AS VARIÁVEIS DETERMINANTES PARA A EVASÃO EM CURSOS TÉCNICOS NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA DO BRASIL. UMA ABORDAGEM COM USO DO MÉTODO SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

Revista(s) alvo:

International Journal of Contemporary Educational Research - (https://ijcer.net/index.php/pub/Aboutjournal)

International Journal of Artificial Intelligence in Education – (https://link.springer.com/journal/40593)

RESUMO

A análise dos fatores associados à evasão escolar é um dos tópicos mais estudados no âmbito da Educação Profissional e Tecnológica devido ao seu impacto direto na formação de profissionais qualificados e no desenvolvimento socioeconômico. Na Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil, enfrentar o desafio da evasão requer um conhecimento aprofundado das variáveis que a influenciam, especialmente em cursos técnicos, onde altos índices de abandono podem comprometer a formação prática e a inserção dos estudantes no mercado de trabalho. A compreensão dessas variáveis exige um olhar aprofundado para os cursos e os fatores sociodemográficos, institucionais e individuais dos estudantes. Este estudo propõe-se a identificar e analisar as variáveis determinantes para a evasão em cursos técnicos da Rede Federal, utilizando técnicas de aprendizado de máquina. A abordagem combinou a análise documental e a revisão bibliográfica com o uso do algoritmo CatBoost para a classificação preditiva. O modelo foi treinado com base em variáveis pessoais, demográficas, institucionais e acadêmicas dos estudantes. O modelo obteve desempenho satisfatório, com um recall de 69% e uma área sob a curva ROC (AUC) de 78%. Para interpretar as saídas do modelo e determinar a importância e o impacto de cada variável no resultado, utilizou-se o método SHapley Additive exPlanations (SHAP). Essa metodologia fornece uma interpretação robusta dos fatores de risco, auxiliando gestores e formuladores de políticas públicas na identificação dos estudantes em situação de vulnerabilidade e na implementação de ações preventivas direcionadas. Ao fornecer subsídios para reduzir a evasão em cursos técnicos,

a pesquisa contribui para aumentar a eficácia acadêmica da Rede Federal, mitigar prejuízos sociais e financeiros e expandir as oportunidades de formação profissional no país.

Palavras-chave: evasão escolar; cursos técnicos; rede federal; aprendizado de máquina; SHAP.

ABSTRACT

The analysis of factors associated with school dropout is one of the most studied topics in the field of Professional and Technological Education, due to its direct impact on the training of qualified professionals and socioeconomic development. In the Federal Network of Professional, Scientific, and Technological Education of Brazil, addressing the challenge of dropout requires in-depth knowledge of the variables that influence it, especially in technical courses, where high dropout rates can compromise practical training and the insertion of students into the labor market. Understanding these factors requires an in-depth look at the sociodemographic, institutional, course, and individual factors of students. This study aims to identify and analyze the determining variables for dropout in technical courses in the Federal Network, using machine learning techniques. The approach combined documentary analysis and literature review with the use of the CatBoost algorithm for predictive classification. The model was trained based on students' personal, demographic, institutional, and academic variables. The model performed satisfactorily, with a recall of 69% and an area under the ROC curve (AUC) of 78%. To interpret the model outputs and determine the importance and impact of each variable on the result, the SHapley Additive exPlanations (SHAP) method was used. This methodology provides a robust interpretation of risk factors, helping managers and public policy makers identify students in vulnerable situations and implement targeted preventive actions. By providing support to reduce dropout rates in technical courses, the research contributes to increasing the academic effectiveness of the Federal Network, mitigating social and financial losses, and expanding professional training opportunities in the country.

Keywords: school dropout; technical courses; federal network; machine learning; SHAP.

1 INTRODUÇÃO

A evasão escolar é compreendida como o abandono do curso pelo estudante antes de sua conclusão. Constitui-se como uma das mais persistentes mazelas sociais, com efeitos profundamente prejudiciais para a sociedade. Em nível individual, representa a interrupção de trajetórias formativas, restringindo o acesso a oportunidades de crescimento pessoal e profissional. No âmbito econômico, compromete a constituição de uma força de trabalho

qualificada, resultando em menor produtividade, menor capacidade de inovação e perda de competitividade do país no cenário global (ROMAN et al., 2022).

O abandono escolar na educação profissional e tecnológica (EPT) representa uma barreira relevante para o progresso socioeconômico nacional, principalmente em um cenário de exigência crescente por mão de obra especializada. A desistência antecipada de cursos técnicos ou de formação profissional impede a conclusão da qualificação pelos estudantes, o que prejudica suas perspectivas de ingresso ou retorno ao emprego formal e restringe o acesso a futuras possibilidades de capacitação (HOLTMANN; SOLGA, 2023).

Os estudos sobre os fatores que impactam a conclusão de cursos técnicos são majoritariamente orientados por abordagens que focalizam o comportamento do aluno, enfatizando a identificação de suas motivações individuais para o abandono. Essa perspectiva, contudo, deixa em aberto uma lacuna significativa: a avaliação insuficiente de variáveis contextuais e institucionais que também influenciam ativamente a decisão do estudante de permanecer ou não no curso (PIEPENBURG; BECKMANN, 2021).

No Brasil, em auditoria realizada pelo Tribunal de Contas da União (TCU), em 2024, revelou-se que os cursos técnicos ofertados pela Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) apresentaram uma taxa média de evasão de 41%, evidenciando a gravidade do problema no âmbito da educação profissional pública federal (BRASIL, 2024).

O uso de técnicas de Inteligência Artificial (IA) e aprendizado de máquina tem se consolidado como uma ferramenta sólida e promissora para a identificação de padrões complexos e a previsão de comportamentos a partir da análise de grandes volumes de dados em diversas áreas do conhecimento, como saúde, finanças e ciências ambientais (JORDAN; MITCHELL, 2015).

No campo educacional, esse potencial ganha uma aplicação estratégica. As técnicas de mineração de dados educacionais (EDM), sustentadas por algoritmos de classificação, vêm sendo amplamente utilizadas para prever a probabilidade de evasão. Essa capacidade preditiva transforma a gestão educacional, permitindo uma atuação preventiva e personalizada por parte das instituições de ensino. Ao antecipar casos de abandono em potencial é possível implementar estratégias de retenção muito mais eficazes, direcionando apoio acadêmico e psicossocial aos estudantes em situação de vulnerabilidade (BAKER; SIEMENS, 2014).

Sob essa perspectiva, o algoritmo CatBoost (*Categorical Boosting*) destaca-se como uma ferramenta computacional particularmente eficaz para a predição de evasão em cursos técnicos ofertados pela RFEPCT. Desenvolvido especificamente para lidar com dados que

contêm variáveis categóricas, esse algoritmo supera limitações de outros métodos ao realizar o tratamento interno dessas categorias sem a necessidade de um pré-processamento extensivo (DOROGUSH et al., 2018).

Por outro lado, o método SHAP (SHapley Additive exPlanations) possibilita traduzir as previsões preditivas de modelos complexos, como CatBoost, em contribuições atribuíveis a cada variável, oferecendo explicações tanto locais (por instância) quanto globais (para o modelo inteiro), o que facilita identificar quais características sociodemográficas, acadêmicas e comportamentais mais influenciam a probabilidade de evasão em cursos técnicos (LIU; ZHOU; LIU, 2025; LUNDBERG; LEE, 2017).

Diante desse cenário, este artigo propõe-se a apresentar uma análise dos resultados da modelagem computacional realizada com o algoritmo CatBoost para identificar e explicar as variáveis determinantes da evasão em cursos técnicos da RFEPCT, utilizando o método SHAP (SHapley Additive exPlanations) para interpretar a influência de cada variável. A explicação das variáveis que têm maior impacto sobre a evasão escolar permitirá subsidiar ações institucionais preventivas e orientar a formulação de políticas públicas voltadas à permanência e ao êxito estudantil.

2 REVISÃO DA LITERATURA

A evasão escolar é definida como o ato de desligar-se ou abandonar um curso antes de sua conclusão. Embora essa definição seja objetiva, suscita uma série de reflexões sobre as possíveis causas do fenômeno, que envolvem uma multiplicidade de fatores. Entre eles, destacam-se as características individuais dos estudantes, os contextos familiares e sociais, bem como aspectos estruturais do sistema educacional e a percepção de relevância dos conteúdos curriculares oferecidos (BRASIL, 2023).

Estudos já realizados sobre a evasão escolar confirmam, ainda que parcialmente, os fatores que contribuem para o afastamento dos estudantes da EPT. Aqueles que não concluem um curso profissionalizante frequentemente atribuem sua desistência à ausência de apoio social, apontando esse fator como central para a decisão de abandonar os estudos (MEEUWISSE; SEVERIENS; BORN, 2010).

Para Dekker, Pechenizkiy e Vleeshouwers (2009) indicadores acadêmicos são fundamentais e podem ser utilizados de forma eficaz para prever o sucesso de um estudante. Berens et al. (2019) realizaram um amplo estudo utilizando dados acadêmicos de instituições

de ensino, demonstrando que métodos de aprendizado de máquina podem alcançar alta acurácia na predição precoce da evasão estudantil.

Os dados quantitativos referentes às matrículas e às taxas de evasão nos cursos técnicos da RFEPCT são sistematicamente consolidados e divulgados através da plataforma Nilo Peçanha (PNP), conforme estabelecido pelo Ministério da Educação (BRASIL, 2023). Esses indicadores são publicados anualmente, após o encerramento do período letivo, constituindo-se como importante ferramenta de monitoramento para gestores educacionais.

No ano de 2023, a eficiência acadêmica dos cursos técnicos dessa rede apresentou avanços em relação aos anos anteriores, especialmente no indicador de conclusão, que atingiu 47,02%, o maior percentual em comparação aos quatro anos anteriores. No entanto, a evasão escolar permaneceu em nível elevado, registrando 41,34% das matrículas do período (BRASIL, 2023).

A aplicação de técnicas de aprendizado de máquina para prever a evasão escolar tem sido bastante investigada na literatura, revelando tanto avanços metodológicos significativos quanto resultados práticos (GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020; HETTIARACHCHI; HARSHANATH, 2025; RASTROLLO-GUERRERO; GÓMEZ-PULIDO; DURÁN-DOMÍNGUEZ, 2020). Entre os métodos mais empregados destacam-se as árvores de decisão (MURTHY, 1998), as redes neurais artificiais (HASTIE; TIBSHIRANI; FRIEDMAN, 2009; MITCHELL, 1997), o Naïve Bayes (DOMINGOS; PAZZANI, 1997), o Random Forest (BREIMAN, 2001), o K-Nearest Neighbors (MITCHELL, 1997), a regressão logística (LONG, 1997) e as máquinas de vetores de suporte (BURGES, 1998). Esses algoritmos têm demonstrado grande potencial na identificação de padrões complexos em dados educacionais, contribuindo para a compreensão e predição do desempenho e da evasão estudantil.

Os métodos de aprendizado de máquina (AM) são geralmente classificados em três abordagens principais: supervisionada, não supervisionada e por reforço. No aprendizado supervisionado, o modelo é treinado com exemplos rotulados, visando aprender uma função de mapeamento capaz de generalizar para novos dados. O aprendizado não supervisionado busca identificar padrões ou agrupamentos em dados não rotulados, sendo aplicado em tarefas como redução de dimensionalidade e clustering. Já o aprendizado por reforço envolve a interação de um agente com o ambiente em que ações são avaliadas por recompensas ou penalidades, com o objetivo de maximizar uma função de retorno ao longo do tempo (RUSSELL; NORVIG, 2010).

Conforme fundamentado por Hastie, Tibshirani e Friedman (2009), o aprendizado de máquina estrutura-se em torno de três tarefas principais: a classificação, que associa observações a categorias pré-definidas; a regressão, empregada para prever valores contínuos; e o agrupamento (*clustering*), que organiza dados em grupos homogêneos sem o uso de rótulos prévios, sendo fundamental para segmentação e análise exploratória.

As aplicações de classificadores de aprendizado de máquina mostram-se eficazes em tarefas de classificação binária, especialmente em contextos preditivos como a identificação da evasão escolar (KRÜGER; BRITTO; BARDDAL, 2023).

Entre os diversos algoritmos existentes, os que utilizam *gradient boosting* têm ganhado destaque. O CatBoost é um algoritmo de aprendizado de máquina com foco especial no tratamento eficiente de variáveis categóricas. Diferencia-se de outros algoritmos de *boosting* por incorporar técnicas avançadas como *Ordered Boosting* e *Target Statistics*, que evitam o *target leakage* durante o treinamento. Além disso, o CatBoost realiza a conversão automática de atributos categóricos, reduzindo a necessidade de pré-processamento. O modelo é robusto, apresenta alta acurácia e tem se mostrado eficaz mesmo em conjuntos de dados heterogêneos e desbalanceados (DOROGUSH et al., 2018; PROKHORENKOVA et al., 2017).

2.1 SHAP (SHapley Additive exPlainations)

Aplicações recentes no campo educacional mostram que, quando combinado a classificadores de alto desempenho como o CatBoost, o SHAP revela padrões que permitem priorizar intervenções dirigidas aos estudantes mais vulneráveis (LIU; ZHOU; LIU, 2025).

O método SHAP (Shapley Additive Explanations) fundamenta-se na teoria dos valores de Shapley, da Teoria dos Jogos Cooperativos, com o objetivo de quantificar a contribuição de cada variável de entrada na previsão de um modelo de aprendizado de máquina. A explicação fornecida pelo SHAP segue a decomposição aditiva da forma:

$$f(x) = E[f(x)] + \sum_{i=1}^{M} \phi_i$$

onde $f(x)$ representa a predição do modelo para a instância x, $E[f(x)]$ é o valor esperado da predição no conjunto de dados, e ϕi é o valor SHAP da variável i, que indica sua contribuição marginal à predição (LIU et al., 2024; LUNDBERG; LEE, 2017).

Os valores SHAP quantificam a contribuição marginal de cada variável de entrada para a predição feita pelo modelo, permitindo uma decomposição aditiva da predição. Essa abordagem possibilita a visualização da importância global das variáveis e fornece explicações individualizadas para cada observação, com isso é possível obter maior transparência e confiança nos modelos (MOLNAR, 2022).

Este processo permite interpretar individualmente cada predição feita pelo modelo, fornecendo transparência e confiabilidade, especialmente em contextos sensíveis. A abordagem torna-se particularmente eficaz ao explicar modelos complexos de forma rigorosamente matemática e alinhada com princípios de justiça algorítmica promovidos pela Inteligência Artificial Explicável (XAI) (LUNDBERG et al., 2020).

A visualização dos valores SHAP por meio de ferramentas gráficas desempenha um papel essencial na interpretação dos resultados dos modelos de aprendizado de máquina. Dentre os principais recursos, destacam-se o *summary_plot*, o *heatmap*, o *beeswarm*, o *waterfall*, o *force_plot* e o *dependence_plot*, cada um oferecendo diferentes perspectivas sobre o impacto das variáveis de entrada nas predições (SHAP, 2025).

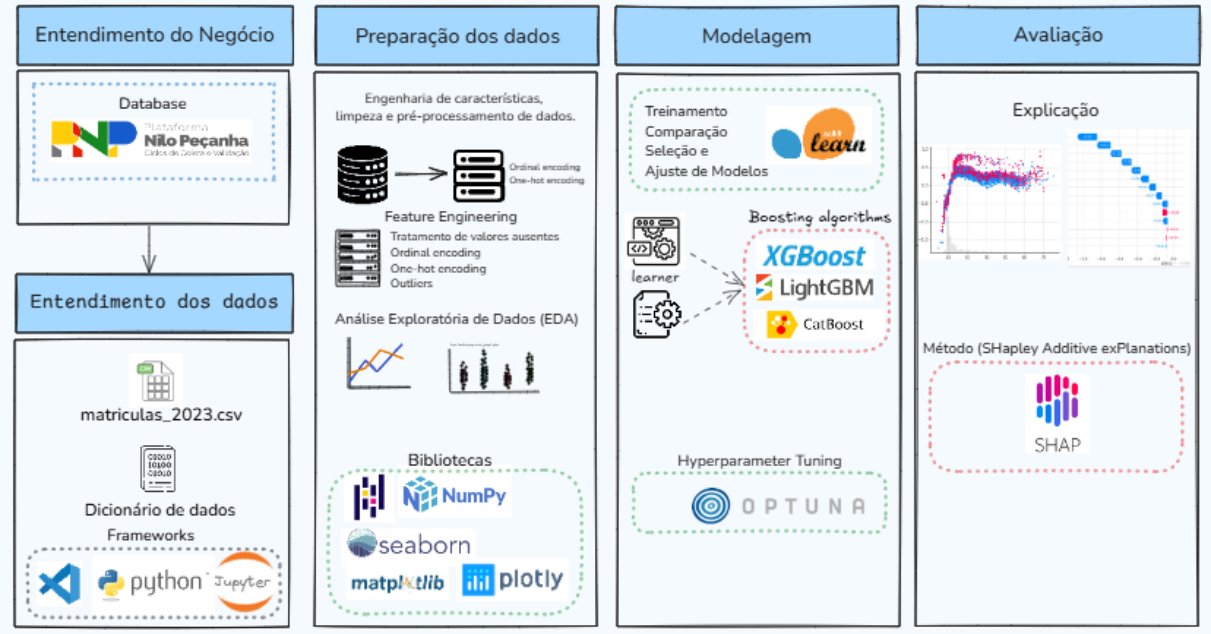
3 MATERIAIS E MÉTODOS

O fluxo metodológico deste estudo foi organizado em cinco fases. Iniciou-se com a compreensão do contexto de negócio, seguida pelo entendimento dos dados. Na etapa seguinte, procedeu-se à preparação dos dados, envolvendo engenharia de atributos e uma análise exploratória. Com a base de dados tratada, procedeu-se à modelagem e comparação dos modelos e, por fim, a avaliação e a explicação com base no algoritmo de melhor desempenho.

O desenvolvimento da solução de aprendizado de máquina foi conduzido conforme *framework* CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (IBM, 2023). O CRISP-DM estabelece um processo interativo e adaptativo que segmenta um projeto em etapas bem-delimitadas (entendimento de negócio, entendimento de dados, preparação dos dados, modelagem, avaliação e implementação), assegurando uma abordagem robusta e reproduzível, independente das ferramentas computacionais empregadas (MARTINEZ, 2019; WITTEN et al., 2023). Como uma metodologia, ela contempla a descrição das fases típicas de um projeto, as tarefas envolvidas em cada fase e a explicação dos relacionamentos entre essas tarefas.

A figura 1 apresenta uma visão geral do fluxo de execução dessas etapas, destacando as atividades realizadas e os artefatos utilizados em cada fase do processo.

Figura 1 - Etapas de elaboração do projeto baseado no CRISP-DM



Fonte: elaborado pelo autor, 2025.

Neste artigo, o foco é a quinta etapa da metodologia que consiste na avaliação do modelo que apresentou melhor desempenho preditivo, o CatBoost, direcionado a analisar a contribuição das variáveis determinantes para a predição da evasão escolar em cursos técnicos na RFEPCT com uso do método SHAP. A etapa de implementação não foi contemplada neste artigo.

3.1 Coleta dos dados

O objetivo foi a coleta, exploração e análise inicial dos dados, com o objetivo de obter uma compreensão abrangente sobre o fenômeno da evasão, identificando padrões, inconsistências e características relevantes.

O conjunto de dados utilizado nesta pesquisa tem origem na Plataforma Nilo Peçanha (PNP), do Ministério da Educação, disponibilizado no portal do MEC (BRASIL, 2024). O acesso aos microdados específicos foi realizado por meio do Portal de Dados Abertos do Governo Federal (BRASIL, 2023).

A base de dados abrange as matrículas de alunos da RFEPCT do ano de 2023, totalizando 145.831 registros. Esses registros incluem informações demográficas, institucionais e acadêmicas associadas aos cursos técnicos. Cabe ressaltar que os dados já se apresentam classificados com base na variável que define a categoria da situação de matrícula.

A análise restringiu-se às matrículas em cursos de nível técnico. Essa delimitação tem fundamento legal na Lei nº 11.892/2008, que institui a Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT). Conforme estabelecido pela legislação, no mínimo 50% das vagas oferecidas pelas instituições desta rede devem ser destinadas à educação profissional técnica de nível médio (BRASIL, 2008). Ademais, a carga horária mínima de 800 horas, obrigatória para esses cursos, possibilita um acompanhamento anual da trajetória acadêmica dos discentes.

Para complementar a análise, incorporou-se a base de dados oficial do Instituto Brasileiro de Geografia e Estatística (IBGE) sobre regiões metropolitanas (IBGE, 2023). A integração desses dados teve como objetivo identificar quais unidades de ensino da RFEPCT estão situadas nessas áreas. O propósito foi investigar a possível influência da localização geográfica na probabilidade de evasão escolar.

3.1.1 Descrição dos dados

A tabela a seguir apresenta a descrição das variáveis contidas no conjunto de dados utilizado na análise. O objetivo é oferecer uma visão clara e estruturada dos atributos que compõem a base, facilitando a compreensão e a reprodutibilidade do estudo. Caracterizar a diversidade e o nível de granularidade das informações disponíveis é essencial para compreender a complexidade da base e para orientar decisões relacionadas à codificação de variáveis categóricas, à identificação de possíveis inconsistências e à preparação dos dados para a modelagem preditiva.

Tabela 1 - Descrição das variáveis do conjunto de dados

Variável	**Descrição**	**Tipo**
Categoria situação (Alvo)	Situação do aluno em relação à matrícula: *em curso*, *concluído* ou *evadido*.	Categórica nominal
Cor/raça	Cor ou raça autodeclarada do aluno *Branca*, *Preta*, *Parda*, *Amarela* ou *Indígena ou Não informado*.	Categórica nominal
Idade	Idade do aluno no momento da matrícula ou da coleta dos dados, em anos. Varia entre 4 e 84 anos.	Numérica discreta
Sexo	Gênero do aluno: *Masculino*, *Feminino* ou *Não informado*.	Categórica nominal
Renda familiar	Faixas salariais ordenadas: $0 < RFP \leq 0,5$; $0,5 < RFP \leq 1,0$; $1,0 < RFP \leq 1,5$; $1,5 < RFP \leq 2,5$; $2,5 < RFP \leq 3,5$; $RFP > 3,5$; *Não declarada*.	Categórica ordinal
Modalidade de ensino	Modalidade em que o aluno estuda: *presencial* ou *educação a distância*.	Categórica nominal

Variável	Descrição	Tipo
Tipo de oferta	Forma de ingresso no curso técnico: *Subsequente, Proeja - Subsequente, Concomitante, Proeja – Concomitante, Integrado ou Proeja - Integrado*.	Categórica nominal
Turno	Turno de realização das aulas: *Matutino*, *Vespertino*, *Noturno* ou *Integral*.	Categórica nominal
Nome de curso	Nome do curso técnico no qual o aluno está matriculado. Registro de 140 diferentes cursos técnicos ofertados.	Categórica nominal
Eixo tecnológico	Área de conhecimento ou eixo tecnológico ao qual o curso pertence. Agrupamento dos cursos em 13 eixos tecnológicos, conforme o Catálogo Nacional de Cursos Técnicos (CNCT).	Categórica nominal
Carga horária mínima	Carga horária mínima exigida para conclusão do curso, expressa em horas. Três categorias de carga horária mínima exigida: 800, 1.000 e 1.200 horas, conforme o CNCT.	Numérica discreta
UF (Unidade da Federação)	Estado da instituição de ensino (ex.: *SP*, *RJ*, *BA*). Representação das 27 Unidades Federativas do Brasil.	Categórica nominal
Município	Município onde está localizada a unidade de ensino. Registro de matrículas em 550 municípios distintos.	Categórica nominal
Região	Corresponde às cinco grandes regiões geográficas do Brasil: *Norte*, *Nordeste*, *Centro-Oeste*, *Sudeste* e *Sul*.	Categórica nominal
Instituição	Nome da instituição da Rede Federal à qual a unidade pertence. Conjunto de 62 instituições da RFEPCT.	Categórica nominal
Unidade de ensino	Nome da unidade de ensino onde o curso técnico é ofertado. Matrículas em 618 unidades de ensino distribuídas em todo o território nacional.	Categórica nominal
Região metropolitana da UE (Unidade de Ensino)	Indica se a unidade de ensino está localizada em região metropolitana, com os valores *Sim* ou *Não*.	Categórica nominal

Fonte: elaborado pelo autor, 2025.

3.2 Pré-processamento, limpeza, transformação, engenharia de atributos e análise exploratória dos dados.

A análise exploratória do conjunto de dados foi conduzida para compreender a estrutura de dados e identificar possíveis inconsistências antes da fase de modelagem. Esta etapa foi fundamental para garantir a qualidade dos dados, orientar as técnicas de pré-processamento e definir estratégias adequadas para a construção de modelos preditivos. A exploração abrangeu múltiplas dimensões, incluindo a análise de valores nulo, a contagem de valores únicos por variável, a visualização da distribuição de registros por atributo, a identificação de *outliers*, a análise da distribuição dos dados, a caracterização da variável alvo e a investigação da relação entre as variáveis preditoras e a variável alvo (HAN; KAMBER; PEI, 2022; PYLE, 1999).

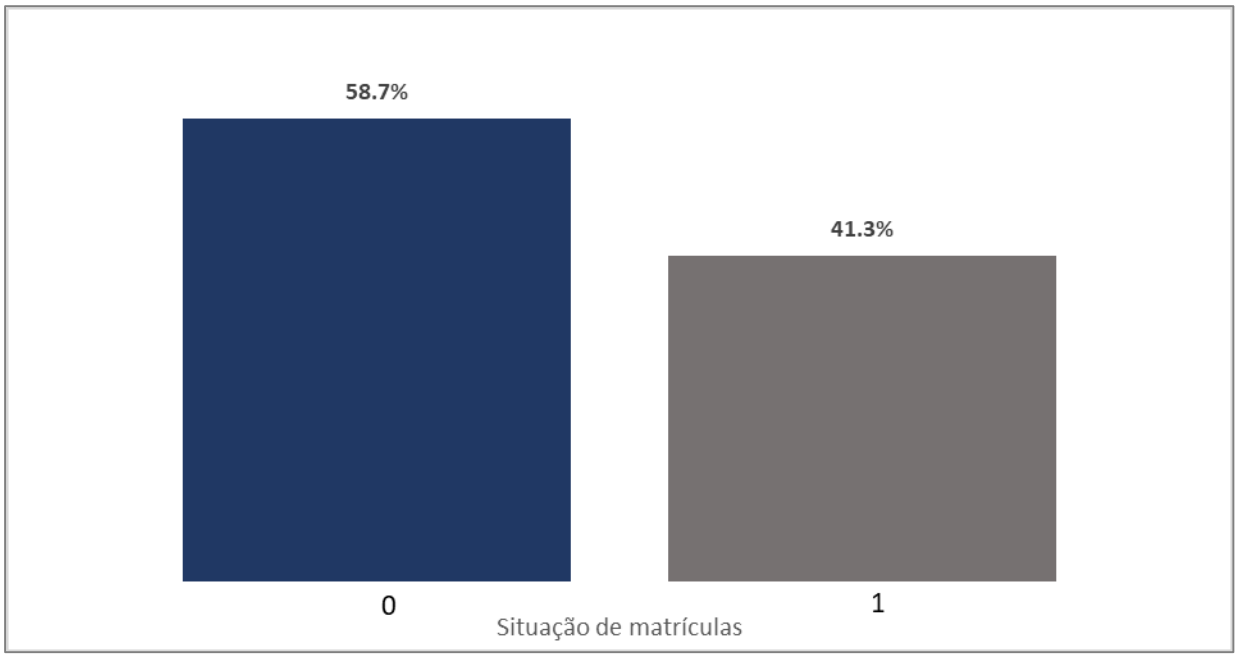
Para enriquecer a perspectiva geográfica da análise, integrou-se ao conjunto principal uma base externa do IBGE sobre regiões metropolitanas. A variável alvo CATEGORIA_SITUACAO foi transformada em um formato binário, onde evadido foi codificado como 1 e não evadido como 0, permitindo assim sua aplicação em algoritmos de classificação supervisionada (PANG et al., 2021). No pré-processamento, as variáveis categóricas foram codificadas utilizando-se as técnicas *OneHotEncoder* ou *OrdinalEncoder*, conforme sua natureza nominal ou ordinal, respectivamente. A divisão do conjunto de dados em treino e teste foi realizada por meio da técnica *Stratified KFold* para preservar a proporção da variável alvo e assim evitar viés na avaliação dos modelos (GAUDREAULT; BRANCO; GAMA, 2021).

Uma etapa relevante consistiu na substituição de valores na coluna "Cor/Raça" para assegurar a consistência categórica, fundamental para análises estatísticas confiáveis (BRASIL, 2023; HAN; KAMBER; PEI, 2022). Assim como, visando à eficiência computacional, as colunas "município" e "unidade_de_ensino" foram excluídas devido à sua alta cardinalidade, que demandaria um custo elevado para codificação e armazenamento (DOMINGUES et al., 2022; HAN; KAMBER; PEI, 2022).

Outra etapa importante foi o tratamento dado a variável ordinal "Renda Familiar", que se aplicou o *OrdinalEncoder*, preservando a relação de ordenamento inerente às suas faixas de renda familiar dos estudantes (HAN; KAMBER; PEI, 2022; PEDREGOSA et al., 2011). Já as variáveis categóricas nominais, "categoria_situacao", "cor_raca", "sexo", "modalidade_de_ensino", "tipo_de_oferta", "turno", "eixo_tecnologico", "nome_do_curso", "instituição", "uf", "regiao" e "região_metropolitana_ue" foram transformadas em numéricas por meio do *OneHotEncoder* da biblioteca Scikit-learn. Por fim, para mitigar a multicolinearidade, fenômeno que pode comprometer a estabilidade e interpretabilidade de modelos sensíveis a correlações, adotou-se a estratégia de remoção de categoria de referência n-1, excluindo-se uma das colunas *dummy* resultantes da codificação (JAMES et al., 2021).

Finalizada a etapa de engenharia de atributos, o conjunto de dados final apresentou um leve desbalanceamento entre as classes da variável alvo, conforme ilustrado na figura 2.

Figura 2 - Percentual de evadidos registrado na base de dados após tratamento dos dados



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A análise da distribuição da variável alvo evidencia uma taxa de evasão de 41,3% no em cursos técnicos da RFEPCT, em contraste com 58,7% de estudantes que permaneceram em situação de matrícula ativa ou concluíram seus cursos.

3.3 Análise com SHAP

Com o objetivo de conferir interpretabilidade aos resultados do modelo de melhor desempenho preditivo, Catboost, utilizou-se o método SHAP (*SHapley Additive exPlanations*) em sua versão 0.49.1 para a linguagem Python. A técnica permitiu explicar as saídas do modelo de aprendizado de máquina selecionado a partir das contribuições atribuídas a cada variável. Com isso, foi possível compreender o impacto de cada atributo individualmente na decisão de evadir ou não. Essa abordagem entrega transparência e confiabilidade às análises, aspectos fundamentais para utilização de modelos preditivos em contextos educacionais (LÓPEZ, 2021).

Para a apresentação dos resultados utilizou-se visualizações, como *summary plot, heatmap, beeswarm, waterfall, force plot e dependence plot*, o que permitiu uma análise detalhada dos atributos que impactam a evasão dos estudantes.

4 RESULTADOS E DISCUSSÃO

4.1 Modelagem dos modelos de aprendizado de máquina

O processo de modelagem utilizou algoritmos de aprendizado de máquina com o objetivo de prever a evasão escolar a partir de um conjunto de variáveis explicativas. Inicialmente, os dados foram divididos de forma estratificada em dois subconjuntos: 80% para o conjunto de treinamento e 20% para o conjunto de teste. Essa divisão assegurou que os dados de teste não fossem utilizados durante a fase de ajuste dos modelos (ZHANG et al., 2021; GERON, 2019; KUHN; JOHNSON, 2013).

O controle da aleatoriedade foi realizado com o parâmetro *random_state* da função *train_test_split()*, de modo a garantir que a amostragem fosse reproduzível em execuções futuras do código. Esse procedimento assegurou a consistência dos experimentos e viabilizou a comparação objetiva entre diferentes configurações de modelos (HANA; LOFSTEAD, 2022).

Durante o pré-processamento, aplicou-se a função *fit_transform()* apenas sobre os dados de treino, enquanto os dados de teste foram submetidos unicamente ao método *transform()*.

Para manter a proporcionalidade da variável alvo entre os subconjuntos, a divisão foi realizada com o parâmetro *stratify=y* na função *train_test_split()*, o que é fundamental em conjuntos de dados desbalanceados, como é o caso deste estudo, em que há mais alunos não evadidos (em curso ou concluído) (BISHOP, 2021; KUHN; JOHNSON, 2019).

Na avaliação e comparação dos modelos, foi adotada a técnica de validação cruzada estratificada (*K-Fold*), que assegurou a manutenção da distribuição da variável alvo em cada partição (KUHN; JOHNSON, 2019; ZHANG et al., 2021).

Como métrica principal de avaliação, utilizou-se o *Receiver Operating Characteristic – Area Under the Curve* (ROC-AUC), uma vez que a acurácia pode apresentar resultados enganosos em situações de desbalanceamento entre as classes. O ROC-AUC permitiu avaliar a capacidade do modelo em distinguir entre as classes, independentemente de sua distribuição HAN; KAMBER; PEI, 2022).

O objetivo dessa etapa foi identificar o algoritmo de melhor desempenho para as etapas subsequentes, que incluíram a seleção de atributos, o ajuste de hiperparâmetros e a avaliação final. Foram aplicadas técnicas de seleção de características e ajuste de hiperparâmetros com o objetivo de aprimorar o desempenho preditivo dos modelos e reduzir a complexidade computacional (HUTTER; HOOS; LEYTON-BROWN, 2019).

Para estimar a evasão escolar utilizou-se variáveis sociodemográficas, econômicas e acadêmicas. Para isso, foram treinados e comparados diversos algoritmos de aprendizado de máquina, incluindo: modelos lineares, como Regressão Logística, Linear SVC e Support Vector Machine (SVM); modelos em instância, como K-Nearest Neighbors (KNN); modelos baseados em árvore de decisão, como Decision Tree e Random Forest; e modelos de *gradient boosting*, como XGBoost, LightGBM e CatBoost.

Após a seleção do modelo com melhor desempenho, realizou-se a avaliação final utilizando o conjunto de teste, a fim de analisar o comportamento do modelo em um cenário mais próximo das condições reais de produção. Além disso, para interpretar os resultados e compreender as contribuições individuais das variáveis na predição, empregaram-se técnicas baseadas em SHAP (Shapley Additive Explanations), com intuito de gerar explicações tanto locais (por instância) quanto globais (sobre o modelo como um todo), promovendo maior transparência e interpretabilidade do modelo final (LUNDBERG; LEE, 2017).

4.2 Análise dos modelos de aprendizado de máquina

A avaliação do desempenho dos modelos ocorreu após o treinamento dos modelos de aprendizado de máquina. Foram testados nove algoritmos: Support Vector Machine (SVM), Random Forest, Linear SVC, K-Nearest Neighbors (KNN), Regressão Logística, Árvore de Decisão, XGBoost, LightGBM e CatBoost.

A avaliação dos modelos baseou-se no desempenho médio nas dobras da validação cruzada, seguido pelo teste final neste conjunto de dados, visando verificar sua capacidade preditiva fora da amostra de treinamento. Os resultados permitiram uma comparação objetiva entre os algoritmos, subsidiando a escolha do modelo mais adequado para a tarefa de realizar a predição (HASTIE: TIBSHIRANI: FRIEDMAN, 2009).

Inicialmente, foi calculada a pontuação de desempenho dos nove classificadores considerando as métricas de acurácia, precisão, recall, F1-score e ROC-AUC após a aplicação do ajuste de hiperparâmetros.

Tabela 2 - Desempenho dos modelos de aprendizado de máquina aplicando hiperparâmetros

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**ROC-AUC**
CatBoost	0,72	0,65	0,69	0,67	0,78
XGBoost	0,71	0,63	0,69	0,66	0,71
LightGBM	0,70	0,63	0,69	0,66	0,70
Linear SVC	0,67	0,60	0,62	0,61	0,66
SVM	0,68	0,64	0,52	0,57	0,66

Random Forest	0,67	0,61	0,59	0,60	0,66
KNN	0,66	0,60	0,55	0,57	0,65
Logistic Regression	0,66	0,59	0,63	0,61	0,66
Decision Tree	0,65	0,57	0,60	0,59	0,54

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela 2 apresenta uma comparação do desempenho dos algoritmos após a aplicação de técnicas de otimização de hiperparâmetros, destacando os modelos CatBoost, XGBoost e LightGBM como os de melhor desempenho, especialmente em relação à métrica ROC-AUC. O CatBoost destacou-se, alcançando o valor mais elevado (0,78), o que indica sua superior capacidade preditiva. Esses resultados sugerem que os ajustes de hiperparâmetros teve papel significativo no aperfeiçoamento do desempenho dos modelos, especialmente em cenários de desbalanceamento de classes (BENTÉJAC; CSÖRGŐ; MARTÍNEZ-MUÑOZ, 2021; HAN; KAMBER; PEI, 2022; KUHN; JOHNSON, 2019; RASCHKA; MIRJALILI, 2022).

Realizou-se também a comparação dos principais algoritmos de aprendizado de máquina considerando as métricas de avaliação após a aplicação da técnica de balanceamento de classes SMOTE (*Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002). Observou-se que, mesmo após o balanceamento, os modelos CatBoost, XGBoost e LightGBM continuaram a apresentar o melhor desempenho. O CatBoost destacou-se, alcançando ROC-AUC de 0,78 e acurácia de 0,72. Já os modelos XGBoost e LightGBM apresentaram valores semelhantes em acurácia (0,71 e 0,70, respectivamente), precisão (0,63) e recall (0,69), demonstrando consistência em múltiplas métricas e estabilidade do desempenho após o balanceamento das classes.

Os resultados demonstraram que os algoritmos baseados em *boosting* (CatBoost, XGBoost e LightGBM) apresentaram desempenho superior mesmo após o balanceamento da base de dados por meio do SMOTE. Entretanto, em problemas como a evasão escolar, que requer atenção na escolha das métricas de avaliação mais adequadas, a acurácia não deve ser utilizada como principal critério de avaliação, visto que pode ocultar erros relevantes, sobretudo em cenários com classes desbalanceadas (SOKOLOVA; LAPALME, 2009).

Dessa forma, as métricas recall e F1-score mostraram-se mais adequadas para o problema em questão. O recall, ou sensibilidade, avaliou a capacidade do modelo em identificar corretamente os estudantes que efetivamente evadiram. Sua relevância decorreu do fato de que deixar de identificar um aluno com risco de evasão poderia gerar consequências educacionais e sociais significativas. O F1-score, que representa a média harmônica entre precisão e recall, foi utilizado para equilibrar a taxa de falsos positivos e a capacidade de detecção de alunos evadidos, oferecendo uma medida mais robusta de desempenho em cenários com classes

desbalanceadas (BAKARIWIE; ASAMOAH; DUWIEJUAH, 2025; JORDAN; MITTELMAN, 2021).

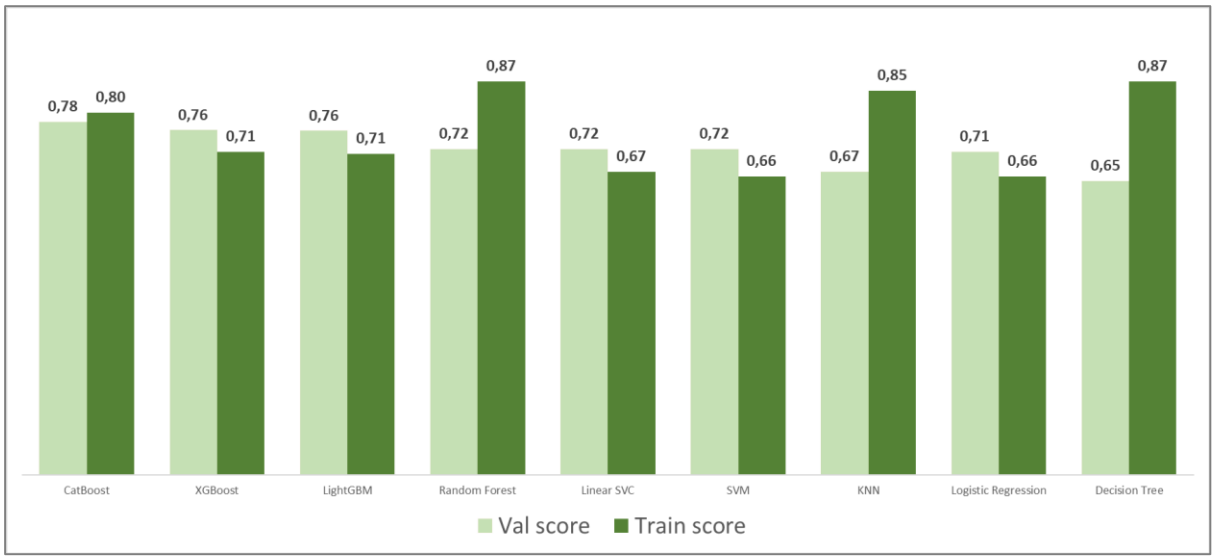
A análise demonstrou que o modelo CatBoost apresentou o melhor desempenho em recall (0,69) e F1-score (0,67) no cenário com ajuste de hiperparâmetros, indicando sua superioridade na identificação da classe minoritária, sem comprometer o equilíbrio geral do modelo. Já a métrica ROC-AUC mostrou-se uma ferramenta eficaz para avaliar a capacidade discriminatória global dos modelos, sendo que o CatBoost alcançou o valor mais elevado (0,78). Contudo, seu uso é recomendado apenas como suporte à análise, não devendo ser considerado como critério único para tomada de decisão (CHAWLA et al., 2002).

4.3 Desempenho dos modelos

A avaliação do desempenho dos modelos foi conduzida mediante a técnica de validação cruzada, considerando-se os valores médios dos *scores* de validação (*val score*) e de treinamento (*train score*). A validação cruzada fornece uma estimativa direta do erro de teste para um modelo, com base em uma única série de dados de treinamento. Para essa análise, o valor médio de validação representou o desempenho geral do modelo ao longo dos diferentes *folds*, constituindo-se em um indicador importante da capacidade de generalização (JAMES et al., 2021).

O escore de treinamento constituiu uma métrica fundamental para a avaliação de sobreajuste (*overfitting*) do modelo. Quando este valor se apresenta significativamente superior ao escore de validação, é possível inferir que o modelo está excessivamente adaptado aos dados de treinamento, resultando em uma baixa capacidade de generalização para novas amostras (RASCHKA, 2020).

Figura 3 - Desempenho do treinamento dos modelos com cross validation (com parâmetros)



Fonte: elaborado pelo autor com biblioteca python, 2025.

A figura 3 apresenta os *scores* médios de desempenho obtidos pelos modelos de aprendizado de máquina empregados, com a aplicação de validação cruzada e ajuste de hiperparâmetros. As métricas referem-se aos conjuntos de treinamento e validação, com o objetivo de avaliar a capacidade de generalização dos algoritmos na tarefa de predição da evasão escolar.

Observa-se que mais uma vez os modelos baseados em *boosting*, CatBoost, XGBoost e LightGBM apresentaram bom equilíbrio entre os *scores* de treino e validação, com destaque para o CatBoost, que alcançou valores de 0,80 (treino) e 0,78 (validação). De fato, algoritmos de *gradient boosting* demonstram robustez em problemas de classificação desbalanceada, devido à sua capacidade de iterativamente corrigir erros de previsão em observações minoritárias (HANCOCK; KHOSHGOFTAAR, 2020).

A análise comparativa entre diferentes algoritmos de classificação supervisionada apontou o modelo CatBoost como o algoritmo de melhor desempenho para a predição da evasão escolar. Os resultados obtidos ao longo do processo de avaliação, levou em consideração métricas relevantes para o contexto do problema, tais como precisão, *recall*, F1-score e ROC-AUC.

O CatBoost apresentou valores superiores nessas métricas em relação aos demais algoritmos testados, demonstrando maior capacidade de identificar corretamente os casos de evasão (classe positiva) e manter um equilíbrio adequado entre precisão e sensibilidade. Dessa forma, o desempenho consistente do CatBoost justifica sua adoção como o modelo preditivo final neste estudo.

Um diferencial relevante do algoritmo é sua capacidade nativa de lidar com variáveis categóricas, característica particularmente vantajosa em conjuntos de dados com grande presença desse tipo de atributo. O CatBoost preserva a semântica e as relações intrínsecas entre as categorias, convertendo-as em valores númericos durante o processo de *boosting* (PROKHORENKOVA et al., 2017). Esta capacidade eliminou a necessidade de técnicas convencionais de pré-processamento como *LabelEncoding* ou *OneHotEncoding*, simplificando o pipeline de preparação dos dados e minimizando riscos de perda informacional ou explosão dimensional (Catboost, 2025).

4.4 Desempenho do algoritmo CatBoostClassifier

A avaliação do modelo foi elaborada utilizando o conjunto de teste (*X_test*), que simula um cenário operacional real ao ser composto por dados não vistos durante o treinamento. Esta abordagem assegura uma estimativa mais confiável do desempenho preditivo do modelo em condições práticas de implantação.

O modelo foi configurado com os hiperparâmetros *auto_class_weights*='Balanced', para compensar o leve desbalanceamento observado na variável alvo, e *eval_metric*='AUC', priorizando assim a capacidade discriminativa do modelo em distinguir entre alunos com probabilidades de evasão. Considerando que o objetivo principal deste estudo é classificar os alunos com base no risco de evasão, e não realizar classificações binárias, a calibração das probabilidades geradas não se mostrou necessária. As probabilidades produzidas pelo modelo já fornecem uma ordenação válida e interpretável, sendo suficientes para fins de priorização (NICULESCU-MIZIL; CARUANA, 2005).

Tabela 3 - Métricas de avaliação do modelo CatBoost no conjunto de teste

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
0 (Não evadido)	0,77	0,74	0,75	17.102
1 (Evadido)	0,65	0,69	0,67	12.049
Acurácia geral			**0,72**	29.151
Média macro	0,71	0,71	0,71	
Média ponderada	0,72	0,72	0,72	

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A Tabela 3 sumariza o desempenho do modelo CatBoost no conjunto de teste. O modelo alcançou uma acurácia global de 72%, demonstrando bom desempenho na tarefa de

classificação. Entretanto, uma análise isolada por classe oferece informações mais substantivas, particularmente relevantes dado o desbalanceamento observado na variável alvo (evadido).

Para a classe majoritária (0 - Não evadido), o modelo registrou uma precisão de 77% e um recall de 74%, consolidados em um F1-score de 0.75. Estes valores indicam uma performance sólida e confiável na identificação de estudantes que permanecem nos cursos. Para a classe minoritária (1 - Evadido), a performance, ainda que inferior, manteve-se razoável, com precisão de 65%, recall de 69% e F1-score de 0.67%.

A robustez do modelo frente ao desbalanceamento é corroborada pelas métricas de agregação. As médias macro e ponderada para precisão, recall e F1-score foram de 0.71 e 0.72, respectivamente. Esta proximidade entre as médias atesta que o modelo não penaliza excessivamente a classe minoritária, mantendo um equilíbrio satisfatório em seu poder discriminativo.

Os resultados obtidos confirmam a adequação do modelo CatBoost para a tarefa de classificação binária de risco de evasão escolar, oferecendo um subsídio confiável para uma tomada de decisão orientada por dados. Conforme destaca Fawcett (2006), em problemas com distribuição assimétrica entre as classes, métricas como o F1-score e a área sob a curva ROC (ROC-AUC) são essenciais para complementar a análise da acurácia, fornecendo uma visão mais fidedigna do desempenho preditivo.

Considerando a natureza desbalanceada do conjunto de dados, com uma predominância de instâncias da classe "não evadido", a avaliação do modelo foi baseada em um conjunto abrangente de métricas, selecionadas para capturar diferentes dimensões do desempenho:

Acurácia: mede a proporção total de previsões corretas;

Precisão: mensura a confiabilidade das previsões positivas;

Recall: mede a capacidade do modelo de identificar corretamente os casos de evasão;

F1-score: representa a média harmônica entre precisão e recall, sendo particularmente informativo em cenários de desbalanceamento;

ROC-AUC (Área sob a Curva ROC): avalia a capacidade discriminativa do modelo em distinguir entre as duas classes;

Índice de Gini: derivado diretamente da ROC-AUC, quantifica o poder de discriminação do modelo;

PR-AUC (Área sob a Curva Precisão-Recall): métrica recomendada para conjuntos de dados desbalanceados, por focar no desempenho da classe de interesse minoritária; e

Brier Score: avalia a calibração e a precisão das probabilidades previstas;

Tabela 4 - Métricas de desempenho global do modelo CatBoost

Métrica	**Valor**
Acurácia	0,7183
Precisão	0,7213
Recall	0,6893
F1-score	0,6692
ROC-AUC	0,7785
Índice de Gini	0,5569
PR-AUC	0,7139
Brier Score	0,1908

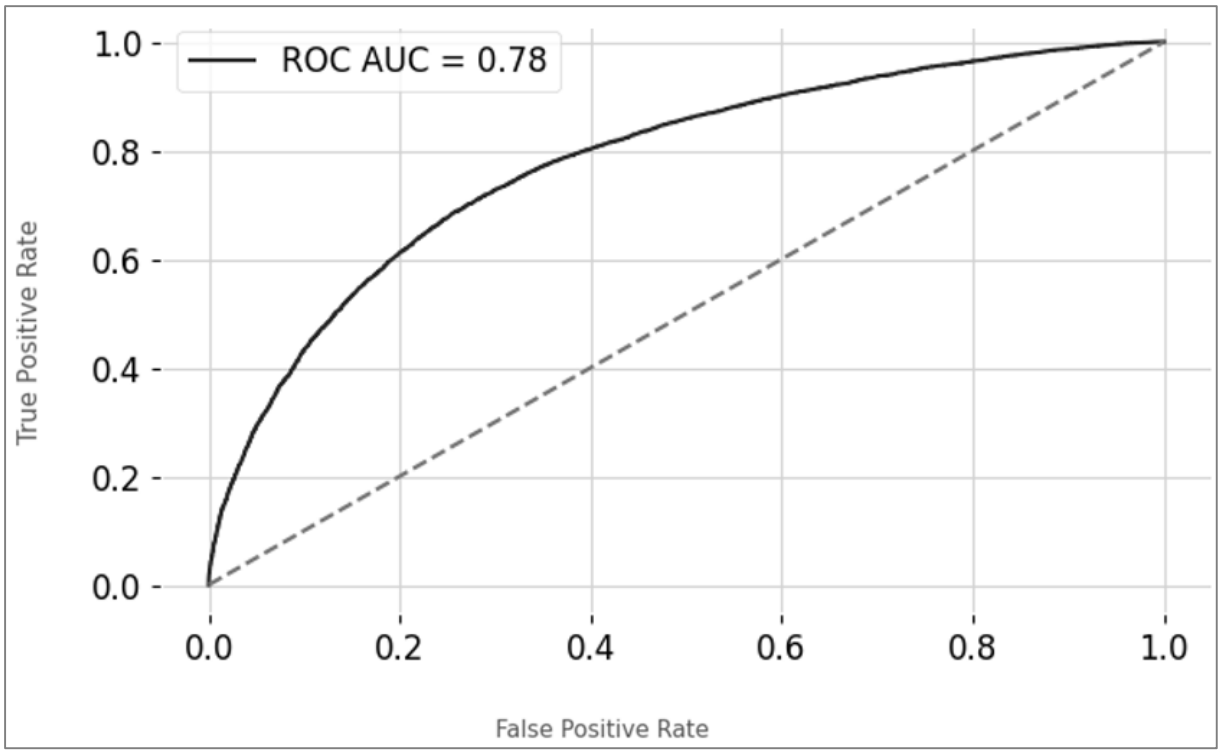
Fonte: elaborado pelo autor com biblioteca Python, 2025.

A tabela 4 consolida as principais métricas de avaliação do modelo CatBoost ajustado, demonstrando sua capacidade preditiva para identificação de risco de evasão em cursos técnicos. O modelo alcançou uma acurácia de 72%, desempenho considerado adequado para problemas de classificação com dados reais e moderadamente desbalanceados.

A análise das métricas específicas revela um equilíbrio satisfatório entre precisão (0,72) e recall (0,69), indicando que o modelo mantém boa capacidade de identificar casos positivos (evadidos) enquanto contém a ocorrência de falsos positivos. O cálculo da precisão foi realizado mediante o parâmetro *average*='weighted', abordagem que pondera o desempenho de cada classe pela sua representatividade no conjunto de dados, assegurando uma avaliação mais fidedigna face ao desbalanceamento observado entre as classes (PEDREGOSA et al., 2011).

O F1-score de 0,67, que sintetiza precisão e recall, corrobora a estabilidade do modelo diante da assimetria na distribuição das classes. A métrica ROC-AUC de 0,78 (Gini = 2 × ROC-AUC − 1), figura 4, evidencia uma discriminabilidade consistente entre evadidos e não evadidos (FAWCETT, 2006), sendo complementada pelo Índice de Gini de 0,56, metricamente derivado da ROC-AUC. A PR-AUC de 0,71 confirma o bom desempenho na classe minoritária, enquanto o Brier Score de 0,19 atesta a adequada calibração das probabilidades preditas.

Figura 4 - Curva ROC do modelo CatBoost para predição de evasão escolar



Fonte: elaborado pelo autor com biblioteca Python, 2025.

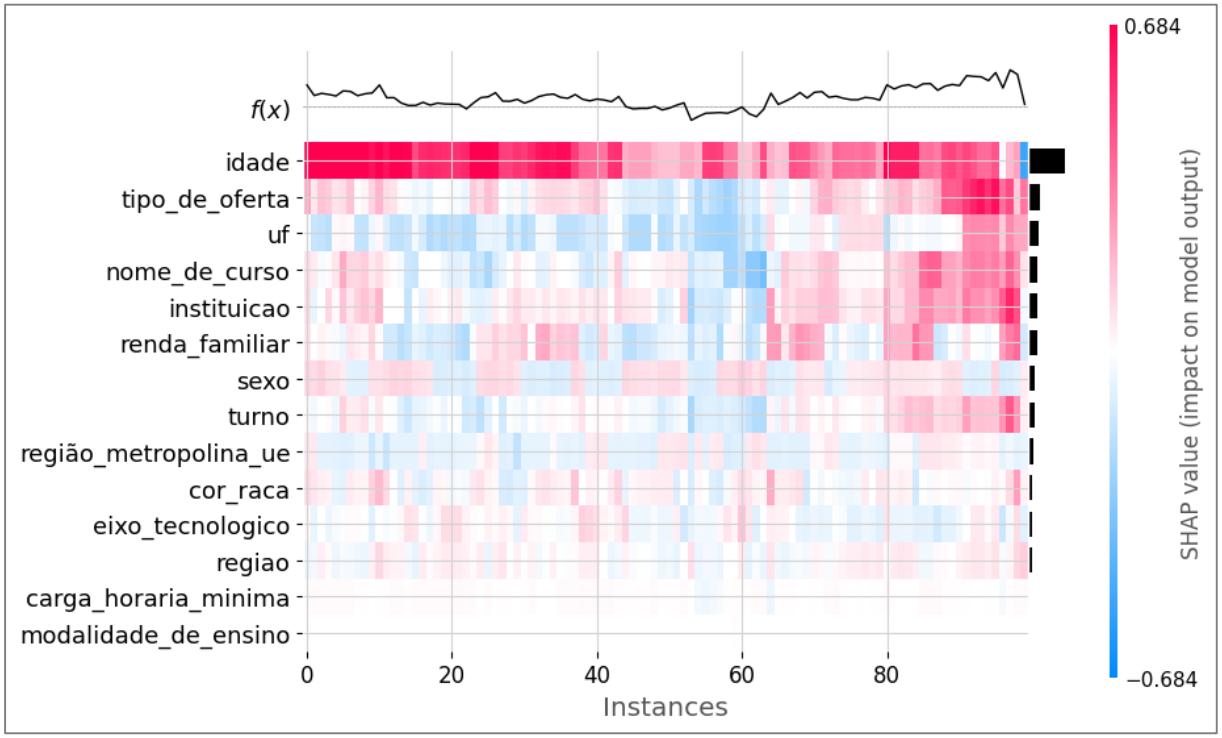
4.5 Interpretando os resultados do CatBoost com SHAP

A interpretação dos resultados do modelo CatBoost foi realizada com o auxílio do método SHAP (*SHapley Additive exPlanations*) (SHAPLEY, 1953). Esta ferramenta conecta a contribuição de cada variável à predição efetuada, atribuindo um valor de Shapley que quantifica a influência de cada atributo individualmente na decisão final do modelo.

A abordagem revela o impacto específico de cada característica em uma predição individual. Estudos recentes consolidam o SHAP como um método robusto para explicar a saída de qualquer modelo de aprendizado de máquina, promovendo a transparência e a interpretabilidade (LIU; ZHOU; LIU, 2025; ZHANG et al., 2023). No contexto desse estudo, essas qualidades são consideradas essenciais para a aplicação ética e confiável de modelos preditivos em cenários sensíveis, como o educacional (LUNDBERG; LEE, 2017; LUNDBERG et al., 2020).

O método foi utilizado como apoio na identificação e apresentação da importância das variáveis que influenciam a evasão escolar. O modelo permitiu destacar, de forma individualizada, os atributos que impactam na evasão ou na permanência dos estudantes na RFEPCT, contribuindo para uma análise mais interpretativa e direcionada do fenômeno.

Figura 5 - Mapa de calor dos valores SHAP para o modelo testado



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 5 apresenta o mapa de calor dos valores SHAP referentes aos cem primeiros registros do conjunto de teste, permitindo interpretar o impacto das quatorze variáveis de entrada nas previsões individuais geradas pelo modelo supervisionado CatBoost. Essa visualização integra informações sobre a importância global e o efeito direcional de cada variável na saída do modelo, proporcionando uma compreensão mais detalhada de seu comportamento preditivo (LUNDBERG; LEE, 2017).

A variável idade exerceu o maior impacto preditivo sobre o modelo, apresentando valores SHAP predominantemente positivos (em vermelho). Esse padrão indica que idades mais elevadas tendem a aumentar a probabilidade predita pelo modelo $f(x)$. A consistência e a intensidade da coloração reforçam a influência significativa dessa variável em múltiplas observações, evidenciando seu papel central na geração das previsões.

Entretanto, o fato da idade ter sido apontada pelo modelo SHAP como a variável de maior impacto preditivo não significa que ela seja a principal causa da evasão dos estudantes na RFEPCT. A identificação de correlações e sua causalidade exigiria análises adicionais com métodos específicos voltados à inferência casual.

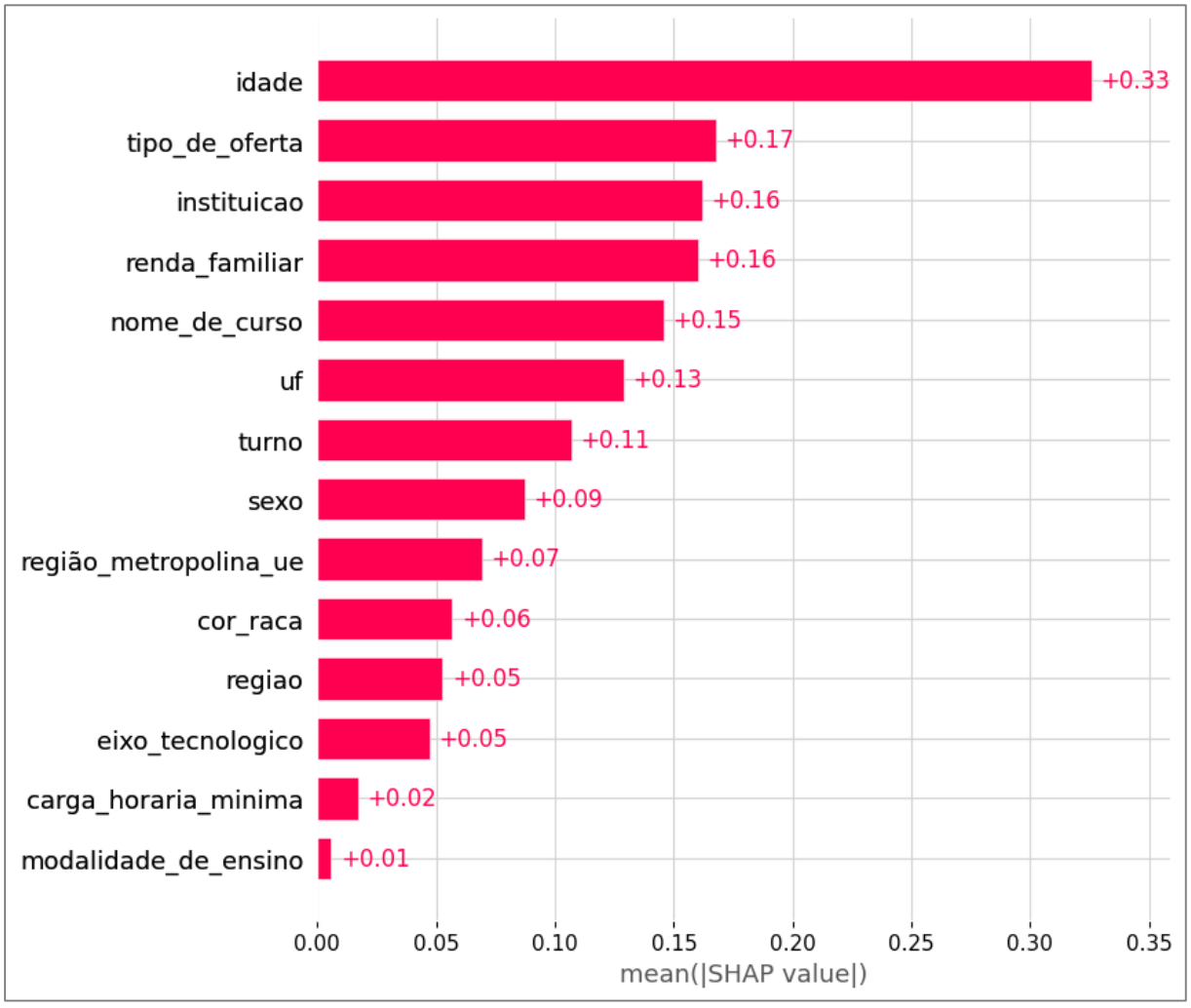
Variáveis como nome_de_curso, instituicao e renda_familiar apresentaram impacto moderado nas previsões, enquanto outras, como modalidade_de_ensino e carga_horaria_minima, mostraram-se menos influentes, com valores SHAP próximos de zero

e ausência de padrões consistentes de contribuição. Esse comportamento indica que o modelo atribui pouca importância a essas variáveis na explicação das previsões.

A linha superior do gráfico representa as predições individuais do modelo, possibilitando visualizar o efeito cumulativo das variáveis em cada instância analisada. A dispersão das colorações ao longo do eixo horizontal evidencia a heterogeneidade das contribuições entre as observações, característica compatível com dados educacionais ou socioeconômicos, onde interações complexas são esperadas e interações complexas entre atributos são frequentemente observadas.

A seguir, a importância das variáveis é analisada por meio de um gráfico de barras, que apresenta o valor médio absoluto dos valores SHAP para cada variável. Essa representação permite identificar o grau de contribuição de cada variável nas predições geradas pelo modelo, evidenciando quais fatores exercem maior influência no processo decisório do algoritmo.

Figura 6 - Importância média das variáveis com base nos valores SHAP absolutos



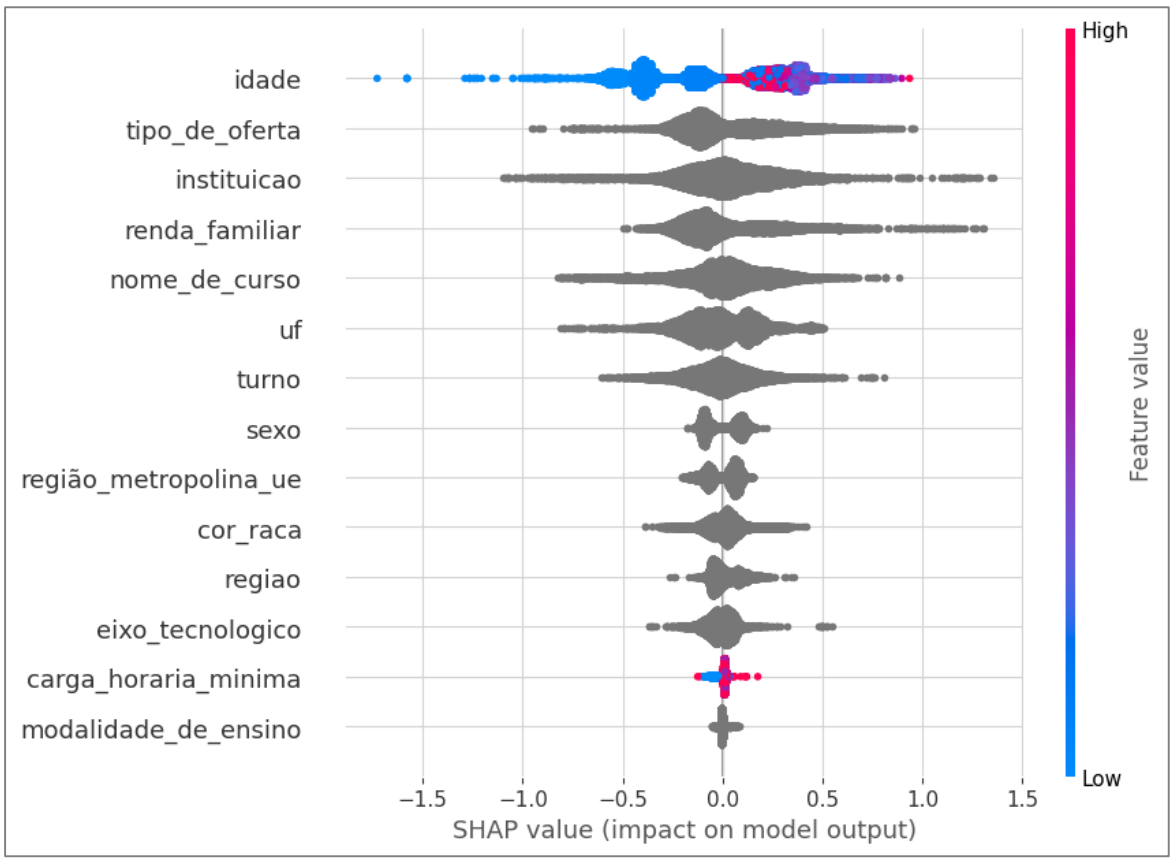
Fonte: elaborado pelo autor com biblioteca Python, 2025.

Conforme ilustrado pela magnitude das contribuições médias, a variável idade destaca-se como a mais influente, com um valor SHAP médio de aproximadamente 0,33. Esse resultado sugere uma forte associação entre a idade do aluno e a variável alvo, posicionando-a como o principal fator preditivo. Em seguida, as variáveis tipo_de_oferta (0,17), instituicao (0,16) e renda_familiar (0,16) exibem contribuições similares, indicando que características institucionais, organizacionais e socioeconômicas também desempenham um papel substancial e complementar nas previsões do modelo.

Outras variáveis com destaque relevante são nome_de_curso (0,15), uf (0,13) e turno (0,11), revelando que o modelo é sensível a aspectos geográficos e específicos da oferta educacional. As variáveis sexo (0,09) e regiao_metropolitana_ue (0,07) demonstram uma importância moderada, cujos impactos podem refletir padrões sistemáticos de desigualdade ou comportamento entre diferentes grupos populacionais.

As demais variáveis como cor_raca, regiao, entre outras, apresentaram menor impacto médio, situando-se abaixo de 0,06. Por fim, carga_horaria_minima e modalidade_de_ensino aparecem como as variáveis de menor influência no modelo, com valores próximos de 0,02 e 0,01, o que pode indicar baixo poder discriminativo ou correlação reduzida com a variável-alvo.

Figura 7 - Sumário com distribuição dos valores SHAP por variável



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 7 apresenta um *summary_plot* (SHAP *beeswarm*), que possibilita a visualização da distribuição do impacto de cada variável preditora sobre as saídas do modelo. Cada ponto no gráfico corresponde a uma instância do conjunto de dados, e sua posição no eixo horizontal reflete o valor SHAP, ou seja, a contribuição da variável para aumentar ou reduzir a predição do modelo para aquela instância específica (LÓPEZ, 2021; LUNDBERG; LEE, 2017). As cores dos pontos indicam a magnitude do valor da variável: valores mais baixos estão em azul e valores mais altos em vermelho.

A variável idade exibe a maior variação de impacto, com valores SHAP que variam significativamente entre negativos e positivos, demonstrando forte influência sobre a saída do modelo. Observa-se que valores baixos de idade (em azul) estão predominantemente associados a impactos negativos, enquanto valores altos (em vermelho) tendem a exercer impactos positivos, sugerindo uma possível correlação direta com a variável-alvo.

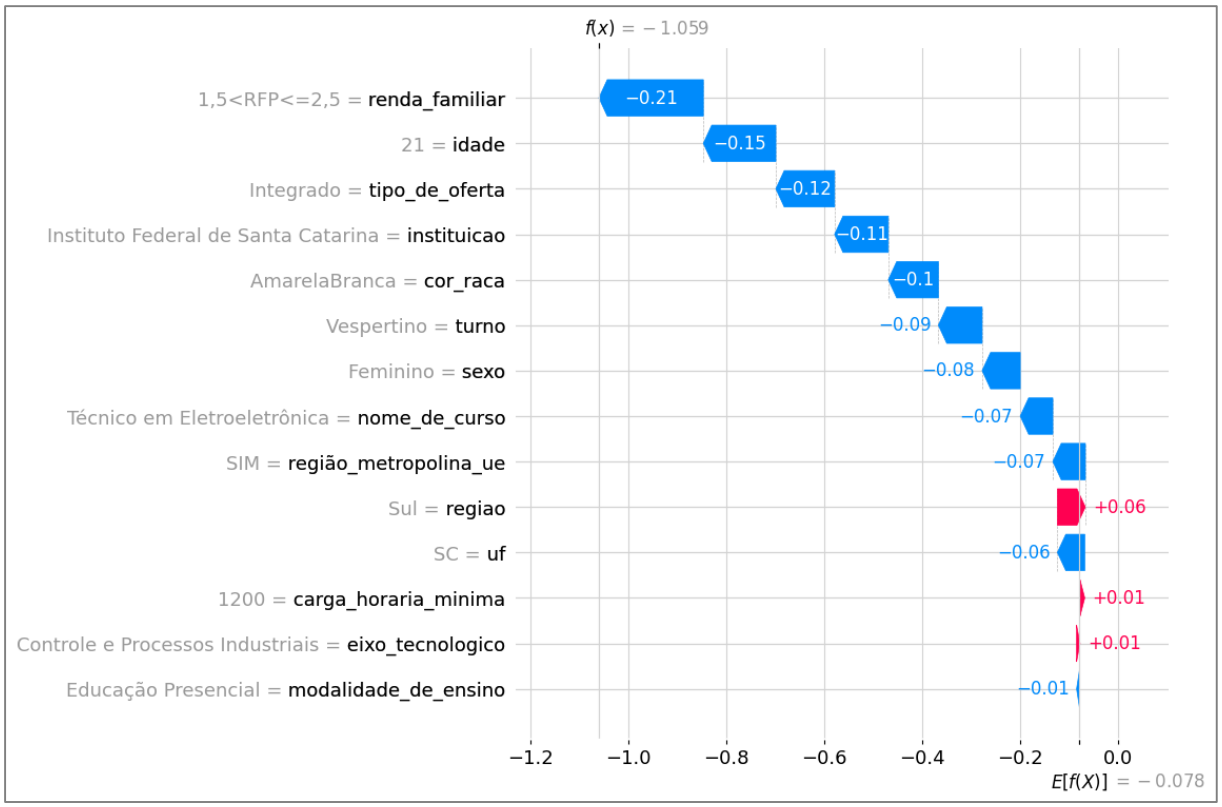
Variáveis, como tipo_de_oferta, instituicao, renda_familiar, nome_de_curso e uf apresentam dispersões moderadas dos valores SHAP, refletindo impactos relevantes, porém com menor amplitude em relação à variável idade. O padrão visual indica que diferentes níveis dessas características afetam o modelo de maneira diferenciada, embora com baixa intensamente (LUNDBERG et al., 2020).

Outras variáveis como carga_horaria_minima e modalidade_de_ensino revelam impactos concentrados próximos de zero, com pouca dispersão, sugerindo baixa relevância nas decisões do modelo. Esse tipo de informação é útil para redução de dimensionalidade e simplificação do modelo, especialmente em cenários regulados ou com foco em interpretabilidade (MOLNAR, 2022).

A visualização mostrou-se eficaz para identificar relações não lineares e interações complexas entre variáveis. A sobreposição de pontos com diferentes colorações em variáveis como cor_raca e região_metropolitana_ue pode sugerir efeitos de heterogeneidade e variações condicionais, que seriam difíceis de detectar por métodos convencionais.

A seguir, por meio do gráfico de cascata, será visualizado a contribuição individual de cada variável na predição da evasão individualmente.

Figura 8 - Influência individual dos atributos no resultado do modelo preditivo com SHAP [posição:10]



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 8 apresenta um gráfico SHAP, utilizado para interpretar localmente a predição realizada pelo modelo para uma matrícula específica do conjunto de dados de teste, no caso a matrícula posição [10], correspondente a um aluno que não evadiu. O formato do gráfico destaca de maneira clara quais variáveis influenciam positiva ou negativamente o valor final da predição. O valor final da função de decisão do modelo foi $f(x) = -1{,}059$, inferior à média global das predições $E[f(x)] = -0{,}078$, indicando uma forte inclinação do modelo para a permanência desse aluno no curso.

As variáveis com maior contribuição para essa classificação (valores SHAP negativos) foram: renda familiar entre 1,5 e 2,5 salários mínimos (−0,21), idade de 21 anos (−0,15), tipo de oferta integrada (−0,12) e instituição "Instituto Federal de Santa Catarina" (−0,11).

Outros fatores relevantes incluem: raça/cor branca ou amarela (−0,10), turno vespertino (−0,09), sexo feminino (−0,08) e curso técnico em Eletroeletrônica (−0,07), os quais também contribuíram para a previsão de não evasão.

Por outro lado, variáveis como região Sul (+0,06), carga horária mínima de 1200 horas (+0,01) e o eixo tecnológico (Controle e Processos Industriais) apresentaram contribuição positiva, ainda que pequena, no aumento do risco de evasão.

Essa análise local evidencia a utilidade do SHAP para explicar a lógica subjacente a uma predição individual, identificando os fatores protetivos e de risco que caracterizam perfis específicos de alunos (MOLNAR, 2022).

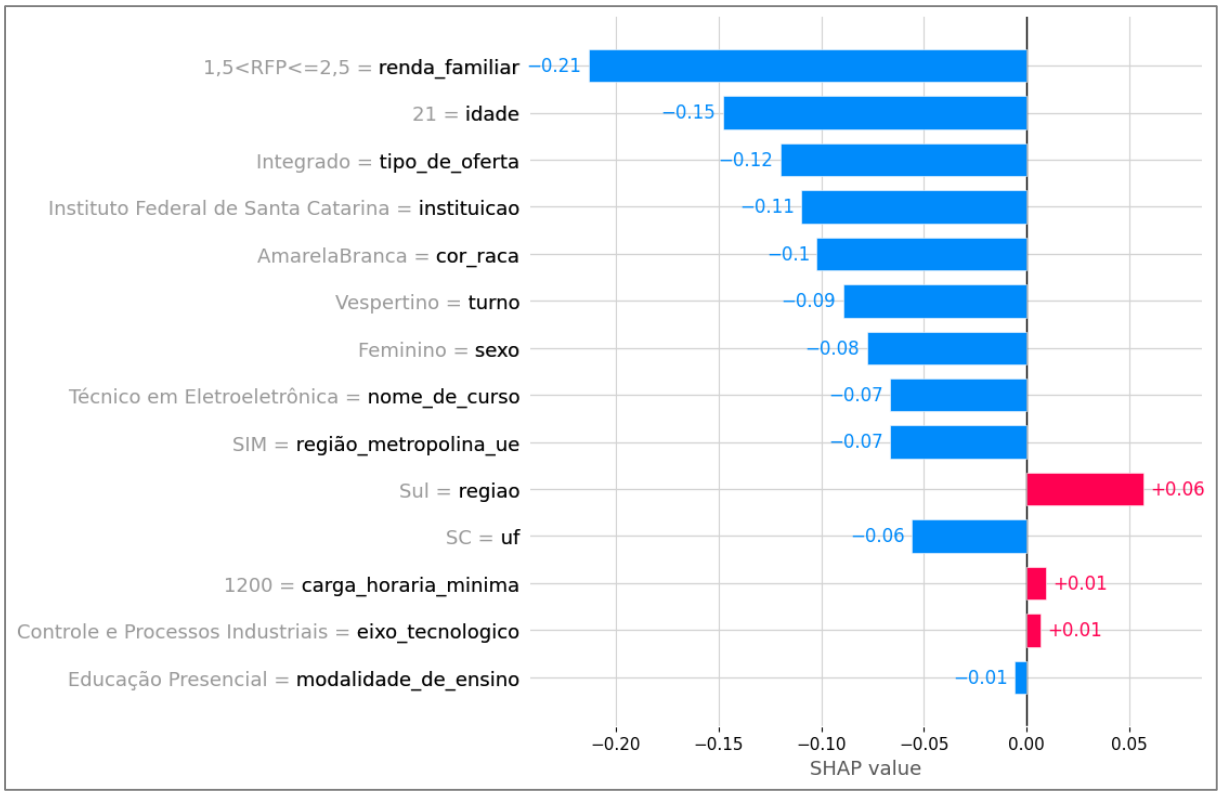
Figura 9 - Influência dos atributos no resultado do modelo preditivo com gráfico de força SHAP [posição:10]

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 9 resume visualmente a influência de cada variável individual sobre a predição do modelo para um aluno específico que não evadiu. Ele complementa a análise anterior ao mostrar de forma linear como cada característica do estudante contribui cumulativamente para o afastamento do valor-base (−0,07794) em direção à predição final (−1,06), sinalizando forte tendência de permanência.

As setas azuis apontam as variáveis que reduzem o risco de evasão como, a renda familiar intermediária, idade de 21 anos, tipo de oferta integrada, turno vespertino, sexo feminino e instituição. A única variável com leve influência contrária (em vermelho) é a região Sul, embora seu impacto seja mínimo frente ao conjunto de fatores que favoreceram a permanência.

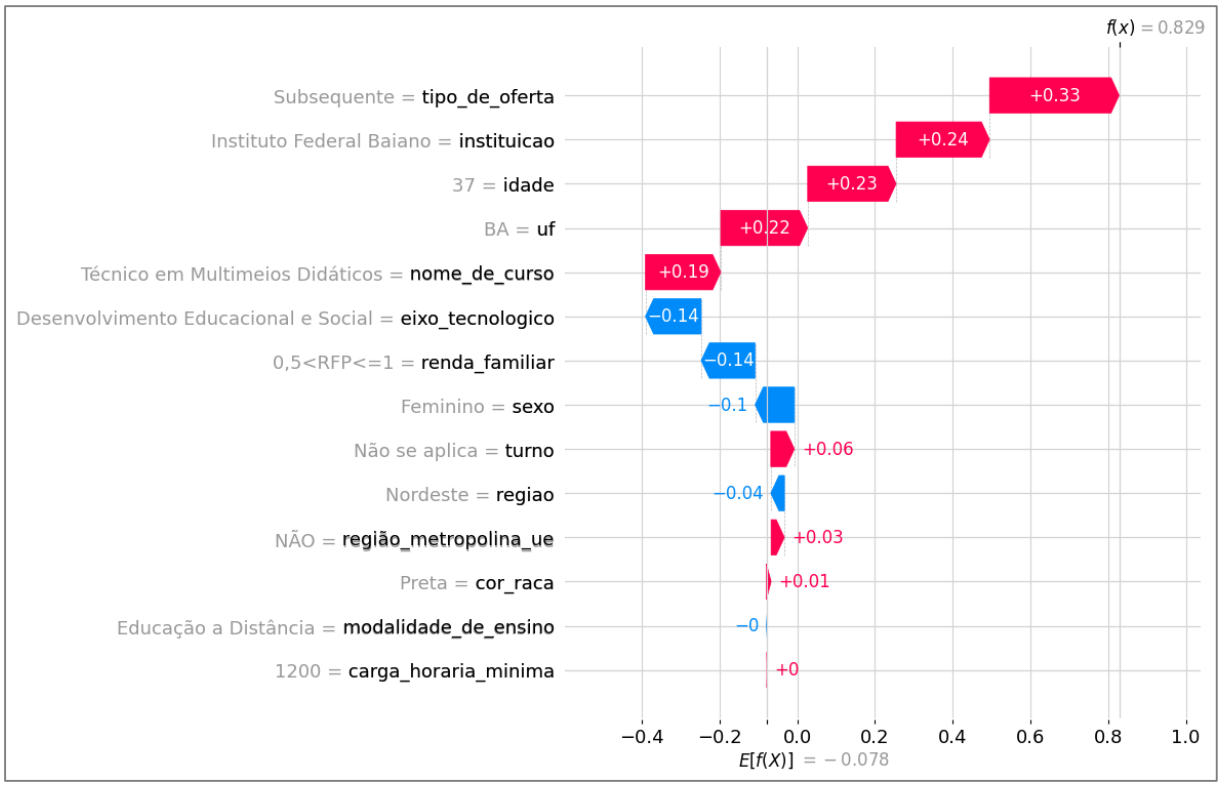
Figura 10 - Influência dos atributos no resultado do modelo preditivo com gráfico de barra SHAP [posição:10]



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A visualização pelo gráfico de barra SHAP contribui para a interpretação do modelo preditivo ao quantificar, em valores absolutos médios de Shapley, a importância relativa de cada variável na predição de evasão escolar. Diferentemente dos gráficos anteriores que indicam a direção da influência para um aluno específico, este resumo global permite hierarquizar os fatores segundo sua contribuição média na tomada de decisão do modelo.

Figura 11 - Influência individual dos atributos no resultado do modelo preditivo com SHAP [posição:15445]



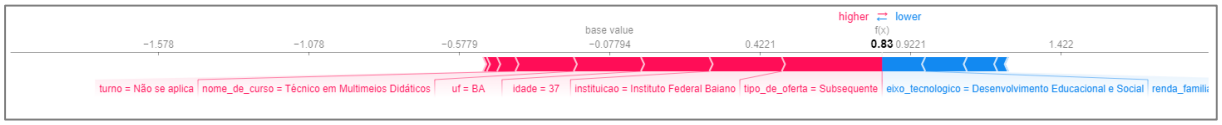
Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 11 apresenta um gráfico SHAP, utilizado para interpretar localmente a predição realizada pelo modelo para uma matrícula específica do conjunto de dados de teste, no caso a matrícula posição [15445], correspondente a um aluno que evadiu. O formato do gráfico destaca de forma clara quais variáveis influenciam positiva ou negativamente o valor final da predição. O valor final da função preditiva $f(x)$=0,829 representa uma forte inclinação do modelo para classificar o estudante como evadido, partindo de uma média esperada de $E[f(x)]$ =−0,078.

As variáveis que aumentaram substancialmente a probabilidade de evasão foram: oferta subsequente (+0,33), vínculo com o Instituto Federal Baiano (+0,24), idade de 37 anos (+0,23), residir no estado da Bahia-BA (+0,22) e o curso Técnico em Multimeios Didáticos (+0,19).

Outras variáveis como renda familiar entre 0,5 e 1 salário mínimo (−0,14), o eixo tecnológico Desenvolvimento Educacional e Social (−0,14), e o sexo feminino (−0,10) atuaram como fatores de proteção a evasão, mas com intensidade insuficiente para contrariar os demais fatores.

Figura 12 - Influência dos atributos no resultado do modelo preditivo com gráfico de força SHAP [posição:15445]



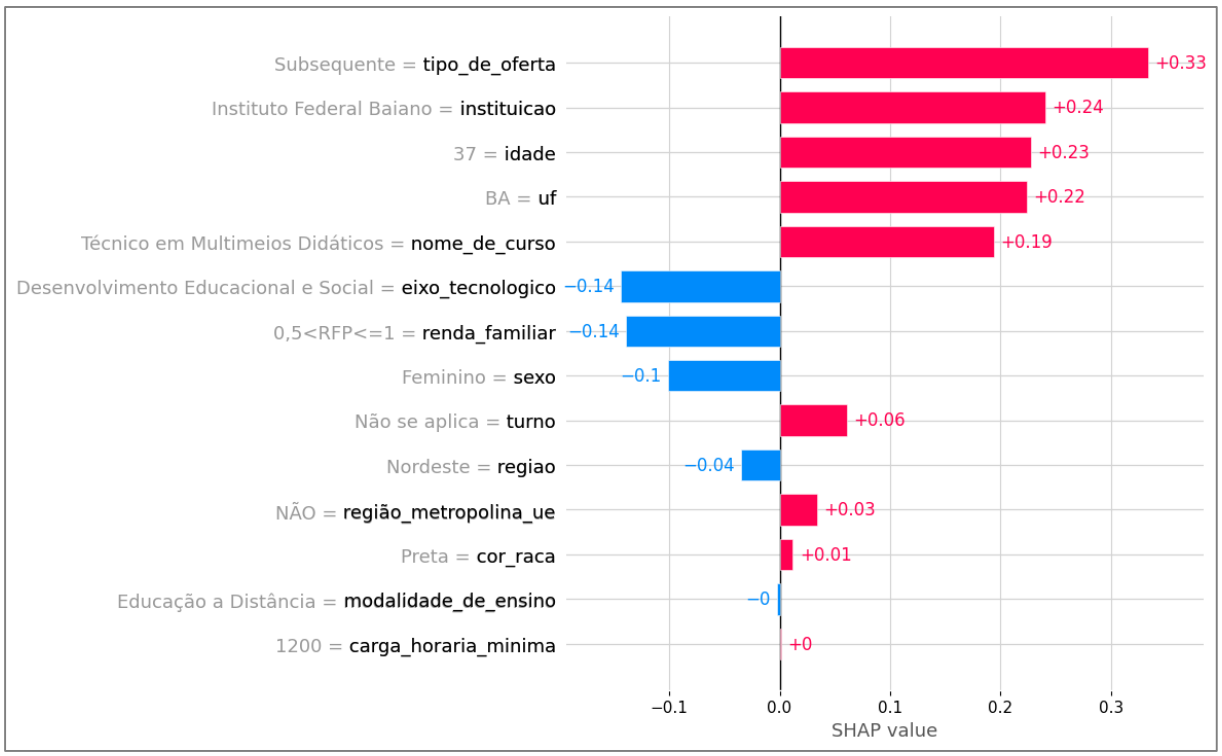
Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 12 resume visualmente a direção e intensidade da contribuição individual de cada variável para o valor final da predição de evasão. A estrutura horizontal, com o ponto de partida na média global da predição (base value= −0,07794), mostra como os atributos do estudante atuaram para deslocar o valor de $f(x)$ até +0,83, resultado em sua classificação como evadido.

As variáveis representadas em vermelho indicam contribuição positiva para a evasão, sendo as mais relevantes: turno não se aplica, curso Técnico em Multimeios Didáticos, UF = BA, idade = 37, instituição = Instituto Federal Baiano e tipo de oferta = Subsequente. Essas características exercem forças decisivas para deslocar a predição em direção à classificação de evasão.

Enquanto que as variáveis em azul representam fatores que protegem do risco de evasão, como eixo tecnológico Desenvolvimento Educacional e Social, renda familiar entre 0,5 e 1 salário mínimo e a região Nordeste. No entanto, sua influência foi insuficiente para reverter a tendência imposta pelas demais variáveis.

Figura 13 - Influência dos atributos no resultado do modelo preditivo com gráfico de barra SHAP [posição:15445]

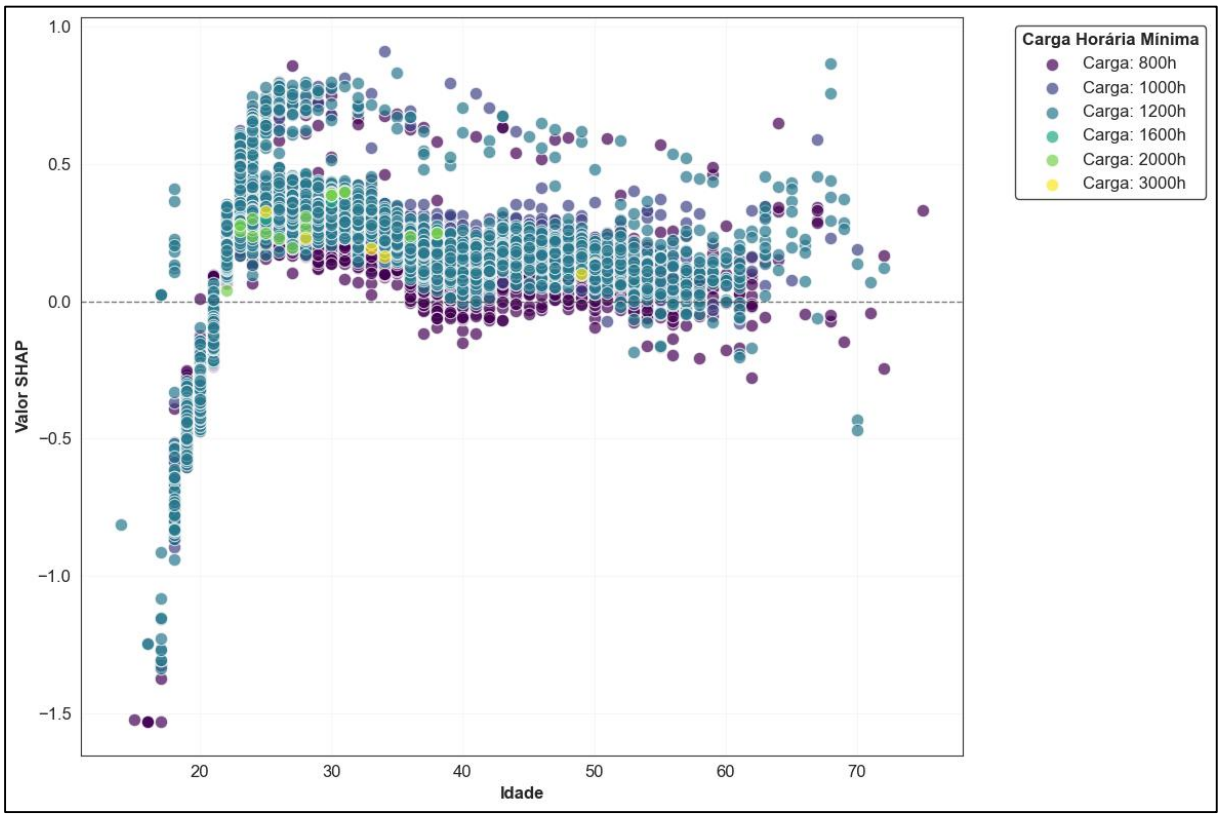


Fonte: elaborado pelo autorn com biblioteca Python, 2025.

A visualização apresentada pelo gráfico de barra SHAP demonstra o impacto médio de cada variável em todas as predições feitas pelo modelo. Este gráfico complementa as análises anteriores ao sintetizar, de forma comparativa, quais atributos são sistematicamente mais relevantes no comportamento geral do modelo de classificação.

O gráfico a seguir apresenta a relação entre a variável idade e carga_horaria_minima, permitindo analisar o efeito individual dessa variável nas previsões do modelo CatBoost. Cada ponto do gráfico representa uma observação do conjunto de dados, sendo que a posição no eixo vertical indica o impacto da idade na saída do modelo, enquanto a coloração dos pontos reflete a variação da variável carga_horaria_minima.

Figura 14 - Impactos da idade no modelo preditivo com gráfico de dispersão SHAP - separado por Carga horária mínima



Fonte: elaborado pelo autor com biblioteca Python, 2025.

A figura 14 apresenta um gráfico em que o eixo horizontal (x) representa a idade dos alunos, e o eixo vertical (y) exibe o valor SHAP, que indica a contribuição da variável idade para a predição do modelo em relação à evasão escolar. Valores SHAP positivos indicam que a idade contribui para aumentar a probabilidade de evasão, ao passo que valores negativos sugerem uma contribuição para reduzir essa probabilidade. Cada ponto do gráfico corresponde a uma observação individual, sendo as cores utilizadas para diferenciar as categorias de carga

horária mínima dos cursos. A linha horizontal tracejada em zero funciona como referência neutra, permitindo visualizar de forma clara em quais faixas etárias a idade exerce impacto positivo ou negativo sobre a decisão preditiva do modelo.

A interpretação dos valores SHAP apresentada neste estudo não deve ser entendida como causal, pois o foco principal da pesquisa foi a aplicação de técnicas de Aprendizado de Máquina, com ênfase na avaliação do desempenho preditivo e na capacidade interpretativa dos modelos a partir dos dados disponíveis da RFEPCT. Nesse contexto, os valores SHAP foram utilizados como métricas explicativas locais e globais, indicando a contribuição relativa das variáveis para as decisões do modelo, sem estabelecer relações de causa e efeito. As variáveis analisadas devem ser vistas como variáveis proxy, que registram padrões e relações presentes nos dados observacionais, refletindo estruturas latentes e condicionais do fenômeno da evasão, mas não seus mecanismos causais implícitos (LUNDBERG; LEE, 2017; MOLNAR, 2022). Assim, ressalta-se a importância de cautela ao extrapolar os resultados para inferências causais, que demandariam delineamentos metodológicos específicos (PEARL, 2009; HERNÁN; ROBINS, 2020).

5 CONCLUSÃO

Este estudo percorreu as etapas fundamentais de entendimento, preparação e modelagem de dados para aplicar aprendizado de máquina à predição de evasão na RFEPCT. Na fase de avaliação, a análise das variáveis determinantes revelou os principais fatores que influenciam a evasão em cursos técnicos. A execução criteriosa da metodologia, mesmo diante do grande volume de dados, demonstrou a viabilidade da proposta em um projeto real de IA, validando sua robustez e relevância prática.

Como principal contribuição, este estudo implementou o método SHAP para explicar as variáveis determinantes na predição de evasão em cursos técnicos da RFEPCT. A abordagem permitiu interpretar o comportamento do modelo CatBoost, que obteve o melhor desempenho preditivo, identificando os fatores sociodemográficos, características do curso e fatores individuais com maior influência na probabilidade de evasão dos estudantes.

As interpretações fornecidas pelo SHAP validaram que as previsões do modelo CatBoost estão alinhadas com o comportamento teórico da evasão em cursos técnicos. Ao quantificar a contribuição marginal de cada variável para o resultado final, o método revela as relações subjacentes capturadas pelo modelo de aprendizado de máquina. Dessa forma, o SHAP

oferece uma abordagem que simula a causalidade e explica o funcionamento interno do modelo, aumentando significativamente a transparência e a confiança do usuário nas predições de evasão.

Reconhece-se que a evasão em cursos técnicos é um fenômeno com várias causas, cuja prevenção e acompanhamento demandam ações integradas e políticas públicas orientadas à permanência escolar. Nesse contexto, a identificação transparente dos fatores de risco por meio de modelos preditivos confiáveis pode subsidiar intervenções mais assertivas, contribuindo para a redução desse problema social.

Esta solução tecnológica apresenta contribuições tanto teóricas quanto práticas para o enfrentamento da evasão escolar. No âmbito teórico, subsidia a formulação de políticas institucionais mais eficazes, promovendo uma gestão educacional baseada em dados e a identificação precoce de estudantes em situação de risco. Na prática, oferece uma metodologia que, a partir de variáveis categóricas e contínuas, explica por meio do método SHAP o risco individual de evasão de cada aluno, viabilizando assim intervenções personalizadas e tempestivas por parte da gestão educacional.

Cabe ressaltar que este estudo está condicionado às limitações inerentes à base de dados utilizada, a qual compreende exclusivamente matrículas finalizadas no ano de 2023 e apresenta um conjunto restrito de variáveis disponibilizadas para acompanhamento pela Rede Federal.

Como desdobramentos futuros desta pesquisa, recomenda-se a ampliação da robustez dos modelos preditivos e o aprofundamento da investigação sobre os fatores determinantes da evasão em cursos técnicos. Entre as possibilidades de continuidade do estudo, sugere-se: a utilização de séries históricas de matrículas para avaliar a estabilidade temporal das predições; a exploração de outras técnicas e algoritmos de aprendizado de máquina; a definição de novos parâmetros para otimização dos modelos testados; e a incorporação de variáveis adicionais para verificar possíveis ganhos no poder preditivo.

Apesar da utilização do SHAP como método explicativo neste trabalho, reconhece-se a existência de abordagens alternativas, como o LIME (Local Interpretable Model-agnostic Explanations). Sugere-se, portanto, a comparação das interpretações geradas por diferentes métodos, uma vez que a relevância das variáveis pode variar conforme a técnica explicativa adotada.

Por fim, os resultados aqui apresentados não apenas acrescentam novos conhecimentos à comunidade científica, como também demonstram a aplicabilidade prática de modelos

preditivos e explicáveis no enfrentamento da evasão escolar, oferecendo subsídios para políticas educacionais baseadas em evidências.

Declaração de IA Generativa e tecnologias assistidas por IA em processo de escrita

Durante a preparação deste trabalho, o autor utilizou o ChatGPT-4 e DeepSeek Latest Version para melhorar a legibilidade e a linguagem. Após o uso desta ferramenta, os autores revisaram e editaram o conteúdo conforme necessário e assumem total responsabilidade pelo conteúdo da publicação.

REFERÊNCIAS

BAKARIWIE, Amiru; ASAMOAH, Dominic; DUWIEJUAH, Abudu Ballu. Prevention of student attrition: a data-backed approach to school counselling using Delphi technique and multiple classification algorithms. Discover Education, v. 4, n. 1, p. 1–13, 31 jul. 2025. Disponível em: https://doi.org/10.1007/s44217-025-00494-7. Acesso em: 28 maio 2025.

BAKER, R. S.; SIEMENS, P. S. Educational data mining and learning analytics. In: SAWYER, R. K. (ed.). The Cambridge Handbook of the Learning Sciences. 2. ed. Cambridge: Cambridge University Press, 2014. p. 253–274.

BERENS, P.; DOHMEN, T.; FALK, A.; HUFFMAN, D.; SUNDE, U. Predicting student dropout: Evidence from administrative data and machine learning methods. Munich: CESifo Working Paper No. 7259, 2019. Disponível em: https://ideas.repec.org/p/ces/ceswps/_7259.html. Acesso em: 29 maio 2025.

BESSEY, D.; BACKES-GELLNER, U. Regional unemployment and educational attainment in vocational training. Economics of Education Review, v. 44, p. 1–18, 2015. Disponível em:

BERNARDI, D. P.; et al. Comparative study of gradient boosting algorithms for predictive modeling. Expert Systems with Applications, v. 173, p. 114617, 2021.https://doi.org/10.1016/j.econedurev.2014.10.003. Acesso em: 29 maio 2025.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2021.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha – Microdados da Eficiência Acadêmica – 2023. Brasília: MEC, 2023. Disponível em: https://dadosabertos.mec.gov.br/pnp/item/261-2023-microdados-eficiencia-academica. Acesso em: 31 maio 2025.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica (SETEC). Relatório anual da educação profissional e tecnológica no Brasil. Brasília, DF: MEC, 2023.

BRASIL. Presidência da República. Lei nº 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica. Diário Oficial da União: seção 1, Brasília, DF, ano 145, n. 251, p. 1, 30 dez. 2008. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm. Acesso em: 31 maio 2025.

BRASIL. Centro de Gestão e Estudos Estratégicos (CGEE). A educação profissional e tecnológica no Brasil: análise e subsídios para políticas públicas. Brasília, DF: CGEE, 2023. Disponível em: https://www.cgee.org.br. Acesso em: 27 maio 2025.

BRASIL Tribunal de Contas da União. Auditoria operacional na Rede Federal de Educação Profissional, Científica e Tecnológica: Relatório de Auditoria. Brasília: TCU, 2024.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

BURGES, Christopher J. C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, v. 2, n. 2, p. 121–167, 1998.

CATBOOST documentation. Categorical features | CatBoost. Disponível em: https://catboost.ai/docs/en/features/categorical-features. Acesso em: 10 jun. 2025.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321–357, 2002. Disponível em: https://www.jair.org/index.php/jair/article/view/10302. Acesso em: 10 jun. 2025.

DEKKER, G. W.; PECHENIZKIY, M.; VLEESHOUWERS, J. M. Predicting students drop out: a case study. In: INTERNATIONAL CONFERENCE ON EDUCATIONAL DATA MINING, 2., 2009, Córdoba. Proceedings... Córdoba: [s.n.], 2009. p. 41–50.

DOMINGUES, R.; BUENO, T. M.; LENGLER, L.; SANTOS, R. M.; FERREIRA, A. C.; OLIVEIRA, M. R. Data preprocessing techniques for machine learning. Journal of Big Data, v. 9, n. 1, p. 1–26, 2022.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, v. 29, n. 2–3, p. 103–130, 1997.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULYAEV, Andrey. CatBoost: gradient boosting with categorical features support. In: Proceedings of the Workshop on ML Systems at NIPS 2018. Montréal: NeurIPS, 2018. Disponível em: https://arxiv.org/abs/1810.11363. Acesso em: 20 de maio de 2025.

FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861–874, 2006.

GAUDREAULT, Jean-Gabriel; BRANCO, Paula; GAMA, João. An Analysis of Performance Metrics for Imbalanced Classification. In: DISCOVERY SCIENCE: 24th International Conference, DS 2021, Halifax, NS, Canada, October 11–13, 2021. Proceedings. Cham: Springer, 2021. p. 67–77. DOI: https://doi.org/10.1007/978-3-030-88942-5_6.

GÉRON, A. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Sebastopol, CA: O'Reilly Media, 2023.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 3. ed. San Francisco: Morgan Kaufmann, 2011.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 4th ed. Amsterdam: Elsevier, 2022.

HANA, Ahmed; LOFSTEAD, Jay. Managing randomness to enable reproducible machine learning. In: ACM SYMPOSIUM ON PRINCIPLES OF DISTRIBUTED COMPUTING, 41., 2022, [S.l.]. Proceedings [...]. New York: ACM, 2022. Disponível em: <https://dl.acm.org/doi/10.1145/3526062.3536353\>. Acesso em: 4 ago. 2025.

HANCOCK, J. T.; KHOSHGOFTAAR, T. M. CatBoost for big data: an interdisciplinary review. Journal of Big Data, v. 7, p. 94, 2020. Disponível em: https://doi.org/10.1186/s40537-020-00369-8. Acesso em: 15 jun. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The elements of statistical learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. ed. New York: Springer, 2017.

HERNÁN, Miguel A.; ROBINS, James M. Causal inference: what if. Boca Raton: Chapman & Hall/CRC, 2020.

KUHN, M.; JOHNSON, K. Feature Engineering and Selection: A Practical Approach for Predictive Models. Boca Raton: Chapman & Hall/CRC, 2019.

HETTIARACHCHI, D. S. S.; HARSHANATH, S. M. B. Predictive modeling for identifying early warning signs of underperformance in vocational education. In: INTERNATIONAL CONFERENCE ON ADVANCED RESEARCH IN COMPUTING (ICARC), 5., 2025, Belihuloya, Sri Lanka. Anais... Belihuloya: IEEE, 2025. p. 1–6. DOI: 10.1109/ICARC64760.2025.10962962.

HOLTMMANN, A. C.; SOLGA, H. Dropping or stopping out of apprenticeships: the role of performance- and integration-related risk factors. Zeitschrift für Erziehungswissenschaft, v. 26, p. 469–494, 2023. DOI: https://doi.org/10.1007/s11618-023-01151-1.

HUTTER, F.; HOOS, H. H.; LEYTON-BROWN, K. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer, 2019. Disponível em: ttps://doi.org/10.1007/978-3-030-05318-5. Acesso em: 27 maio 2025.

IBGE – Instituto Brasileiro de Geografia e Estatística. Recortes metropolitanos e aglomerações urbanas. Rio de Janeiro: IBGE, [2023]. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/18354-recortes-metropolitanos-e-aglomeracoes-urbanas.html. Acesso em: 31 maio 2025.

IBM. Introdução ao CRISP-DM. IBM Documentation, 2023. Disponível em: https://www.ibm.com/docs/pt-br/spss-modeler/18.4.0?topic=guide-introduction-crisp-dm. Acesso em: 15 jun. 2025.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. An Introduction to Statistical Learning: With Applications in R. 2. ed. New York: Springer, 2021.

JORDAN, M. I.; MITCHELL, T. M. Machine learning: trends, perspectives, and prospects. Science, v. 349, n. 6245, p. 255–260, 2015.

JORDAN, M. I.; MITTELMAN, J. Machine Learning: A Probabilistic Perspective. Cambridge, MA: The MIT Press, 2021.

KRÜGER, João Gabriel Corrêa; BRITTO, Alceu; BARDDAL, Jean Paul. An explainable machine learning approach for student dropout prediction. 2023. Disponível em: https://doi.org/10.2139/ssrn.4253068. Acesso em: 15 jun. 2025.

KUHN, M.; JOHNSON, K. Applied Predictive Modeling. New York: Springer, 2013.

KUHN, M.; JOHNSON, K. Feature Engineering and Selection: A Practical Approach for Predictive Models. 2nd ed. Boca Raton: CRC Press, 2019.

LIU, Z., ZHOU, X. e LIU, Y. (2025). Previsão da evasão escolar usando aprendizagem de conjunto com análise de IA explicável baseada em SHAP. Journal of Social Systems and Policy Analysis, 2(3), 111–132. https://doi.org/10.62762/JSSPA.2025.321501

LÓPEZ, Fernando. SHAP: Shapley Additive Explanations. Towards Data Science, 11 jul. 2021. Disponível em: https://towardsdatascience.com/shap-shapley-additive-explanations-5a2a271ed9c3. Acesso em: 15 jun. 2025.

LONG, J. Scott. Regression models for categorical and limited dependent variables. Thousand Oaks: Sage Publications, 1997.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, v. 30, 2017. Disponível em: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. Acesso em: 21 maio 2025.

LUNDBERG, S. M.; ERION, G.; CHEN, H. et al. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence, v. 2, p. 56–67, 2020. DOI: https://doi.org/10.1038/s42256-019-0138-9.

MARTINEZ, W. L. Computational statistics handbook with MATLAB. 3. ed. Boca Raton: Chapman and Hall/CRC, 2019.

MEEUWISSE, M.; SEVERIENS, S. E.; BORN, M. P. Goals and perceptions of the learning environment of successful and unsuccessful ethnic minority students in university education. Learning and Individual Differences, v. 20, n. 5, p. 587–591, 2010a. Disponível em: https://doi.org/10.1016/j.lindif.2010.07.002. Acesso em: 29 maio 2025.

MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997.

MOLNAR, Christoph. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2. ed. [S.l.]: Christoph Molnar, 2022. Disponível em: https://christophm.github.io/interpretable-ml-book/. Acesso em: 21 maio 2025.

MURTHY, S. K. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery, v. 2, n. 4, p. 345–389, 1998.

NICULESCU-MIZIL, A.; CARUANA, R. Predicting good probabilities with supervised learning. In: Proceedings of the 22nd international conference on Machine learning. ACM, 2005. p. 625–632.

PANG, G.; SHEN, C.; CAO, L.; VAN DEN HENGEL, A. Deep learning for anomaly detection: a review. ACM Computing Surveys, v. 54, n. 2, p. 1–38, 2021. Disponível em: https://doi.org/10.1145/3439950. Acesso em: 27 maio 2025.

PEARL, Judea. Causality: models, reasoning, and inference. 2. ed. Cambridge: Cambridge University Press, 2009.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; et al. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf. Acesso em: 23 maio 2025.

PIEPENBURG, J. G.; BECKMANN, J. The relevance of social and academic integration for students' dropout decisions: evidence from a factorial survey in Germany. European Journal of Higher Education, v. 12, n. 3, p. 255–276, 2021. Disponível em: https://doi.org/10.1080/21568235.2021.1930089. Acesso em: 15 maio 2025.

PROKHORENKOVA, L.; GUSEV, G.; VOROBEV, A.; DOROGUSH, A. V.; GULIN, A. CatBoost: unbiased boosting with categorical features. NeurIPS, 2017. Disponível em: https://arxiv.org/abs/1706.09516. Acesso em: 15 jun. 2025.

PYLE, Dorian. Data preparation for data mining. San Francisco: Morgan Kaufmann, 1999.

ROMAN, N. V.; DAVIDSE, P. E.; HUMAN-HENDRICKS, A.; BUTLER-KRUGER, L.; SONN, I. K. School dropout: intentions, motivations and self-efficacy of a sample of South Africa youth. Youth, v. 2, n. 2, p. 126–137, 2022. Disponível em: https://doi.org/10.3390/youth2020010. Acesso em: 14 jun. 2025.

RASCHKA, S. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 3rd ed. Birmingham: Packt Publishing, 2020.

RASCHKA, S.; MIRJALILI, V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 4th ed. Birmingham: Packt Publishing, 2022.

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

RUSSELL, Stuart J.; NORVIG, Peter. Inteligência artificial. 3. ed. São Paulo: Elsevier, 2010.

SHAP. Catboost tutorial — SHAP latest documentation. Disponível em: https://shap.readthedocs.io/en/latest/example_notebooks/tabular_examples/tree_based_models/Catboost%20tutorial.html. Acesso em: 17 jun. 2025.

SHAPLEY, L. S. 1953. A value for n-person games. Contributions to the Theory of Games, 307–317. Santa Monica, CA: RAND Corporation.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427–437, 2009.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. Data Mining: Practical Machine Learning Tools and Techniques. 5. ed. Cambridge, MA: Morgan Kaufmann, 2023.

6 ARTIGO 3

PLATAFORMA PREVIA: DESENVOLVIMENTO E IMPLEMENTAÇÃO DA APLICAÇÃO WEB DE PREDIÇÃO DA EVASÃO EM CURSOS TÉCNICOS NA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA DO BRASIL

Revista(s) alvo:

Journal of Social Systems and Policy Analysis - (https://www.icck.org/article/abs/jsspa.2025.321501)

Revista Brasileira da Educação Profissional e Tecnológica (RBEPT) A2 - (https://www2.ifrn.edu.br/ojs/index.php/RBEPT/about/submissions)

RESUMO

A evasão nos cursos técnicos da Rede Federal de Educação Profissional, Científica e Tecnológica do Brasil é causada por diversos fatores e demanda uma série de medidas para enfrentá-la. A evasão nesta rede de ensino reduz as chances de inserção no mundo do trabalho e limita as oportunidades de emprego em um cenário de alta demanda por mão de obra qualificada. É imprescindível a adoção de sistemas e ferramentas tecnológicas que possam identificar, de maneira preditiva e preventiva, os alunos propensos à evasão, por meio da análise de dados acadêmicos e comportamentais. O monitoramento contínuo e multidimensional é fundamental para embasar ações eficazes de combate à evasão nas instituições dessa. Este artigo descreve a implementação da plataforma web PrevIA, cujo objetivo é predizer a evasão em cursos técnicos na Rede Federal. A pesquisa utiliza análise bibliográfica e documental, utilizando métodos de inteligência artificial, com uso do algoritmo CatBoostClassifier. O modelo foi treinado com base em dados pessoais, demográficos e acadêmicos dos estudantes. Como resultado, a plataforma possibilita a realização de simulações, com a entrada de novos dados do estudante e exibe o percentual de probabilidade de evasão no curso técnico selecionado. O modelo CatBoost, treinado e serializado para uso em aplicação web, obteve um índice de recall de 69%, uma acurácia de 72% e uma área sob a curva ROC (AUC) de 78%. A plataforma mostrou ser confiável ao calcular a taxa de evasão referemte ao curso técnico, evidenciando sua capacidade de antecipar casos potenciais de evasão e auxiliar nos processos de tomada de decisão das instituições. A principal contribuição do estudo consiste na proposição de uma ferramenta web que pode auxiliar gestores e formuladores de políticas públicas na identificação de perfis de risco e na implementação de ações preventivas.

Palavras-chave: PrevIA; aprendizado de máquina; evasão escolar; cursos técnicos; Rede Federal.

ABSTRACT

Dropout rates in technical courses offered by the Federal Network for Professional, Scientific, and Technological Education of Brazil are caused by several factors and require a series of measures to address them. Dropout rates in this network compromise entry into the workforce and limit employment opportunities in a scenario of high demand for skilled labor. It is essential to adopt technological systems and tools that can predictively and preventively identify students prone to dropout through the analysis of academic and behavioral data. Continuous and multidimensional monitoring is essential to support effective actions to combat dropout in Federal Network institutions. This article describes the implementation of the PrevIA web platform, which aims to predict dropout in technical courses in the Federal Network. The research uses bibliographic and documentary analysis, employing artificial intelligence methods, with the use of the CatBoostClassifier algorithm. The model was trained based on students' personal, demographic, and academic data. As a result, the platform enables simulations to be carried out, with the entry of new student data, and displays the percentage probability of dropout in the selected technical course. The CatBoost model, trained and serialized for use in web applications, achieved a recall rate of 69%, an accuracy of 72%, and an area under the ROC curve (AUC) of 78%. The platform proved to be reliable in calculating the dropout rate in technical courses, demonstrating its ability to anticipate potential cases of dropout and assist institutions in their decision-making processes. The main contribution of the study is the proposal of an online tool that can assist managers and public policy makers in identifying risk profiles and implementing preventive actions.

Keywords: PrevIA; machine learning; school dropout; technical courses; Federal Network.

1 INTRODUÇÃO

A formação profissional desempenha um papel fundamental para o fortalecimento da força de trabalho de um país, ao fornecer aos indivíduos habilidades práticas e conhecimentos especializados voltados para o mundo do trabalho. Esse tipo de educação é especialmente relevante em países em desenvolvimento, pois contribui diretamente para a diminuição do desemprego e para a promoção do crescimento econômico (PETNUCHOVA et al., 2012).

A interrupção precoce dos estudos em cursos técnicos ou em programas de aprendizagem profissional não apenas compromete o projeto de vida do estudante e a sua empregabilidade, mas também impacta negativamente a competitividade da economia, na

medida em que a formação profissional é um dos pilares centrais para a inovação e a produtividade de um país. (UNESCO, 2023).

Para Bessey et al (2015) a evasão escolar se define como um entrave ao desenvolvimento socioeconômico, especialmente em um cenário de aceleradas transformações no mundo do trabalho, que intensifica a demanda por mão de obra qualificada. Constitui um dos desafios mais persistentes da educação contemporânea, produzindo impactos multidimensionais. No nível institucional, a evasão representa perdas significativas de recursos humanos e financeiros, comprometendo a eficiência dos investimentos públicos e a sustentabilidade das políticas educacionais (TOMLINSON; WALKER, 2022; VAARMA; LI, 2024).

No âmbito da educação profissional e tecnológica (EPT), a evasão escolar apresenta-se como um problema de alcance global, despertando crescente interesse acadêmico e político. Pesquisas internacionais evidenciam que diferentes países enfrentam desafios semelhantes na identificação das causas e na adoção de estratégias de permanência e conclusão dos cursos (OECD, 2020; UNESCO, 2018; UNESCO, 2023).

O fenômeno da evasão escolar desperta interesse em diversos segmentos educacionais, muito por conta de sua complexidade e abrangência. Sua ocorrência gera preocupações tanto em instituições públicas quanto privadas, dado que a saída precoce de alunos acarreta implicações nas esferas social, acadêmica e econômica (BAGGI; LOPES, 2011).

A literatura sobre evasão em cursos técnicos aponta consenso quanto à influência de fatores socioeconômicos, institucionais e individuais na permanência estudantil. Contudo, persistem divergências sobre quais variáveis exercem maior peso na decisão de abandono (YI et al., 2015). Predominam análises voltadas ao comportamento individual do estudante, enquanto aspectos contextuais e institucionais ainda carecem de investigação mais aprofundada (BÖHN; DEUTSCHER, 2022).

Nesse contexto, o aprendizado de máquina tem se destacado como uma ferramenta promissora para identificar padrões e prever comportamentos a partir da análise de dados. No contexto educacional, algoritmos de classificação têm sido amplamente aplicados para prever a probabilidade de evasão escolar, possibilitando ações preventivas e estratégias de retenção estudantil mais eficazes (BAKER; SIEMENS, 2014; CHO; YU; KIM, 2023; VAARMA; LI, 2024).

Sob essa perspectiva, o aprendizado supervisionado, tem se mostrado eficaz na predição de eventos educacionais, como a evasão escolar. Entre os modelos de aprendizado de máquina, o CatBoost destaca-se por sua elevada capacidade de generalização e pelo tratamento eficiente

de variáveis categóricas, sem necessidade de processo complexo de codificação, o que reduz vieses e melhora o desempenho do modelo (DOROGUSH; ERSHOV; GULYAEV, 2018; MDUMA; KALEGELE; MACHUVE, 2019; PROKHORENKOVA et al., 2018). Assim, o CatBoost apresenta-se como um algoritmo robusto para a previsão da evasão em cursos técnicos ofertados pela Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT), contribuindo para a identificação antecipada de estudantes em risco e para o aprimoramento das estratégias de permanência estudantil.

Diante desse cenário, este artigo propõe-se a apresentar a plataforma web PrevIA - Predição de Evasão na Rede Federal com Inteligência Artificial (IA), capaz de predizer o percentual de evasão em cursos técnicos ofertados pela RFEPCT. A plataforma utiliza o algoritmo de aprendizado de máquina CatBoostClassifier para calcular a probabilidade de evasão dos estudantes a partir de novos dados inseridos na ferramenta. O objetivo é fornecer subsídios para a implementação de ações institucionais preventivas e orientar a formulação de políticas públicas voltadas à permanência e ao êxito estudantil.

2 REVISÃO DA LITERATURA

2.1 Evasão escolar

A evasão escolar, que se refere à interrupção do vínculo do aluno com o curso antes de sua conclusão, seja por abandono, desistência formal, cancelamento de matrícula, transferência ou desligamento por razões administrativas ou disciplinares, é um fenômeno de alta complexidade e com múltiplas facetas, que tem sido bastante discutido no campo educacional nas últimas décadas (TINTO, 1975). Em países em desenvolvimento, a criação de um sistema educacional livre desse problema ainda é um objetivo distante, considerando a magnitude dos desafios socioeconômicos que afetam diretamente a permanência dos alunos (MDUMA; KALEGELE; MACHUVE, 2019).

Böhn e Deutscher (2022) afirmam que os fatores relacionados à evasão escolar são reconhecidos e constituem um obstáculo significativo ao desenvolvimento educacional. Isso se deve à interação dinâmica entre variáveis de natureza sociocultural, econômica e institucional.

De acordo Silva, Filho e Fernandes (2024), a evasão escolar revela a ausência de condições adequadas para a permanência dos estudantes. A persistência dos elevados índices evidencia a necessidade de aprofundar as investigações sobre as causas da evasão e de fortalecer

a implementação de políticas públicas educacionais que assegurem a continuidade e a conclusão dos estudos. Nesse sentido, é fundamental que as ações governamentais priorizem estratégias integradas de acompanhamento pedagógico, apoio socioeconômico e valorização do estudante.

A evasão escolar é um desafio global enfrentado por instituições de ensino, cujos impactos negativos repercutem não apenas na trajetória dos estudantes, mas também na comunidade e nas próprias unidades educacionais. Diante dessa problemática, torna-se necessário adotar medidas preventivas que permitam a identificação precoce de discentes em situação de risco, bem como a implementação de estratégias de intervenção direcionadas e efetivas (RUMBERGER, 2018).

2.2 Rede Federal de Educação Profissional, Científica e Tecnológica

Na organização da oferta de educação profissional e tecnológica do Brasil, encontra-se a Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT). Criada em 2008 pela Lei nº 11.892, de 29 de dezembro, a Rede Federal constitui um marco na ampliação, interiorização e diversificação da educação profissional e tecnológica no país (BRASIL, 2008).

Reconhecida pela qualidade do ensino ofertado, pela diversidade de cursos e por sua atuação junto à população e aos arranjos produtivos locais, a Rede Federal busca potencializar as características regionais em termos de trabalho, cultura e lazer (BRASIL, 2023).

Integrante do sistema federal de ensino vinculado ao Ministério da Educação (MEC), a RFEPCT é composta por diversos tipos de instituições, como os Institutos Federais de Educação, Ciência e Tecnologia, a Universidade Tecnológica Federal do Paraná, Centros Federais de Educação Tecnológica (Cefets), escolas técnicas vinculadas a universidades federais e o Colégio Pedro II (BRASIL, 2008).

Em 2024, a Rede Federal era formada por 64 Instituições sendo, 38 Institutos Federais, 02 Centros Federais de Educação Tecnológica (Cefet), a Universidade Tecnológica Federal do Paraná (UTFPR), 22 escolas técnicas vinculadas às universidades federais e o Colégio Pedro II. Considerando os respectivos campi associados a estas instituições federais, tem-se ao todo 686 unidades de ensino distribuídas entre as 27 unidades federativas do país (BRASIL, 2024).

Essas instituições possuem autonomia administrativa, patrimonial, financeira, didático-pedagógica e disciplinar.

No âmbito do MEC, compete à Secretaria de Educação Profissional e Tecnológica (Setec) o planejamento e o desenvolvimento da RFEPCT, incluindo a garantia de adequada disponibilidade orçamentária e financeira (BRASIL, 2023).

2.3 Evasão na Rede Federal de EPCT

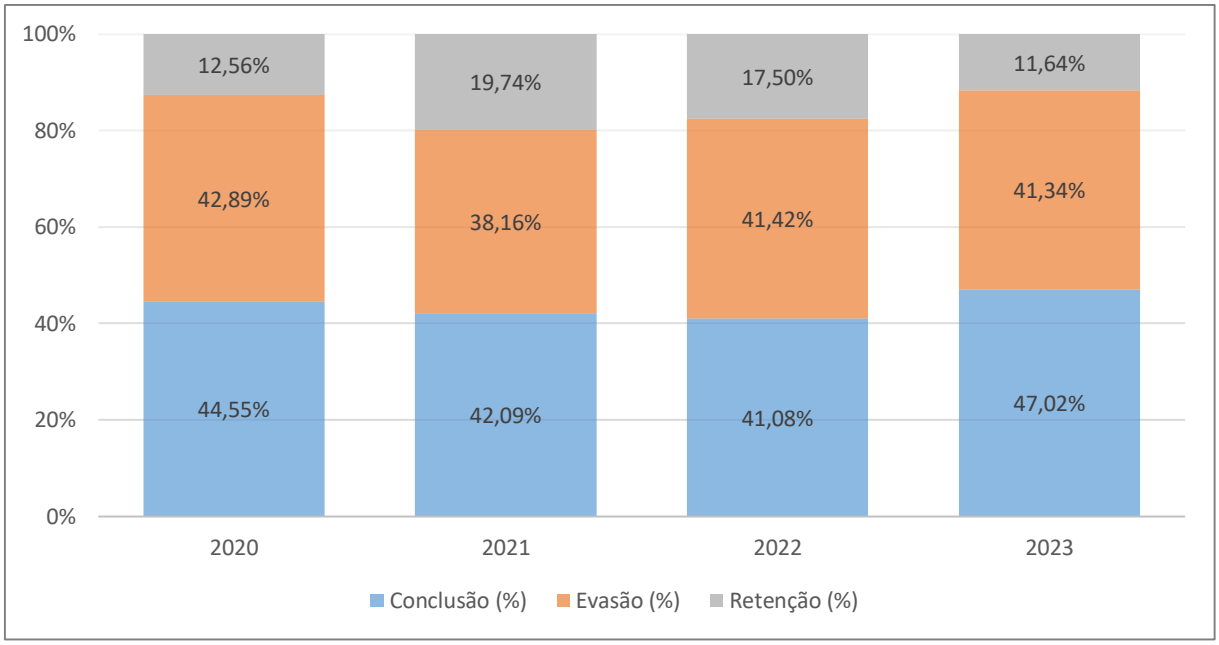
Os dados quantitativos referentes às matrículas e às taxas de evasão nos cursos técnicos da RFPCT são sistematicamente consolidados e divulgados através da plataforma Nilo Peçanha (PNP), conforme estabelecido pelo MEC (BRASIL, 2024). Esses indicadores são publicados anualmente, após o encerramento do período letivo, constituindo-se como importante ferramenta de monitoramento para gestores educacionais.

O Plano Nacional de Educação (PNE) 2014-2024, prorrogado até 2025, estabelece uma meta ambiciosa relacionada à educação profissional técnica de nível médio no Brasil, ao prever que pelo menos 90% dos matriculados concluam seus cursos (BRASIL, 2014). Essa meta reflete o compromisso em ampliar a eficiência e a permanência dos estudantes na educação profissional, reduzindo os índices de evasão que historicamente comprometem a efetividade das políticas públicas educacionais (YI et al., 2015).

O Tribunal de Contas da União (TCU) classificou, em 2024, a evasão escolar na Rede Federal de EPCT como um fenômeno complexo, influenciado por múltiplos fatores estruturais, financeiros e institucionais. Segundo auditoria realizada por este Tribunal, diversos elementos impactam negativamente a permanência e o êxito dos estudantes, contribuindo para o aumento das taxas de evasão. Entre os principais desafios identificados estão a insuficiência de recursos para assistência estudantil e alimentação escolar, carência de equipes multiprofissionais, deficiências de infraestrutura e ausência de políticas consolidadas de transporte escolar (BRASIL, 2024).

A figura 1 apresenta as taxas de evasão nos cursos técnicos em turmas concluídas na RFEPCT no período de 2020 a 2023.

Figura 1 - Taxa de eficiência em cursos técnicos na RFEPCT (2020-2023)



Fonte: elaborado pelo autor a partir de dados da Plataforma Nilo Peçanha, 2025.

2.4 Uso de Inteligência Artificial na evasão escolar

O Aprendizado de Máquina (AM) constitui um campo de estudo dedicado à análise, ao desenvolvimento e à aplicação de algoritmos capazes de gerar modelos preditivos ou decisórios de maneira indutiva, utilizando dados como base para a extração de padrões e inferências (KORKMAZ; AYDIN, 2025; ROMERO; VENTURA, 2024).

Para Rastrollo-Guerrero, Gómez-Pulido e Durán-Domínguez (2020), a utilização de técnicas de aprendizado de máquina para prever a evasão escolar tem se consolidado como um tema amplamente explorado na pesquisa educacional, evidenciando não apenas avanços metodológicos significativos, mas também a geração de resultados práticos relevantes. Estudos recentes destacam a capacidade desses modelos em identificar estudantes em risco, permitindo a implementação de intervenções pedagógicas mais assertivas e oportunas (MUSTOFA et al., 2025; SEO et al., 2024).

Entre as técnicas mais comumente utilizadas em aprendizado de máquina destacam-se: árvores de decisão (MURTHY, 1998), redes neurais (MITCHELL, 1997; HASTIE; TIBSHIRANI; FRIEDMAN, 2009), naïve Bayes (DOMINGOS; PAZZANI, 1997), random forest (BREIMAN, 2001), K-Nearest Neighbors (MITCHELL, 1997), regressão logística (LONG, 1997) e máquinas de vetores de suporte (BURGES, 1998). Além disso, técnicas de

baseadas em *Gradient Boosting* como, XGBoost, LightGBM e CatBoost, são atualmente bastante utilizadas (CHEN; GUESTRIN, 2016).

2.4.1 Algoritmos de Gradient Bosting

Os algoritmos de *Gradient Boosting* são abordagens de aprendizado de máquina fundamentadas em técnicas de *ensemble*, as quais integram diversos modelos para criar um modelo preditivo robusto (FRIEDMAN, 2001).

Entre as implementações mais conhecidas estão o XGBoost, o LightGBM e o CatBoost. O XGBoost (*Extreme Gradient Boosting*) é reconhecido por sua eficiência computacional e desempenho de destaque em tarefas de classificação e regressão (CHEN; GUESTRIN, 2016).

O LightGBM, desenvolvido pela Microsoft, destaca-se por sua capacidade de gerenciar grandes volumes de dados e por empregar técnicas como amostragem baseada em gradiente e agrupamento exclusivo de características, aumentando a eficiência e reduzindo o uso de memória (KE et al., 2017).

Por outro lado, o CatBoost, criado pela Yandex, foi desenvolvido para lidar de forma eficiente com variáveis categóricas, incorporando técnicas como o *Ordered Boosting* e *Target Statistics* para prevenir o vazamento de dados durante o treinamento. Além disso, o CatBoost realiza a conversão automática de atributos categóricos, diminuindo a necessidade de pré-processamento. O modelo é robusto, com alta acurácia e tem se mostrado eficiente mesmo em conjuntos de dados variados e desbalanceados (DOROGUSH et al., 2018; PROKHORENKOVA et al., 2018).

2.5 Plataforma web para predição da evasão escolar

O uso de ferramentas baseadas em dados para identificar e prevenir a evasão escolar se estabelece como uma estratégia importante para políticas públicas e gestão educacional. Com essas tecnologias, é possível identificar precocemente alunos em risco, avaliar cenários e priorizar intervenções com base em evidências empíricas. Para Vaarma e Li (2024a), a integração de análises descritivas e previsões automáticas possibilita que gestores e equipes pedagógicas realizem intervenções mais eficazes e direcionadas (NIYOGISUBIZO et al., 2022).

Embora as ferramentas tecnológicas e os algoritmos de predição desempenhem um papel crucial na identificação de padrões associados ao abandono escolar, é fundamental ressaltar a centralidade da ação pedagógica na implementação de intervenções eficazes (QUAYE; HARPER; PENDAKUR, 2019). A análise de dados e as previsões de evasão, por si só, não asseguram a permanência dos estudantes. A sua efetividade depende da capacidade de educadores e gestores em transformar essas informações em estratégias personalizadas que favoreçam a motivação, o engajamento e o acompanhamento contínuo dos alunos (NAUJOKAITIENĖ et al., 2020). Nesse sentido, a integração entre tecnologias de aprendizado de máquina e práticas pedagógicas contextualizadas representa um caminho promissor para o fortalecimento da retenção escolar na educação profissional e tecnológica (NIYOGISUBIZO et al., 2022; VAARMA; LI, 2024).

De acordo com Hettiarachchi e Harshanath (2025), a eficácia dessa estratégia está diretamente relacionada ao monitoramento de variáveis essenciais. O acompanhamento da vida escolar, o engajamento e os fatores socioeconômicos dos estudantes são fundamentais para evitar o baixo desempenho estudantil e a consequente evasão. Ao incorporar modelos preditivos ao processo educacional, as instituições de educação profissional têm a capacidade de abordar essas questões de forma proativa, o que resulta em maior retenção e aumento das taxas de conclusão dos cursos.

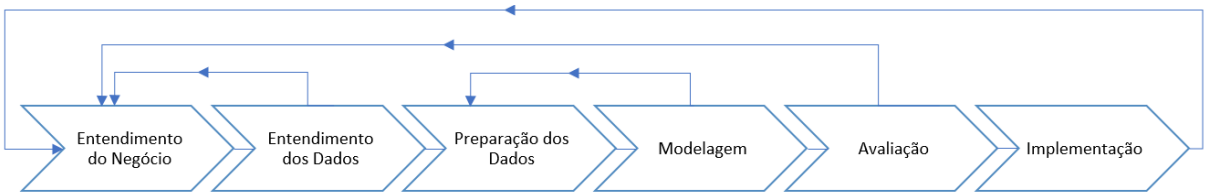
3 MATERIAL E MÉTODOS

A metodologia do projeto foi organizada em seis etapas fundamentais: entendimento do negócio; entendimento dos dados por meio da coleta e pré-processamento; preparação dos dados com ênfase na engenharia de atributos e análise exploratória; criação dos modelos preditivos; avaliação de desempenho; e, finalmente, implementação do modelo escolhido.

A metodologia CRISP-DM (*CRoss Industry Standard Process for Data Mining*) (IBM, 2023; MARTINEZ, 2019) foi empregada para direcionar o desenvolvimento da parte computacional do estudo. O CRISP-DM trata de partes de um problema ao estabelecer um modelo de processo que oferece uma estrutura para a realização de projetos, independentemente da tecnologia empregada (CHAPMAN, 2000 MARBÁN; MARISCAL; SEGOVIA, 2009; SCHRÖER; KRUSE; GÓMEZ, 2021).

O ciclo de vida de projetos de mineração de dados que empregam o CRISP-DM é ilustrado na figura 2.

Figura 2 - Ciclo de vida de projeto com CRISP-DM



Fonte: elaborado pelo autor, adaptado de IBM, 2025.

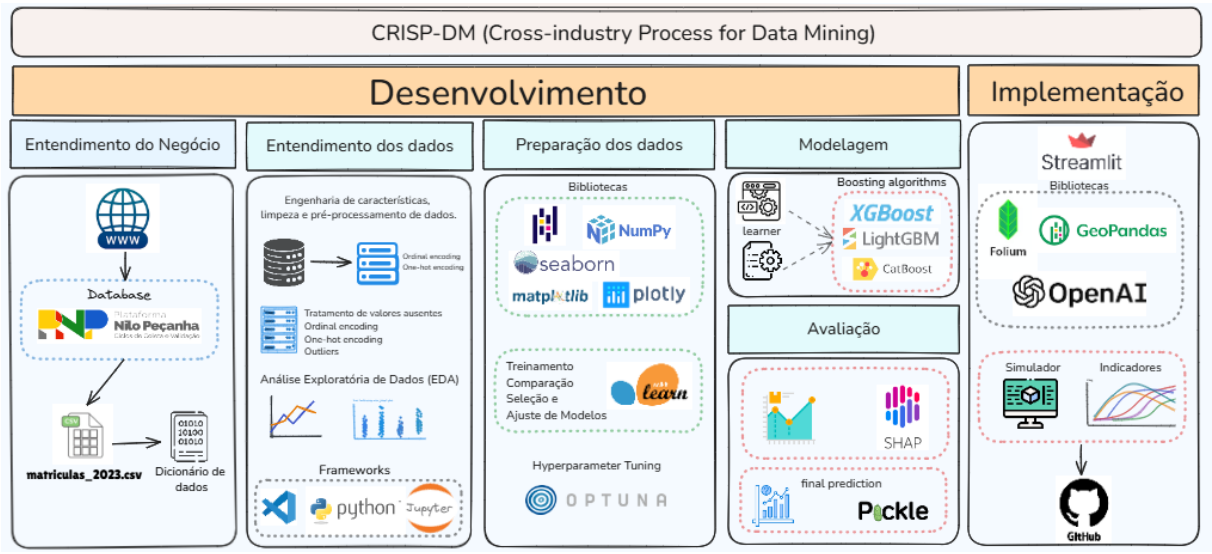
Como uma metodologia, inclui a descrição das etapas típicas de um projeto, as atividades necessárias para cada etapa e a explicação das interações entre essas atividades.

O *pipeline* no projeto, baseado na estrutura CRISP-DM, foi composto pelas seguintes etapas:

- Entendimento do negócio: definição clara do problema de negócio a ser resolvido.
- Entendimento dos dados: coleta de dados e análise inicial para entender sua natureza e qualidade.
- Preparação dos dados: realização da limpeza, transformação, engenharia de atributos e análise exploratória dos dados.
- Modelagem: separação dos dados em conjuntos de treinamento e teste, treinamento dos modelos de AM, realização de comparações e características e ajuste de hiperparâmetros.
- Avaliação: análise crítica do desempenho dos modelos construídos, considerando métricas apropriadas para classificação e avaliação e interpretação do modelo final com base em seu desempenho.
- Implementação: consolidação da análise final e implantação do modelo preditivo selecionado.

A figura 3 apresenta uma visão geral do fluxo de execução dessas etapas, destacando as atividades realizadas e os artefatos utilizados em cada fase do processo.

Figura 3 - Etapas de elaboração do projeto baseado no CRISP-DM



Fonte: elaborado pelo autor, 2025.

Este artigo foca principalmente na sexta etapa da metodologia: a implementação do modelo usando o algoritmo CatBoostClassifier. O objetivo é simular a entrada de novos dados de estudantes e apresentar percentuais de chance de evasão, a fim de prever a evasão escolar em cursos técnicos na RFEPCT.

3.1 Conjunto de dados e pré-processamento de dados

Os dados utilizados neste estudo foram obtidos no Portal de Dados Abertos do Governo Federal (BRASIL, 2024), tendo como fonte a PNP, disponibilizada pelo Ministério da Educação (BRASIL, 2023).

A base contempla informações referentes às matrículas de alunos da RFEPCT em 2023, com um total 145.831 registros. Os dados abrangem informações demográficas, institucionais e acadêmicas vinculadas aos cursos técnicos. É importante ressaltar que os dados já estão classificados com base na variável que se refere à categoria da situação de matrícula.

A análise delimitou-se apenas os dados de matrículas em cursos do tipo técnico. Essa opção se baseia na legislação da RFEPCT, que estabelece, no mínimo, 50% da oferta educacional destinada à educação técnica de nível médio (BRASIL, 2008). Além disso, a exigência de uma carga horária mínima de 800 horas, exigida para esses cursos, possibilita um acompanhamento mais consistente da trajetória acadêmica dos estudantes em base anual.

Além da base principal, foi adicionada ao projeto uma base de dados referente às regiões metropolitanas do Brasil, disponibilizada pelo Instituto Brasileiro de Geografia e Estatística (IBGE) (BRASIL, 2023b). O objetivo dessa integração foi identificar as unidades de ensino

situadas nessas regiões, a fim de investigar a influência da localização geográfica na probabilidade de evasão escolar. A hipótese é que fatores territoriais, como infraestrutura urbana, mobilidade e acesso a oportunidades, podem ter impacto considerável na permanência dos estudantes nos cursos técnicos.

3.1.1 Dicionário de Dados

A descrição das variáveis presentes no conjunto de dados usado na análise é apresentada na tabela a seguir. O objetivo é proporcionar uma visão clara e organizada dos atributos que constituem a base, tornando o estudo mais compreensível e reproduzível.

Tabela 1 - Descrição das variáveis da base de dados

Variável	**Descrição**	**Tipo**
Categoria situação (Alvo)	Situação do aluno em relação à matrícula: *em curso, concluído* ou *evadido*.	Categórica nominal
Cor/raça	Cor ou raça autodeclarada do aluno *Branca, Preta, Parda, Amarela* ou *Indígena ou Não informado*.	Categórica nominal
Idade	Idade do aluno no momento da matrícula ou da coleta dos dados, em anos. Varia entre 4 e 84 anos.	Numérica discreta
Sexo	Gênero do aluno: *Masculino, Feminino* ou *Não informado*.	Categórica nominal
Renda familiar	Faixas salariais ordenadas: $0 < RFP \leq 0,5$; $0,5 < RFP \leq 1,0$; $1,0 < RFP \leq 1,5$; $1,5 < RFP \leq 2,5$; $2,5 < RFP \leq 3,5$; $RFP > 3,5$; *Não declarada*.	Categórica ordinal
Modalidade de ensino	Modalidade em que o aluno estuda: *presencial* ou *educação a distância*.	Categórica nominal
Tipo de oferta	Forma de ingresso no curso técnico: *Subsequente, Proeja - Subsequente, Concomitante, Proeja – Concomitante, Integrado ou Proeja - Integrado*.	Categórica nominal
Turno	Turno de realização das aulas: *Matutino, Vespertino, Noturno* ou *Integral*.	Categórica nominal
Nome de curso	Nome do curso técnico no qual o aluno está matriculado. Registro de 140 diferentes cursos técnicos ofertados.	Categórica nominal
Eixo tecnológico	Área de conhecimento ou eixo tecnológico ao qual o curso pertence. Agrupamento dos cursos em 13 eixos tecnológicos, conforme o Catálogo Nacional de Cursos Técnicos (CNCT).	Categórica nominal
Carga horária mínima	Carga horária mínima exigida para conclusão do curso, expressa em horas. Três categorias de carga horária mínima exigida: 800, 1.000 e 1.200 horas, conforme CNCT.	Numérica discreta

Variável	Descrição	Tipo
UF (Unidade da Federação)	Estado da instituição de ensino (ex.: *SP*, *RJ*, *BA*). Representação das 27 Unidades Federativas do Brasil.	Categórica nominal
Município	Município onde está localizada a unidade de ensino. Registro de matrículas em 550 municípios distintos.	Categórica nominal
Região	Corresponde às cinco grandes regiões geográficas do Brasil: *Norte*, *Nordeste*, *Centro-Oeste*, *Sudeste* e *Sul*.	Categórica nominal
Instituição	Nome da instituição da Rede Federal à qual a unidade pertence. Conjunto de 62 instituições da Rede Federal de EPCT.	Categórica nominal
Unidade de ensino	Nome da unidade da instituição onde o curso técnico é ofertado. Matrículas em 618 unidades de ensino distribuídas em todo o território nacional.	Categórica nominal
Região metropolitana da UE (Unidade de Ensino)	Indica se a unidade de ensino está localizada em uma região metropolitana (sim ou não). Indica se a unidade de ensino está situada em região metropolitana, com os valores *Sim* ou *Não*.	Categórica nominal

Fonte: elaborado pelo autor, 2025.

3.2 Engenharia de atributos: pré-processamento, limpeza, transformação e análise exploratória dos dados.

A análise das propriedades principais do conjunto de dados foi realizada com o propósito de obter um entendimento completo da estrutura informacional e detectar possíveis inconsistências antes da modelagem preditiva A análise exploratória de dados é uma etapa fundamental para garantir a confiabilidade dos dados, orientar as decisões de pré-processamento e definir estratégias de modelagem apropriadas. Nesse contexto, diversos aspectos foram analisados, começando pela avaliação de valores nulos, que possibilitou a identificação de variáveis com dados ausentes e a avaliação de sua proporção em relação ao total de registros (PYLE, 1999). Posteriormente, procedeu-se à contagem de valores únicos por variável, com foco nas variáveis categóricas, a fim de analisar sua cardinalidade e a necessidade de empregar técnicas de codificação específicas, como *one-hot encoding* ou *label encoding*, levando em conta a sensibilidade de certos algoritmos a esse aspecto (GÉRON, 2019). Adicionalmente, a visualização do volume de dados por variável ajudou a identificar inconsistências, desequilíbrios ou erros no preenchimento.

Além disso, métodos como o intervalo interquartil (IQR) e o desvio padrão foram utilizados para identificar outliers, devido ao impacto negativo que valores extremos podem ter no desempenho de modelos de aprendizado de máquina (HAN; KAMBER; PEI, 2022). A

distribuição dos dados também foi analisada, avaliando-se o comportamento estatístico das variáveis numéricas, incluindo média, mediana, dispersão e assimetria e das variáveis categóricas, por meio de frequências absolutas e relativas. A análise da variável alvo foi outro aspecto importante. O foco estava na distribuição das categorias “evadido”, “em curso” e “concluído”, a fim de verificar o balanceamento entre as classes. Finalmente, investigou-se a conexão entre as variáveis explicativas e a variável alvo, com o objetivo de encontrar padrões preditivos significativos.

A etapa de construção e integração de dados incluiu a combinação do conjunto principal com bases externas, incluindo informações geográficas provenientes do IBGE, relacionadas às regiões metropolitanas. Essa integração permitiu o enriquecimento da análise sob uma perspectiva territorial e socioeconômica, ampliando o potencial explicativo do modelo. A variável alvo CATEGORIA_SITUACAO foi convertida para formato binário (evadido = 1; não evadido = 0), de modo a viabilizar sua aplicação em algoritmos de classificação supervisionada (PANG et al., 2021).

No pré-processamento dos dados, as variáveis categóricas foram codificadas de acordo com sua natureza: as ordinais foram convertidas usando a técnica *Ordinal Encoder*, enquanto as nominais foram processadas com *OneHotEncoder*, funções presentes na biblioteca scikit-learn (PEDREGOSA et al., 2011). Em seguida, utilizando a técnica *Stratified KFold*, que mantém a proporção da variável-alvo nas amostras, o conjunto de dados foi dividido em subconjuntos de treino e teste. Esse processo é fundamental para garantir a representatividade e evitar viés na análise de desempenho dos modelos (PEDREGOSA et al., 2011; JAMES; WITTEN; HASTIE; TIBSHIRANI, 2021).

Uma fase importante desse processo foi a padronização e substituição de valores da variável "Cor/Raça", a fim de assegurar a consistência categórica e a validade estatística das análises que se seguiram. É comum em bases educacionais registros incompletos, inconsistentes ou ausentes em variáveis sensíveis, como a autodeclaração étnico-racial. Isso pode afetar a qualidade da modelagem e a interpretação dos resultados (BRASIL, 2023; HAN; KAMBER; PEI, 2022).

A exclusão de colunas também se mostrou essencial devido a critérios de análise e computação. Em projetos de mineração de dados, é aconselhável remover atributos de alta cardinalidade ou baixa relevância preditiva, uma vez que isso simplifica o modelo, melhora o tempo de processamento e reduz a probabilidade de *overfitting* (DOMINGUES et al., 2022; HAN; KAMBER; PEI, 2022). Dessa forma, optou-se por excluir as variáveis “município” e

“unidade_de_ensino”, que possuíam um grande número de categorias distintas, que geraram altos custos computacionais nos processos de codificação e armazenamento. A variável “renda_familiar”, originalmente categórica e ordinal, também foi removida após a criação de uma coluna numérica codificada, a fim de evitar redundância e assegurar maior clareza no conjunto final de atributos.

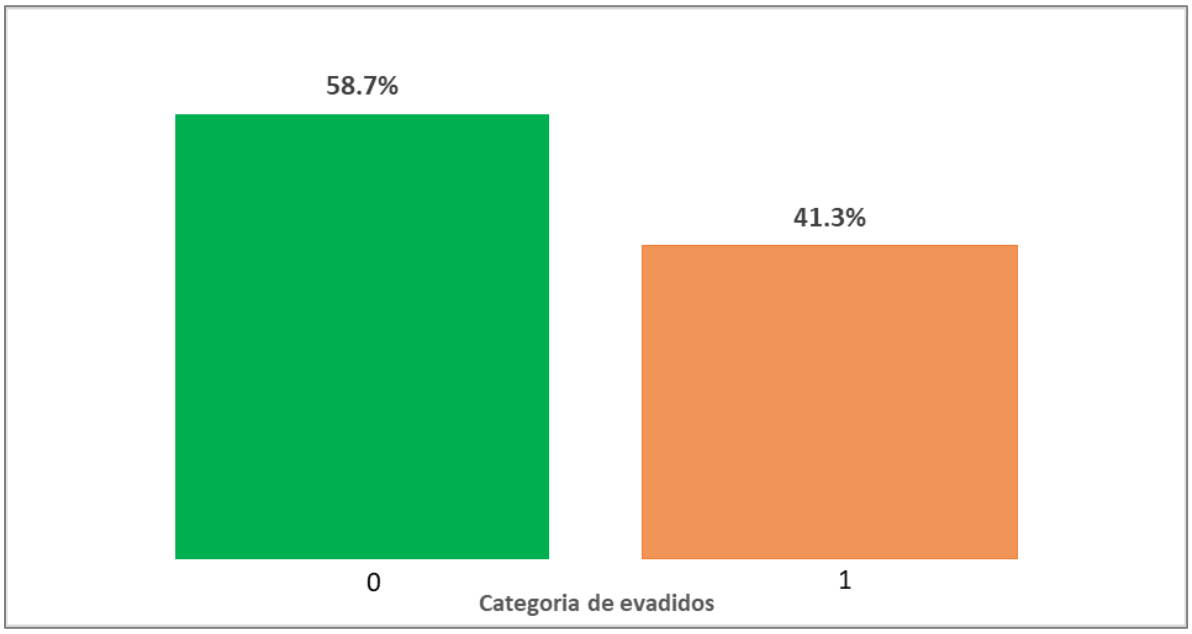
Para tratar a variável "Renda Familiar", empregou-se o *Ordinal Encoder*, dado que suas faixas exibem uma ordenação natural de menor para maior. Essa abordagem transforma categorias em valores inteiros em ordem crescente, mantendo a hierarquia implícita entre as classes. A estratégia foi fundamental para que os modelos de aprendizado supervisionado compreendessem de forma adequada a progressão de renda (PEDREGOSA et al., 2011; HAN; KAMBER; PEI, 2022).

Posteriormente, utilizou-se o *OneHotEncoder* para converter as demais variáveis categóricas nominais em formato numérico, assegurando a compatibilidade com os algoritmos de aprendizado de máquina, que necessitam de dados quantitativos como entrada. O método cria uma variável binária para cada categoria única, atribuindo o valor 1 quando a observação está na categoria correspondente e 0 caso contrário. As variáveis convertidas foram: categoria_situacao, cor_raca, sexo, modalidade_de_ensino, tipo_de_oferta, turno, eixo_tecnologico, nome_do_curso, instituicao, uf, regiao e regiao_metropolitana_ue.

Finalmente, a técnica n–1 foi utilizada para eliminar colunas redundantes resultantes do processo de *one-hot encoding*. Essa abordagem tem como objetivo evitar a multicolinearidade, fenômeno que acontece quando variáveis fortemente correlacionadas afetam a estabilidade e a interpretabilidade dos modelos (JAMES et al., 2021; KUHN; JOHNSON, 2013). Ao eliminar uma coluna de cada conjunto codificado, preserva-se a independência linear entre os atributos, o que confere maior solidez às análises preditivas.

Após a finalização da fase de engenharia de atributos e pré-processamento, o conjunto de dados final apresentou um leve desbalanceamento entre as classes da variável alvo, como mostrado na figura 4. A categoria predominante, referente aos alunos não evadidos, correspondeu a cerca de 58,7% dos registros, ao passo que a categoria minoritária, relacionada aos evadidos, compreendeu 41,3% dos registros. Embora esse desequilíbrio seja moderado, ele requer atenção, pois a distribuição desigual das classes pode gerar vieses no processo de aprendizado dos algoritmos de classificação (KOTSIANTIS, 2013; BRANCO et al., 2016).

Figura 4 - Percentual de evadidos registrado na base de dados utilizada



Fonte: elaborado pelo autor, 2025.

3.3 Etapa da modelagem

Durante a fase de modelagem, foram utilizados métodos de aprendizado de máquina para criar modelos preditivos destinados a reconhecer padrões associados à evasão escolar. Empregaram-se algoritmos tradicionais e de *boosting*, como XGBoost, LightGBM e CatBoost, conhecidos por sua eficácia e performance em tarefas de classificação supervisionada (CHEN; GUESTRIN, 2016; KE et al., 2017; PROKHORENKOVA et al., 2018). O treinamento e a comparação dos modelos foram realizados no *framework scikit-learn* (PEDREGOSA et al., 2011; SCIKIT-LEARN, 2025), utilizando métricas padronizadas e validação cruzada para assegurar a robustez dos resultados. A biblioteca Optuna, que automatiza a busca pelos melhores parâmetros de configuração (AKIBA et al., 2019), foi utilizada para otimizar os hiperparâmetros. Após as devidas modificações, o modelo final foi serializado com uso da biblioteca Pickle versão 3.13.5 do Python, permitindo sua integração à aplicação web desenvolvida.

3.4 Etapa de avaliação

Na fase de avaliação, analisou-se o desempenho dos modelos treinados para escolher o que apresenta a melhor capacidade preditiva e equilíbrio entre as métricas de acurácia, precisão, recall, F1-score e ROC-AUC, de acordo com as recomendações clássicas para a validação de modelos supervisionados (JAMES et al., 2021). A biblioteca SHAP (SHapley Additive

exPlanations) foi empregada para melhorar a interpretabilidade dos resultados, possibilitando a quantificação da contribuição de cada variável nas previsões do modelo e proporcionando transparência e explicabilidade nos processos de decisão algorítmica (LUNDBERG; LEE, 2017). O modelo com melhor desempenho foi validado e, em seguida, integrado à plataforma PrevIA, para a aplicação prática da análise preditiva.

3.5 Etapa de Implementação

A plataforma PrevIA (Predição de Evasão na Rede Federal com IA) foi implementada utilizando bibliotecas consolidadas do ecossistema Python (versão 3.11.4), com foco na acessibilidade e na integração dos recursos analíticos com a interface web. O código foi desenvolvido no Visual Studio Code, ambiente de programação da Microsoft.

Para a elaboração da plataforma optou-se pela biblioteca Streamlit (versão 1.47.0) devido à sua capacidade de desenvolver aplicações web interativas e visualmente atraentes de maneira rápida e descomplicada, qualidades especialmente apropriadas para projetos focados em ciência de dados, aprendizado de máquina e análise interativa. Nesse cenário, o Streamlit teve um papel fundamental ao disponibilizar o modelo preditivo em um ambiente interativo, possibilitando a manipulação intuitiva dos dados e a visualização dos resultados em tempo real.

As bibliotecas GeoPandas (versão 1.1.1) e Folium (versão 0.20.0) foram utilizadas para a análise espacial da evasão, permitindo a criação de mapas interativos e a visualização geográfica da distribuição da evasão nos cursos técnicos da Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT).

Além disso, a API da OpenAI (versão 0.28.0) foi integrada como um recurso auxiliar para a criação de *insights* textuais para análise dos dados de evasão em cursos técnicos.

4 RESULTADOS

4.1 Modelagem computacional

A modelagem incluiu o uso de algoritmos de aprendizado de máquina para prever a evasão escolar, utilizando um conjunto de variáveis explicativas. Primeiramente, os dados

foram separados em dois subconjuntos: 80% para treinamento e 20% para teste. Essa divisão assegurou que os dados de teste não fossem usados durante o treinamento, uma prática recomendada para prevenir o vazamento de dados e possibilitar uma avaliação mais precisa do desempenho dos modelos (GOODFELLOW; BENGIO; COURVILLE, 2016; GERON, 2023; KUHN; JOHNSON, 2013).

Além disso, o controle da aleatoriedade na divisão foi realizado por meio do parâmetro *random_state* da função *train_test_split()*, garantindo que a amostragem pudesse ser reproduzida em execuções futuras do código, assegurando a comparabilidade entre as diferentes configurações de modelos. (HANA; LOFSTEAD, 2022).

Para calcular a evasão escolar, diversos algoritmos de aprendizado de máquina foram treinados e comparados, incluindo: modelos lineares, como Regressão Logística, Linear SVC e Support Vector Machine (SVM); modelos em instância, como K-Nearest Neighbors (KNN); modelos baseados em árvore de decisão, como Decision Tree e Random Forest; e modelos de gradiente boosting, como XGBoost, LightGBM e CatBoost.

No pré-processamento, a função *fit_transform()* foi utilizada apenas nos dados de treino, enquanto os dados de teste foram tratados somente com o método *transform()*. Visou-se que o modelo não fosse afetado por dados do conjunto de teste durante o processo de aprendizado (BISHOP, 2021; KUHN; JOHNSON, 2013). Utilizou-se também a técnica de validação cruzada estratificada (*Cross Validation Stratified K-Fold*), garantindo que a distribuição da variável alvo fosse mantida em cada partição (KUHN; JOHNSON, 2013; BROWNLEE, 2020).

4.2 Avaliação dos modelos

O ROC-AUC score foi a principal métrica utilizada para avaliar os modelos, visto que a acurácia poderia ser enganosa em situações em que as classes estão desbalanceadas. O ROC-AUC possibilitou a análise da habilidade do modelo em diferenciar entre as classes, independente da sua distribuição (FAWCETT, 2006; HAN; KAMBER; PEI, 2022). Outras métricas somaram-se à avaliação do modelo como, precisão, recall e F1-score.

Durante a modelagem, foram empregadas técnicas de seleção de características e ajuste de hiperparâmetros visando melhorar o desempenho dos modelos. Os ajustes foram conduzidos com métodos sistemáticos de busca pelos hiperparâmetros ideais, aprimorando os resultados conforme a métrica de interesse (FRANCESCHI et al., 2025). Para tanto, utilizou-se a biblioteca Optuna versão 4.5.0 em Python para a otimização dos hiperparâmetros do modelo.

4.3 Comparação dos modelos de aprendizado de máquina

Após o treinamento dos modelos de aprendizado de máquina, procedeu-se à fase de teste com o objetivo de avaliar o desempenho geral de cada algoritmo na predição da evasão escolar. Nessa etapa, foi utilizado um conjunto de teste separado, composto exclusivamente por dados não utilizados durante o treinamento, assegurando uma avaliação imparcial e realista da capacidade de generalização dos modelos (MURPHY, 2012). Foram avaliados nove algoritmos: CatBoost, XGBoost, LightGBM, Support Vector Machine (SVM), Random Forest, Linear SVC, K-Nearest Neighbors (KNN), Regressão Logística e Árvore de Decisão.

Os resultados alcançados permitiram uma comparação objetiva entre os algoritmos, apoiando a seleção do modelo mais apropriado para uso em produção.

A tabela 2 exibe as pontuações de desempenho dos nove classificadores, levando em conta as métricas de acurácia, precisão, recall, F1-score e ROC-AUC com ajustes de hiperparâmetros.

Tabela 2 - Desempenho dos modelos de aprendizado de máquina aplicando hiperparâmetros

Modelo	**Acurácia**	**Precisão**	**Recall**	**F1-score**	**ROC-AUC**
CatBoost	0,72	0,65	0,69	0,67	0,78
XGBoost	0,71	0,63	0,69	0,66	0,71
LightGBM	0,70	0,63	0,69	0,66	0,70
Linear SVC	0,67	0,60	0,62	0,61	0,66
SVM	0,68	0,64	0,52	0,57	0,66
Random Forest	0,67	0,61	0,59	0,60	0,66
KNN	0,66	0,60	0,55	0,57	0,65
Logistic Regression	0,66	0,59	0,63	0,61	0,66
Decision Tree	0,65	0,57	0,60	0,59	0,54

Fonte: elaborado pelo autor com biblioteca Python, 2025.

A análise do desempenho comparativo dos algoritmos destacou os modelos CatBoost, XGBoost e LightGBM os quais obtiveram os melhores resultados, particularmente na métrica ROC-AUC. O algoritmo também se destacou na métrica Recall (0,69). Os resultados evidenciam que a otimização de hiperparâmetros contribuiu para o aprimoramento dos modelos, melhorando sua capacidade preditiva considerando o cenário de desbalanceamento de classes (BERNARDI et al., 2021; HAN; KAMBER; PEI, 2022; KUHN; JOHNSON, 2019; HUTTER; HOOS; LEYTON-BROWN, 2019; RASCHKA; MIRJALILI, 2022).

Após a aplicação da técnica de balanceamento de classes SMOTE *(Synthetic Minority Over-sampling Technique*) (CHAWLA et al., 2002) os modelos CatBoost, XGBoost e LightGBM mantiveram-se como os de melhor desempenho. O modelo CatBoost destacou-se com um ROC-AUC de 0,78 e uma acurácia de 0,72, sendo consistentes em várias métricas, o que demonstrou a estabilidade após o balanceamento das classes.

Durante a fase de teste dos algoritmos, os modelos CatBoost, XGBoost e LightGBM se destacaram ao apresentarem os melhores desempenhos em diversas métricas, mesmo com a utilização da técnica de balanceamento SMOTE. Entretanto, em situações específicas, como a predição de evasão escolar, a seleção das métricas exige atenção, pois a acurácia não deve ser utilizada como a principal medida de avaliação, uma vez que pode esconder erros relevantes, especialmente em casos de classes desbalanceadas (SOKOLOVA; LAPALME, 2009).

A métrica ROC-AUC provou ser uma ferramenta eficiente para analisar a habilidade discriminatória dos modelos, particularmente no caso do CatBoost, que obteve o valor mais alto (0,78) (HASTIE; TIBSHIRANI; FRIEDMAN, 2009). Ao considerar a evasão escolar, a seleção do modelo mais adequado deve dar prioridade àqueles que apresentam também os melhores resultados em termos de recall e F1-score, uma vez que essas métricas estão diretamente ligadas ao objetivo de identificar os alunos com maior risco de evasão (DAVIS; GOADRICH, 2006; FAWCETT, 2006).

4.4 Desempenho dos modelos

O desempenho dos modelos foi avaliado com base nos resultados obtidos por meio da validação cruzada, considerando os valores médios dos *scores* de validação (*val score*) e de treinamento (*train score*). O valor médio de validação, que reflete o desempenho do modelo durante os diversos *folds* da validação cruzada, constitui um indicador relevante da sua capacidade de generalização para dados não vistos (JAMES et al., 2021).

Por outro lado, o escore de treinamento permitiu a avaliação da ocorrência de *overfitting*. Quando o valor do *train score* é muito maior que o valor do *val score*, isso indica que o modelo está se ajustando demais aos dados de treinamento, o que leva a uma baixa capacidade de generalização (RASCHKA; MIRJALILI, 2022).

A figura 5 a seguir mostra o desempenho dos classificadores, com hiperparâmetros ajustados, avaliados de acordo com essas métricas.

Figura 5 - Desempenho do treinamento dos modelos com cross validation e com ajustes de hiperparâmetros

Fonte: elaborado pelo autor, 2025.

Os modelos CatBoost, XGBoost e LightGBM apresentam um bom equilíbrio entre os *scores* de treino e validação, com destaque para o CatBoost, com valores de 0,80 para o treino e 0,78 para a validação. Esse resultado sugere que os algoritmos baseados em métodos de *boosting* apresentaram uma maior habilidade de generalização, mesmo diante de um problema de classificação binária com leve desbalanceamento de classes.

4.5 CatBoostClassifier

Diante do desempenho superior apresentado, o algoritmo CatBoost foi escolhido para prever a evasão escolar. Além de demonstrar a maior habilidade em identificar corretamente os casos de evasão (classe positiva), o modelo também conseguiu manter um equilíbrio satisfatório entre precisão e sensibilidade.

Uma vantagem decisiva do CatBoost é sua capacidade nativa de lidar com variáveis categóricas, tornando-o particularmente vantajoso em conjunto de dados que possuem uma grande presença desse tipo de atributo. Essa característica eliminou a necessidade de aplicar técnicas tradicionais de pré-processamento, como *Label Encoding* ou *One-Hot Encoding*, já que o algoritmo realiza o tratamento interno dessas variáveis de forma eficiente. Com isso, a complexidade do pipeline de preparação dos dados foi reduzida, minimizando o risco de perda

de informação e evitando a alta dimensionalidade gerada pela codificação *One-Hot* (CATBOOST, 2025; PROKHORENKOVA et al., 2018).

O CatBoost foi ajustado com os hiperparâmetros auto_class_weights='Balanced', para compensar o leve desbalanceamento da variável alvo (evadido), e eval_metric='AUC', priorizando a capacidade discriminativa entre os alunos que irão ou não evadir.

Tabela 3 - Métricas de avaliação do modelo CatBoost

Classe	**Precisão**	**Recall**	**F1-score**	**Suporte**
0 (Não evadido)	0,77	0,74	0,75	17.102
1 (Evadido)	0,65	0,69	0,67	12.049
Acurácia geral			0,72	29.151
Média macro	0,71	0,71	0,71	
Média ponderada	0,72	0,72	0,72	

Fonte: elaborado pelo autor com biblioteca Python, 2025.

Nota-se na tabela 3 que o modelo alcançou uma acurácia geral de 72%, o que demonstra um bom desempenho na classificação correta dos dados. Para a classe 0 (não evadido), o modelo obteve uma precisão de 77% e um recall de 74%, gerando um F1-score de 0,75. No que diz respeito à classe 1 (evadido), os resultados mostraram 65% de precisão, 69% de recall e F1-score de 0,67, indicando um desempenho aceitável na detecção de alunos evadidos.

A consistência observada nos resultados das métricas, resultado de um rigoroso processo de validação cruzada e ajustes de hiperparâmetros, reforça a robustez do modelo. Os resultados não somente confirmam a adequação do CatBoost para prever a probabilidade de evasão escolar, como também validam a sua utilização como uma ferramenta analítica capaz de gerar percepções relevantes para fundamentar estratégias proativas de retenção estudantil.

4.6 Implementação da plataforma PrevIA

Após a conclusão da análise e pré-processamento dos dados da evasão escolar na RFEPCT, bem como da modelagem computacional, com a escolha do modelo CatBoostClassifier, com uso de variáveis categóricas, iniciou-se à implementação da ferramenta computacional em ambiente web. Essa etapa permitiu que a aplicação fosse disponibilizada para uso real por gestores e profissionais técnicos envolvidos no acompanhamento dos estudantes, de maneira acessível, interativa e prática em um ambiente de

produção, promovendo o uso da inteligência artificial como instrumento de apoio à tomada de decisão.

É importante ressaltar que o método *predict_proba*, presente na biblioteca *Scikit-learn*, foi empregado. Esse método gera uma matriz que apresenta as probabilidades de cada instância pertencer a cada uma das classes possíveis, possibilitando uma estimativa contínua da probabilidade de um resultado específico (PEDREGOSA et al., 2011). Ao contrário da predição binária convencional, o *predict_proba* permite uma avaliação mais detalhada, sendo particularmente útil em situações em que o nível de confiança da predição é tão importante quanto a classificação em si.

A implementação da plataforma na web foi dividida em três componentes principais: o *back-end*, responsável por carregar o modelo e realizar a predição; o *front-end*, formado pela interface do usuário, com os formulários e visualizações interativas dos resultados; e os serviços de implantação e hospedagem, utilizando Git e GitHub para integração e Streamlit Cloud para a publicação.

Após a conclusão do treinamento do modelo CatBoost, ele foi serializado e salvo no formato Pickle (.pkl), que é a um método de serialização de objetos utilizado em Python. O módulo pickle permite a conversão de estruturas de dados, como modelos treinados, em um fluxo de bytes, possibilitando seu armazenamento e posterior reutilização sem a necessidade de novo treinamento (PYTHON SOFTWARE FOUNDATION, 2025). Isso é útil em ambientes de produção, onde o modelo pode ser carregado rapidamente e utilizado para realizar previsões em tempo real ou em sistemas integrados.

No desenvolvimento de aplicações de aprendizado de máquina, é comum salvar modelos treinados em arquivos .pkl, especialmente ao implementar uma etapa de *deploy* para que o modelo possa ser acessado por usuários finais por meio de interfaces, como plataformas online (THOKALA, 2021).

Quanto ao *front-end*, a aplicação web foi criada com a biblioteca Streamlit versão 1.47.0 em Python, por conta de sua facilidade e eficácia na criação de interfaces para aplicações voltadas ao aprendizado de máquina (SARANGPURE et al., 2023; STREAMLIT COMMUNITY, 2025).

O Streamlit é um *framework* de código aberto que permite o desenvolvimento rápido e interativo de aplicações web focadas em ciência de dados e aprendizado de máquina. Seu diferencial está na facilidade de uso, exigindo conhecimentos intermediários de desenvolvimento web, o que o torna uma ferramenta acessível para cientistas de dados, pesquisadores e profissionais (KHORASANI; ABDOU; FERNÁNDEZ, 2022).

A aplicação web foi hospedada na plataforma Streamlit Cloud e a codificação está acessível publicamente no repositório GitHub no endereço: https://github.com/jabsoncd/PrevIA-Predicao-Evasao-Rede-Federal, promovendo assim, a transparência e a colaboração aberta. A plataforma PrevIA pode ser acessada em seu ambiente de produção por meio do link https://previa-beta.streamlit.app/.

Optou-se pelo Streamlit Cloud devido à sua simplicidade no processo de *deploy*, uma vez que permite a implantação de aplicações diretamente de repositórios GitHub, sem a necessidade de configurar servidores, containers ou pipelines complexos.

A definição pela ferramenta GitHub ocorreu porque, ao tornar o código público, atende-se ao princípio da ciência aberta, o que possibilita que outros pesquisadores validem, melhorem ou reproduzam o trabalho (CHEN; TORO-MORENO; SUBRAMANIAM, 2024). Além disso, o repositório GitHub se conecta de forma nativa ao Streamlit Cloud, o que permite automatizar o *deploy* a partir do código versionado.

A aplicação web foi disponibilizada com duas funcionalidades principais: um simulador de evasão e indicadores de evasão. A aba simulador de evasão permite a entrada de novos registros de estudantes via formulário, gerando predições probabilísticas de risco de evasão com base no modelo CatBoost treinado e salvo. Na aba de indicadores de evasão, os usuários podem explorar variáveis e representações visuais da evasão escolar na RFEPCT em diversos níveis de agregação (regiões, unidades da federação, instituições, grupos sociais e cursos), utilizando gráficos dinâmicos e mapa interativo.

Com essas funcionalidades gestores e equipes pedagógicas podem avaliar cenários alternativos e planejar intervenções direcionadas. A combinação dessas duas dimensões, monitoramento por indicadores e simulação preditiva dos estudantes, define a plataforma PrevIA como uma ferramenta prática que auxilia na tomada de decisão, com o objetivo de reduzir a evasão nos cursos técnicos da RFEPCT.

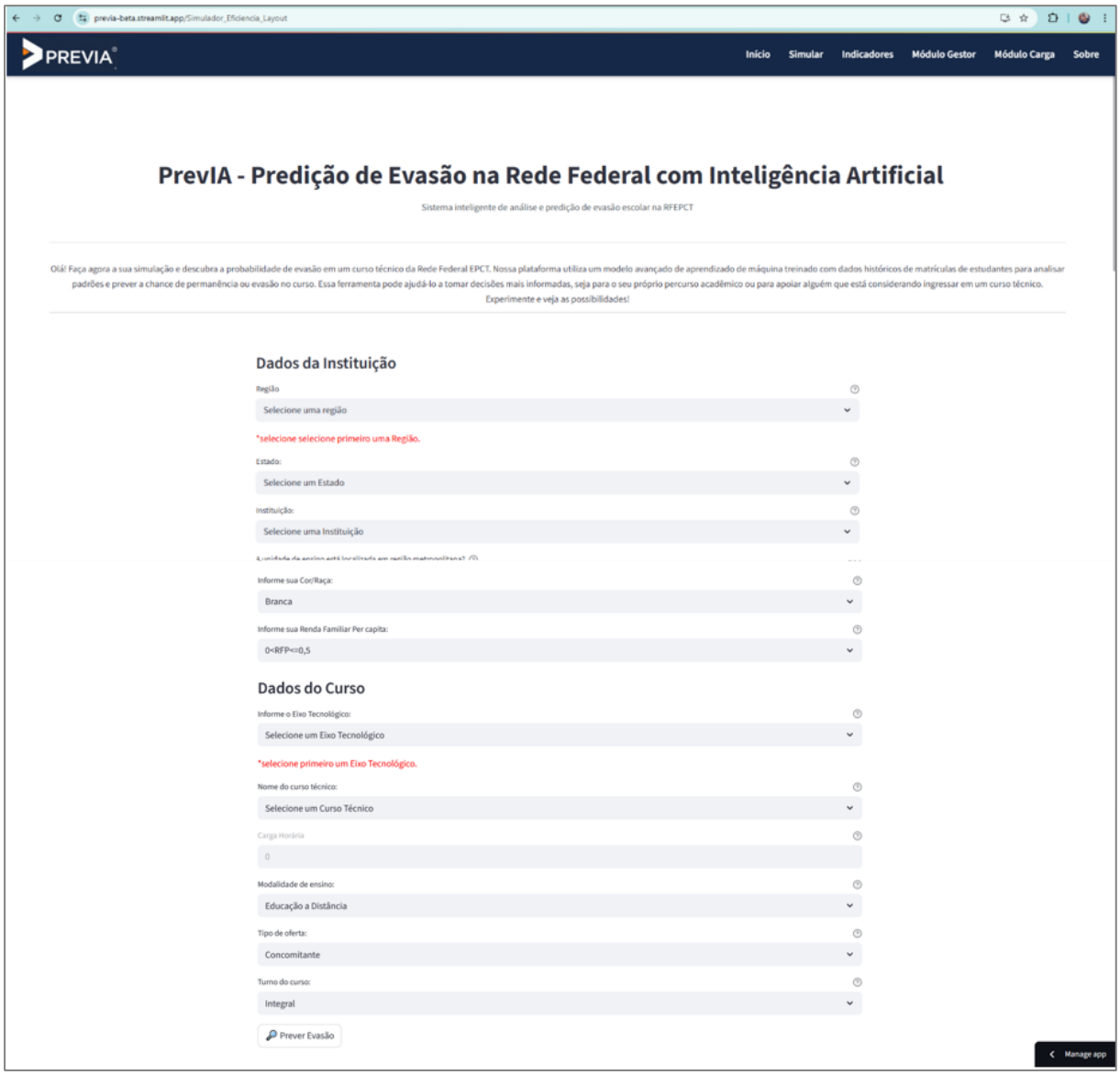
A plataforma web foi organizada com duas funcionalidades principais conforme figura 6:

Figura 6 - Tela inicial da plataforma PrevIA

Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

A primeira funcionalidade é o simulador de evasão: esta página apresenta um formulário que recebe dados relevantes do estudante, mantendo as variáveis contidas no conjunto de dados treinados. As informações fornecidas dizem respeito à localização da instituição de ensino (região, estado, instituição, se está localizada em região metropolitana), às características pessoais do estudante (idade, cor/raça, gênero, renda familiar) e às características do curso desejado (nome, eixo tecnológico, carga horária mínima, modalidade de ensino, tipo de oferta e turno).

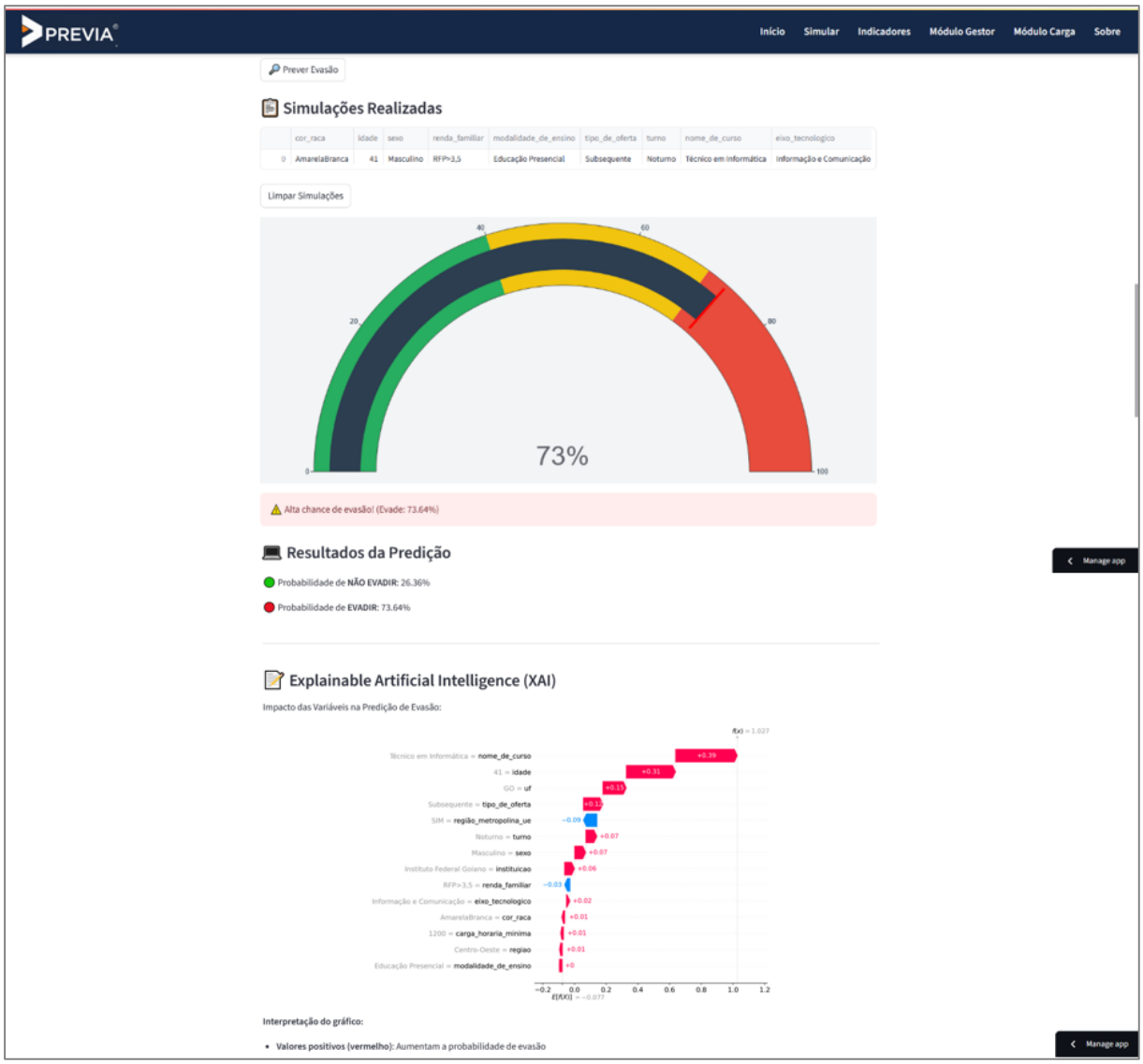
Figura 7 - Tela com o formulário para simulação da probabilidade de evasão – Plataforma PrevIA



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

Após o preenchimento e ao clicar no botão Prever Evasão, as informações são processadas pelo modelo CatBoostClassifier, salvo em formato (.pkl), e a probabilidade de evasão do estudante é retornada na tela. Um gráfico radial *(radial gauge*), exibe o resultado, fornecendo uma representação visual do percentual estimado de evasão.

Figura 8 - Resultado gerado pelo simulador de probabilidade de evasão – Plataforma PrevIA



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

Com base no valor percentual de probabilidade de evasão, o resultado da predição classifica as seguintes categorias de risco de evasão de acordo com a regra estabelecida no código:

Tabela 4 - Categorias de risco de evasão escolar na plataforma PrevIA

Categoria de Risco de Evasão		**Situação do Estudante**
Risco baixo	Menor que 50%	Estudante com perfil de permanência
Risco moderado	Entre 51% e 60%	Estudante com incertezas
Risco considerável	Entre 61% e 70%	Estudante com sinais de evasão
Risco alto	Entre 71% e 90%	Estudante com forte tendência à evasão
Risco muito alto	Acima de 90%	Estudante prestes a evadir

Fonte: elaborado pelo autor, 2025.

É importante destacar que, ao inserir novos dados de um aluno para simular a probabilidade de evasão, a plataforma não armazena essas informações nem as utiliza para treinar o modelo de aprendizado de máquina.

A segunda funcionalidade diz respeito aos indicadores de evasão: essa página exibe informações consolidadas sobre a evasão na RFEPCT, com base nos microdados divulgados referente ao ano de 2023. São exibidos quatro *cards* de indicadores: demográficos, sociais, cursos e o mapa da evasão. Em cada categoria são apresentados diversos indicadores. A página oferece filtros para que o usuário escolha as variáveis desejadas.

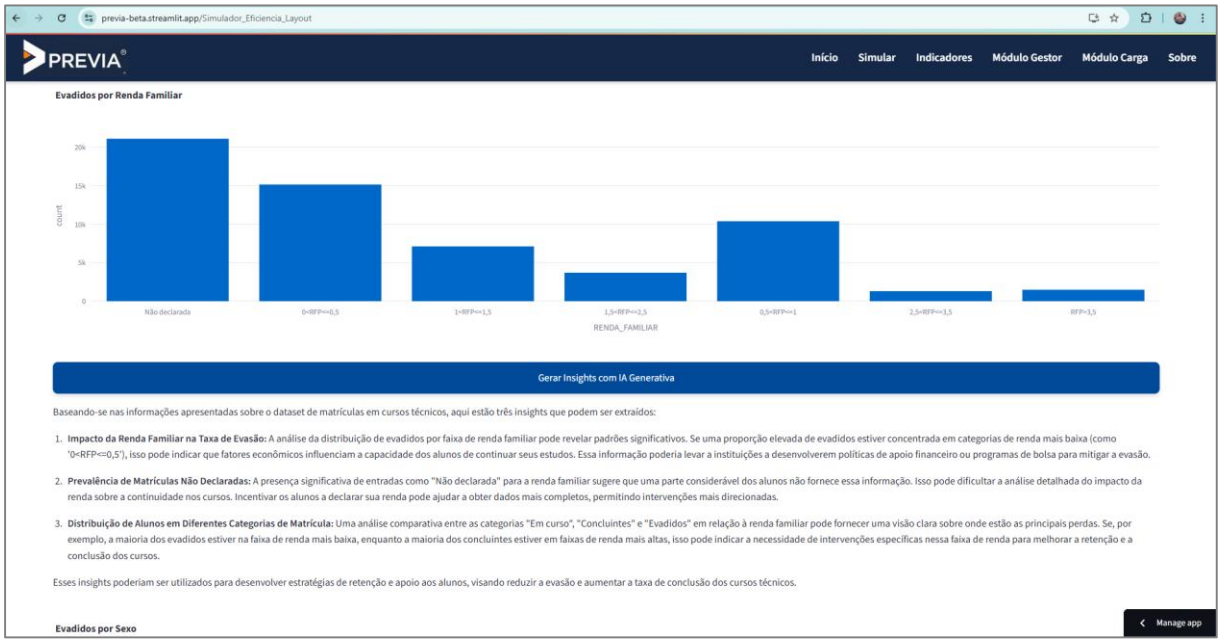
Figura 9 - Tela com indicadores de evasão – Plataforma PrevIA

Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

Cada gráfico conta com um botão chamado Gerar *Insights* com IA Generativa que, ao ser pressionado, possibilita a análise automatizada do conteúdo utilizando a inteligência

artificial GPT-4o-mini, criada pela OpenAI. O modelo fornece três *insights* interpretativos a respeito dos dados exibidos no gráfico escolhido.

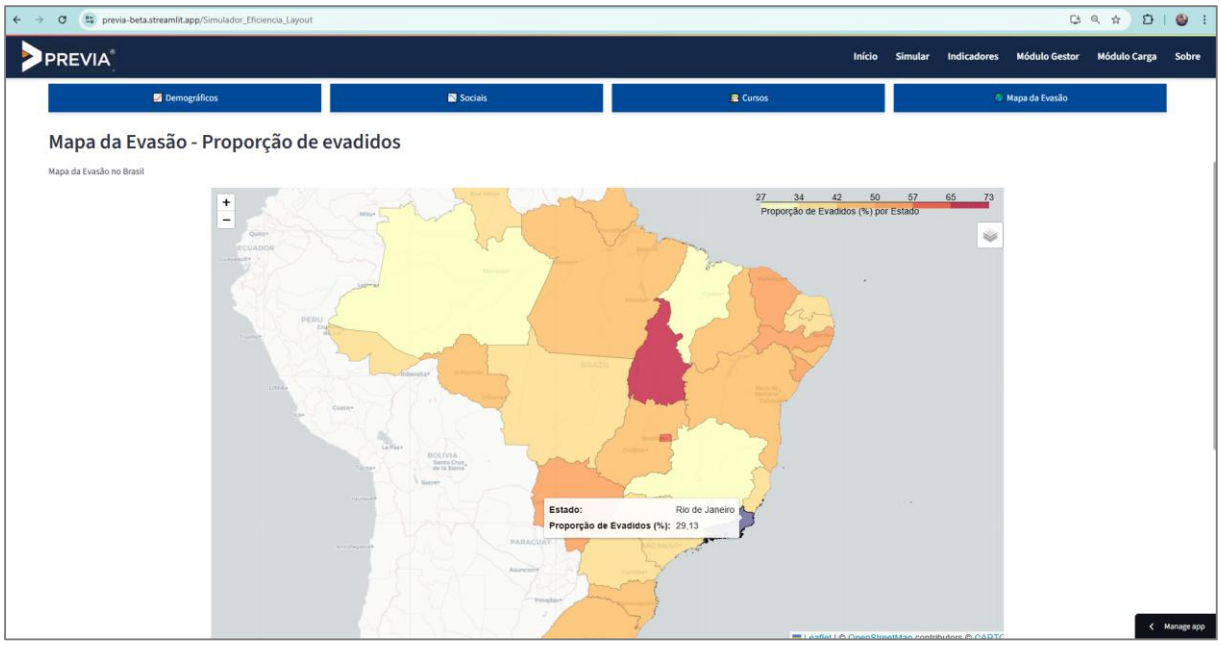
Figura 10 - *Insights* gerados pelo modelo GPT-4o-mini da OpenAI – Plataforma PrevIA



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

O *card* mapa da evasão exibe um mapa de calor dinâmico da evasão no Brasil, considerando a proporção de evadidos de cada estado. A estrutura utilizou as bibliotecas Folium (versão 0.20.0) e GeoPandas (versão 1.1.1). Ao posicionar o cursor sobre a área territorial do Estado é exibido o valor percentual de evadidos.

Figura 11 - Mapa dinâmico de calor com a proporção de evadidos por Estados – Plataforma PrevIA



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

5 DISCUSSÃO

5.1 Estratégias para mitigar a evasão na Rede Federal

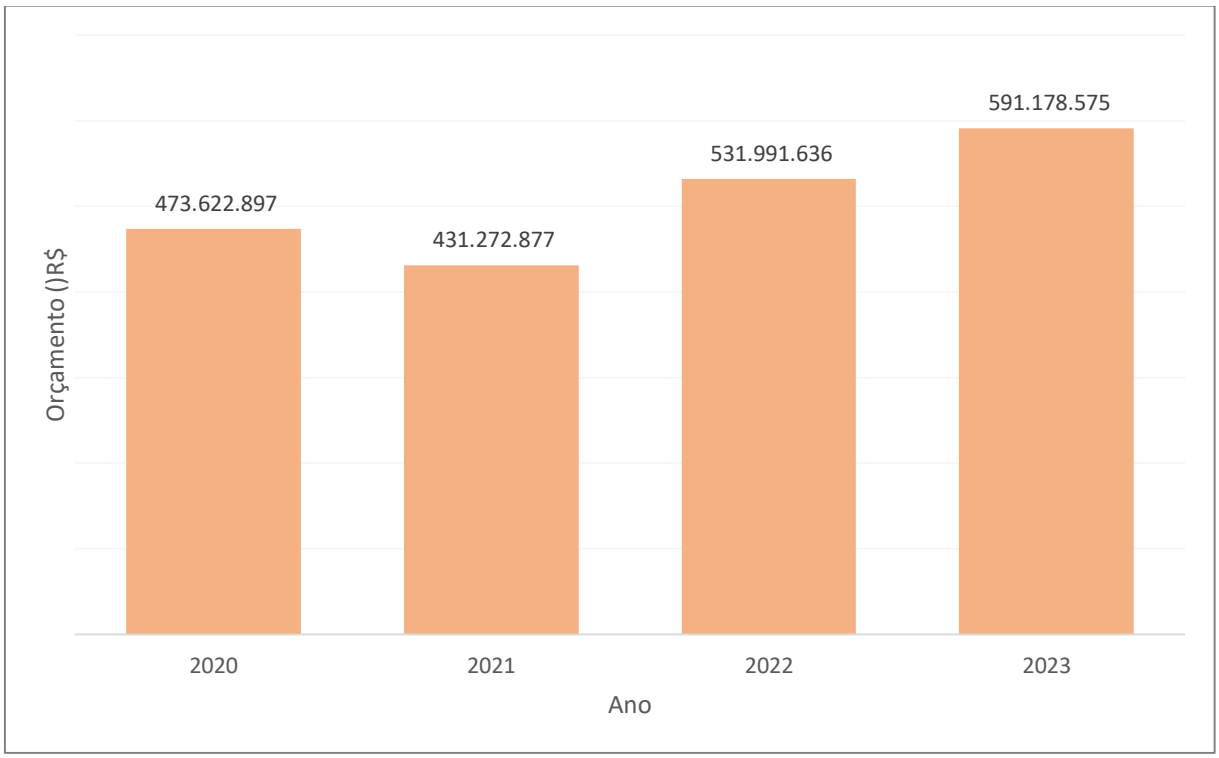
A evasão escolar na Rede Federal de Educação Profissional, Científica e Tecnológica (RFEPCT) representa um dos desafios mais significativos à efetividade das políticas públicas educacionais no Brasil. A análise dos dados por categorias explicativas revela que a evasão não é um fenômeno homogêneo, mas complexo, sendo afetada por fatores socioeconômicos, demográficos, pedagógicos e territoriais (BONALDO; PEREIRA, 2016; BRASIL; 2024).

O índice de evasão apresentado no estudo (41,3%) é significativo e reflete um cenário de fragilidade na trajetória educacional de expressiva parcela de estudantes da Rede Federal. Segundo levantamento feito pelo TCU, os principais obstáculos à permanência e êxito dos estudantes incluem a insuficiência de recursos para assistência estudantil e alimentação escolar, carência de equipes multiprofissionais, deficiências de infraestrutura e ausência de políticas consolidadas de transporte escolar (BRASIL, 2024b).

A assistência estudantil para a Rede Federal, contemplada pela ação orçamentária 2994 do orçamento público federal, surge como uma política estratégica nesse contexto, uma vez que busca mitigar desigualdades de origem e promover a equidade no acesso e permanência (BRASIL, 2025a). A ampliação do orçamento nos últimos anos é um indicativo positivo, mas

ainda insuficiente diante das complexidades que afetam os estudantes da RFEPCT (BRASIL, 2023a).

Figura 12 - Orçamento federal para assistência estudantil na Rede Federal de EPCT período de 2020 a 2023



Fonte: elaborado pelo autor, dados extraídos do SIOP, 2025.

Esse orçamento constitui uma das principais políticas públicas voltadas à promoção da permanência e do êxito dos estudantes em situação de vulnerabilidade socioeconômica. O objetivo é garantir que os Institutos Federal tenham condições mínimas que possibilitem aos estudantes concluírem seus cursos com qualidade (BRASIL, 2023a).

A ampliação do orçamento demonstra um reconhecimento da importância da política de assistência estudantil no enfrentamento à evasão escolar, na luta contra a desigualdade educacional e na garantia da equidade no acesso à educação (BRASIL, 2023).

A Política Nacional de Assistência Estudantil (PNAES), estabelecida pela Lei nº 14.914, de 3 de julho de 2024, representa mais uma estratégia do governo federal para assegurar a permanência e êxito dos estudantes nas instituições federais de educação superior e de educação profissional e tecnológica. Um dos principais objetivos da PNAES é a redução das taxas de retenção e evasão na educação pública federal (BRASIL, 2024c).

Mais recentemente, o Programa Pé-de-Meia, instituído pela Lei nº 14.818, de 16 de janeiro de 2024, surge como uma política de incentivo financeiro-educacional direcionada a

estudantes do ensino médio e da Educação de Jovens e Adultos (EJA) matriculados em escolas públicas, incluindo a RFEPCT. Seu objetivo é incentivar a permanência e a conclusão dos estudos, principalmente entre os jovens em maior situação de vulnerabilidade (BRASIL, 2024d).

Além dos desafios estruturais enfrentados em âmbito nacional, as instituições que compõem a RFEPCT têm implementado políticas institucionais com o objetivo de reduzir a evasão escolar. Essas ações demonstram a capacidade dessa rede em articular ações tanto de forma independente quanto em colaboração com políticas federais.

Há ações de domínio interno das instituições entre as estratégias de enfrentamento à evasão. Dentre elas, intervenções curriculares, como a adequação dos currículos às diretrizes nacionais e à realidade local, iniciativas de reforço de conteúdos, aprimoramento da comunicação institucional, incentivo às atividades esportivas, artísticas e culturais, acompanhamento acadêmico e de frequência, além de práticas de acolhimento e fortalecimento do ambiente escolar. A busca ativa de estudantes que abandonam a escola também são ferramentas comuns de prevenção e recuperação das trajetórias escolares.

Já as ações em alinhamento com as políticas federais abrangem infraestrutura, suporte ao estudante e intermediação com o mundo do trabalho. Medidas como a concessão de bolsas de estudo (para ensino, pesquisa e extensão), disponibilização de refeitórios para estudantes, bibliotecas e programas de transporte escolar são consideradas ações fundamentais para diminuir as desigualdades de acesso e promover a equidade na educação.

A elaboração de Plano Estratégico de Permanência e Êxito Estudantil por diversas instituições da RFEPCT, tem se mostrado um instrumento importante para identificar as principais causas para a evasão, considerando as seguintes diversões: individuais, internas à instituição e externas (KOCSIS; MOLNÁR, 2024).

Essas ações reforçam a compreensão de que a evasão na Rede Federal de EPCT não é tratada de maneira isolada, mas por meio de estratégia sistêmica que inclui políticas públicas estruturantes, fortalecimento institucional e escuta ativa dos indivíduos envolvidos no processo educacional. Portanto, há a compreensão que a redução da evasão exige diagnóstico contínuo, intervenção qualificada e articulação entre setores para garantir que o direito à educação seja, de fato, plenamente efetivado.

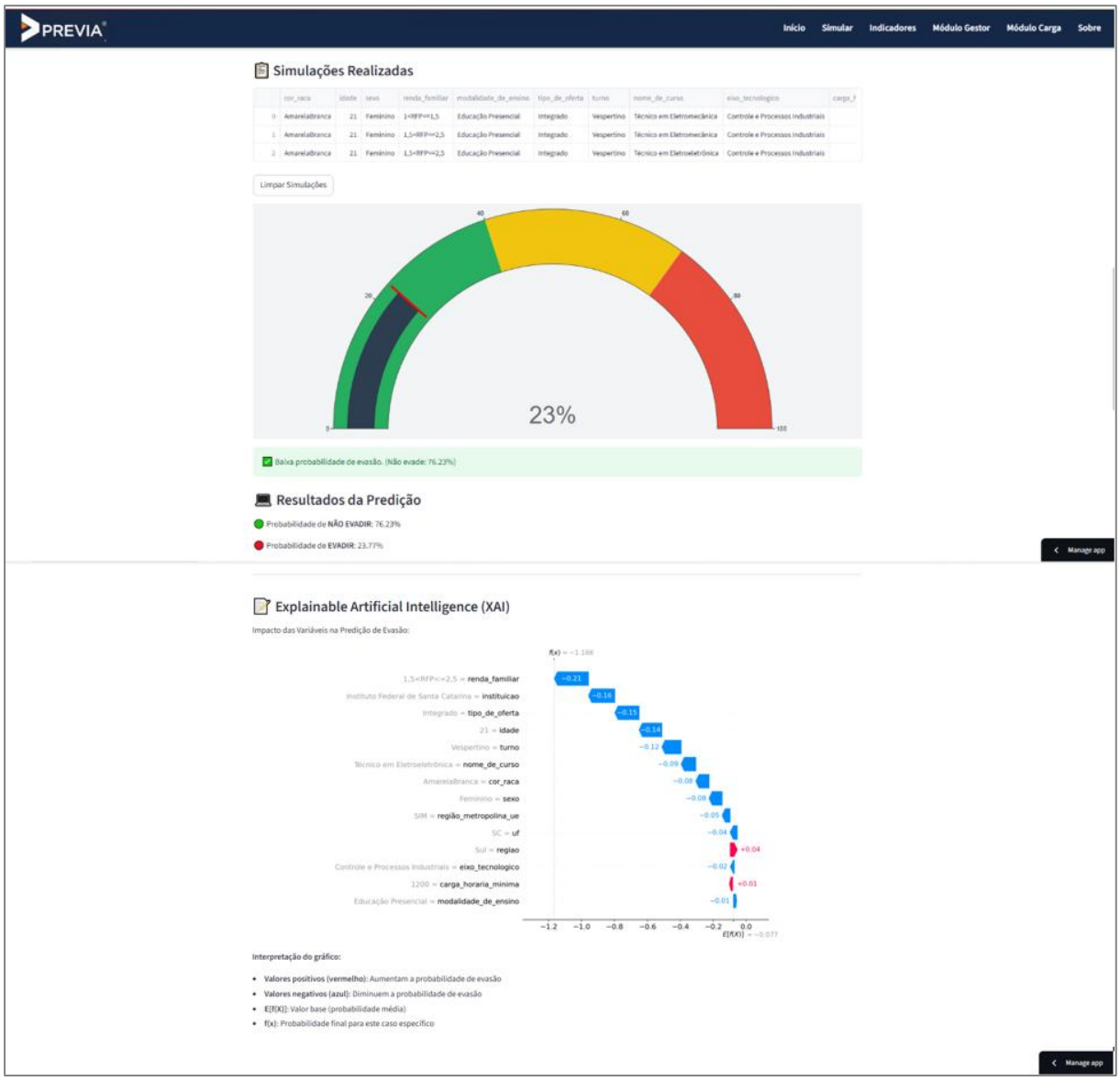
5.2 A contribuição da plataforma PrevIA para mitigar a evasão em cursos técnicos

Conforme apontamento do TCU, além de restrições materiais que impactam a redução da evasão escolar, existem questões gerenciais importantes a serem atendidas nos planos estratégicos das instituições da RFEPCT. Entre essas questões gerenciais, destacam-se a dificuldade de acesso e integração dos dados acadêmicos relacionados à evasão, a falta de indicadores desagregados por cor/raça e renda familiar, a ausência de identificação e classificação dos fatores que mais influenciam a evasão, bem como a inexistência de metas para melhorar as taxas de conclusão dos cursos técnicos. Essas lacunas comprometem a capacidade das instituições de planejar e executar ações efetivas de intervenção eficazes (BRASIL, 2024b).

A construção de um simulador de probabilidade de evasão escolar em cursos técnicos, considerando as variáveis da plataforma Nilo Peçanha que afetam a evasão, visa auxiliar tanto as ações gerenciais quanto às demandas práticas na identificação precoce de novos estudantes que ingressam nas instituições da RFEPCT (SANDOVAL-PALIS; NARANJO; VIDAL; GILAR-CORBI, 2020).

A figura 13 ilustra a aplicabilidade e a confiabilidade da plataforma desenvolvida, conforme demonstrado pelos resultados das métricas de avaliação do modelo preditivo utilizado neste projeto. Os dados apresentados são semelhantes aos de um estudante que não evadiu do curso técnico. Ao realizar a predição na plataforma PrevIA, o resultado indicou uma probabilidade de evasão de 23%. Isso evidencia um perfil de estudante com baixo risco de evadir do curso técnico.

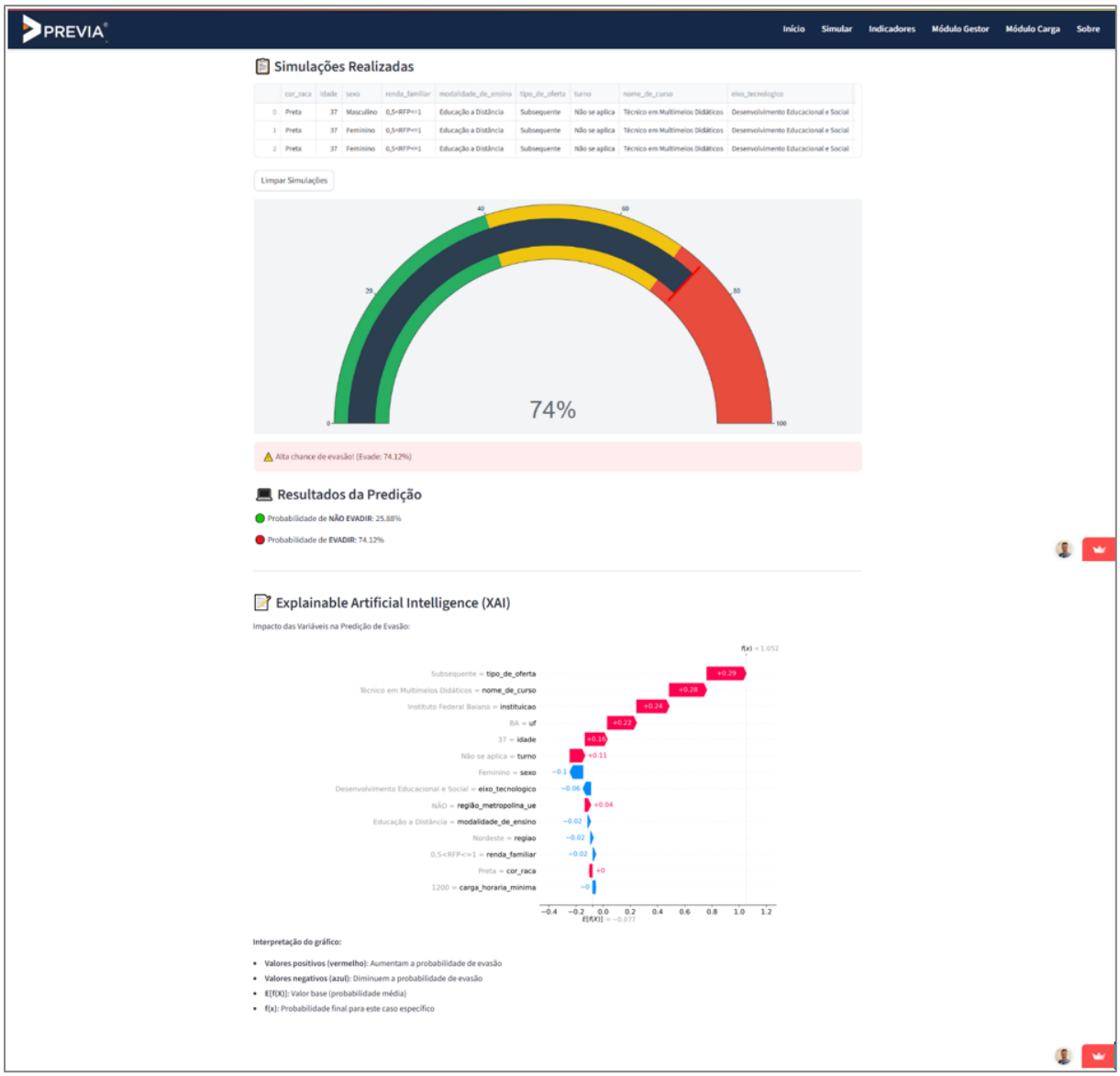
Figura 13 - Tela da plataforma PrevIA com o percentual de probabilidade de aluno que não evadiu no conjunto de dados de teste



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

Em outra simulação foi realizado o procedimento oposto. A figura 14 mostra a inserção de dados semelhantes de um estudante que evadiu do curso técnico. Ao realizar a predição, o resultado indicou uma probabilidade de evasão de 74%. O que evidencia um estudante com alta probabilidade de abandonar o curso técnico.

Figura 14 - Tela da plataforma PrevIA com o percentual de probabilidade de aluno que evadiu no conjunto de dados de teste



Fonte: elaborado pelo autor, tela plataforma PrevIA, 2025.

A capacidade do modelo preditivo de atribuir percentuais de risco de evasão a cada estudante, permitindo uma análise mais detalhada do que uma simples classificação binária (evadido/não evadido) representa um diferencial que possibilita uma gestão mais eficaz e orientada por dados. De forma, os recursos e estratégias da instituição podem ser direcionados prioritariamente aos alunos com maior probabilidade de evasão (CANO; LEONARD, 2019; KORKMAZ; AYDIN, 2025).

A definição de percentuais de probabilidade viabiliza o desenvolvimento de sistemas de alerta capazes de apoiar a tomada de decisão em tempo hábil pelos gestores educacionais (BAKER; SIEMENS, 2014). Portanto, o uso de probabilidades contínuas possibilita a definição de limites personalizados para as intervenções pedagógicas. Dessa forma, estudantes com risco de evasão superior a 70%, conforme indicado pela PrevIA, podem ser imediatamente

direcionados à assistência estudantil, enquanto os estudantes com risco entre 40% e 70% podem receber acompanhamento preventivo. Isso promove o uso eficiente dos recursos institucionais e aumenta a eficácia das estratégias de permanência.

A utilização da plataforma pode trazer diversas contribuições, como a diminuição de custos operacionais devido à otimização dos recursos humanos e tecnológicos; o aumento nas taxas de retenção e permanência dos alunos; a atuação proativa dos setores de assistência estudantil; e a realização de estudos mais focados na compreensão das necessidades específicas dos estudantes.

Além disso, as ferramentas web oferecem a vantagem de serem acessíveis e escaláveis, possibilitando que gestores e educadores de diferentes regiões e contextos acessem dados e *insights* gerados. Isso é particularmente relevante em um país com grandes disparidades regionais e socioeconômicas, como o Brasil, em que a centralização de informações pode ajudar a diminuir as desigualdades educacionais.

Na área da EPT, a implementação da plataforma PrevIA representa uma aplicação prática da modelagem computacional, em conformidade com as boas práticas em ciência de dados voltadas para a gestão de políticas públicas. Desenvolver ferramentas de IA para identificar e prevenir a evasão escolar é essencial para modernizar a administração educacional e promover implementações mais eficazes.

É fundamental considerar as limitações decorrentes do conjunto de variáveis utilizadas, restritas àquelas disponibilizadas pela PNP. Embora esses dados sejam relevantes, eles não abrangem a complexidade integral do fenômeno de estudo. Diversos outros aspectos podem contribuir para uma compreensão mais ampla da evasão escolar na RFEPCT, como o histórico educacional anterior, a escolha de cursos e disciplinas, bem como fatores relacionados à vida acadêmica, como desempenho, frequência e características pessoais do estudante. A incorporação desses atributos possibilita uma análise mais robusta e abrangente, capaz de revelar aspectos que os dados atualmente disponíveis não contemplam.

6 CONCLUSÃO, LIMITAÇÕES E TRABALHOS FUTUROS

A evasão nos cursos técnicos deve ser compreendida como um fenômeno social que demanda estratégias articuladas de prevenção e acompanhamento, sustentadas por ações institucionais e políticas públicas que incentivem a permanência escolar. O uso da inteligência artificial e a implementação de sistemas preditivos de evasão configuram-se como aliados

estratégicos capazes de antecipar riscos e apoiar a tomada de decisões voltadas à redução desse problema.

Este artigo descreveu as etapas de análise e preparação dos dados, correspondentes a uma das fases essenciais da realização da modelagem computacional para realizar aprendizado de máquina. Posteriormente, na etapa de implementação do modelo, foi desenvolvida uma plataforma web utilizando o modelo preditivo que apresentou melhor desempenho, o CatBoost. Com esse modelo foi possível simular a probabilidade de um estudante evadir de um curso técnico na RFEPCT. A execução desses procedimentos exigiu uma execução criteriosa, especialmente diante do grande volume de dados envolvidos. Devido a essa complexidade, a metodologia proposta neste estudo buscou a aplicabilidade em um projeto real de aprendizado de máquina.

A principal contribuição deste estudo é a implementação de uma metodologia baseada em técnicas de aprendizado de máquina para prever a evasão em cursos técnicos da RFEPCT. O objetivo foi criar uma plataforma web que permite simular, de maneira acessível e interativa, a probabilidade de um estudante abandonar os estudos, utilizando dados sociodemográficos, individuais e características do curso pretendido.

Como limitação, cabe informar que o conjunto de dados de treinamento utilizado refere-se unicamente às matrículas em situação de finalização no ano de 2023, divulgados em 2024, e apresenta um número limitado de variáveis disponibilizadas para o acompanhamento e a gestão da RFEPCT. Com isso, é preciso admitir que o uso exclusivo apenas da plataforma não é suficiente para analisar os diversos fatores que levam a evasão em cursos técnicos.

Trabalhos futuros podem ampliar a robustez dos modelos preditivos com a utilização de diferentes técnicas de aprendizado de máquina, bem como outros parâmetros. Além disso, a adição de outras variáveis como frequências, notas, rendimento e participações em projetos, pode contribuir para o melhor entendimento dos fatores relacionados à evasão escolar. Ademais, a plataforma poderá receber novas funcionalidades para gerar informações relevantes sobre os padrões existentes de evasão na RFEPCT. Por exemplo, simulação em lote de estudantes por curso. A coleta de *feedback* de equipes de registros acadêmicos das instituições da RFEPCT quanto à eficácia e facilidade de uso da ferramenta poderá agregar melhorias à plataforma.

A plataforma tecnológica PrevIA apresenta um recurso inovador no enfrentamento da evasão escolar, ao incorporar a IA como componente central para apoiar políticas e ações institucionais mais eficientes, ao fortalecer a gestão orientada a dados e permitir a detecção precoce de estudantes em risco. Ao revelar padrões de evasão, o uso dessa ferramenta amplia a

capacidade de intervenção das instituições principalmente na orientação de ações de permanência e êxito estudantil. Em última análise, para superar a evasão é necessário um processo permanente de diagnóstico, intervenções qualificadas e colaboração entre setores, sendo a IA um aliado estratégico para garantir o direito à educação.

Declaração de IA Generativa e tecnologias assistidas por IA em processo de escrita

Durante a preparação deste trabalho, o autor utilizou o ChatGPT-4 e DeepSeek Latest Version para melhorar a legibilidade e a linguagem. Após o uso desta ferramenta, os autores revisaram e editaram o conteúdo conforme necessário e assumem total responsabilidade pelo conteúdo da publicação.

REFERÊNCIAS

AKIBA, Takuya; SANO, Shotaro; YANASE, Toshihiko; OHTA, Takeru; KOYAMA, Masanori. Optuna: a next-generation hyperparameter optimization framework. In: ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY & DATA MINING (KDD), 25., 2019. Proceedings, p. 2623–2631, 2019. Disponível em: https://doi.org/10.1145/3292500.3330701. Acesso em: 23 jul. 2025.

BAGGI, C. A. S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. Avaliação: Revista da Avaliação da Educação Superior, Campinas, v. 16, n. 2, p. 355-374, jul. 2011.

BAKER, R. S.; SIEMENS, P. S. Educational data mining and learning analytics. In: SAWYER, R. K. (ed.). The Cambridge Handbook of the Learning Sciences. 2. ed. Cambridge: Cambridge University Press, 2014. p. 253–274.

BENTÉJAC, Candice; CSÖRGŐ, Anna; MARTÍNEZ-MUÑOZ, Gonzalo. A comparative analysis of gradient boosting algorithms. Artificial Intelligence Review, v. 54, n. 3, p. 1937–1967, 2021. Disponível em: https://doi.org/10.1007/s10462-020-09896-5. Acesso em: 23 jul. 2025.

BESSEY, D.; BACKES-GELLNER, U. Regional unemployment and educational attainment in vocational training. Economics of Education Review, v. 44, p. 1–18, 2015. Disponível em: https://doi.org/10.1016/j.econedurev.2014.10.003. Acesso em: 29 maio 2025.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2021.

BÖHN, S.; DEUTSCHER, V. Determinants of dropout in vocational education and training: a systematic review. Empirical Research in Vocational Education and Training, v. 14, n. 3, p. 1-25, 2022.

BONALDO, Luciane; PEREIRA, Luis Nobre. Dropout: Demographic profile of Brazilian university students. Procedia - Social and Behavioral Sciences, v. 228, p. 138-143, 2016. ISSN 1877-0428. Disponível em: https://www.sciencedirect.com/science/article/pii/S1877042816309466. DOI: 10.1016/j.sbspro.2016.07.020. Acesso em: 22 jun. 2025.

BRASIL. Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação – PNE e dá outras providências. Diário Oficial da União: seção 1, Brasília, DF, p. 1, 26 jun. 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm. Acesso em: 23 maio 2025.

BRASIL. Lei nº 13.005, de 25 de junho de 2014. Aprova o Plano Nacional de Educação - PNE e dá outras providências. Diário Oficial da União, Brasília, DF, 26 jun. 2014.

BRASIL. Ministério da Educação. Secretaria de Educação Profissional e Tecnológica (SETEC). Relatório anual da educação profissional e tecnológica no Brasil. Brasília, DF: MEC, 2023.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha – PNP. Microdados Eficiência Acadêmica, edição 2023. Brasília, DF: MEC, 2023. Disponível em: https://dadosabertos.mec.gov.br/pnp/item/261-2023-microdados-eficiencia-academica. Acesso em: 14 jun. 2024.

BRASIL. Ministério da Educação. Plataforma Nilo Peçanha (PNP). Brasília, DF: MEC, 2024. Disponível em: https://www.gov.br/mec/pt-br/pnp. Acesso em: 14 jul. 2024.

BRASIL. Portal de Dados Abertos do Governo Federal. Brasília, DF: Governo Federal, 2024. Disponível em: https://dados.gov.br/. Acesso em: 14 jun. 2024.

BRASIL. Instituto Brasileiro de Geografia e Estatística. Recortes Metropolitanos e Aglomerações Urbanas. Brasília, DF, 2024a. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/18354-recortes-metropolitanos-e-aglomeracoes-urbanas.html. Acesso em: 14 jul. 2024.

BRASIL. Presidência da República. Lei nº 11.892, de 29 de dezembro de 2008. Institui a Rede Federal de Educação Profissional, Científica e Tecnológica. Diário Oficial da União: seção 1, Brasília, DF, ano 145, n. 251, p. 1, 30 dez. 2008. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2007-2010/2008/lei/l11892.htm. Acesso em: 31 maio 2025.

BRASIL. Centro de Gestão e Estudos Estratégicos (CGEE). A educação profissional e tecnológica no Brasil: análise e subsídios para políticas públicas. Brasília, DF: CGEE, 2023. Disponível em: https://www.cgee.org.br. Acesso em: 27 maio 2025.

BRASIL. Instituto Brasileiro de Geografia e Estatística (IBGE). Recortes metropolitanos e aglomerações urbanas. Rio de Janeiro: IBGE, [2023b]. Disponível em: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/estrutura-territorial/18354-recortes-metropolitanos-e-aglomeracoes-urbanas.html. Acesso em: 31 maio 2025.

BRASIL. TRIBUNAL DE CONTAS DA UNIÃO. Auditoria operacional na Rede Federal de Educação Profissional, Científica e Tecnológica: relatório de auditoria. Brasília, DF: TCU, 2024.

BRASIL. Tribunal de Contas da União. Acórdão nº 986/2024- Plenário, 22 maio 2024. Brasília, DF: TCU, 2024b. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A986%2520ANOACORDAO%253A2024%2520/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/0. Acesso em: 14 nov. 2024.

BRASIL. CONSELHO NACIONAL DAS INSTITUIÇÕES DA REDE FEDERAL DE EDUCAÇÃO PROFISSIONAL, CIENTÍFICA E TECNOLÓGICA (CONIF). Necessidade de recomposição orçamentária para a Rede Federal de Educação Profissional, Científica e Tecnológica. Brasília, 9 ago. 2023. Disponível em: https://portal.conif.org.br/geral/necessidade-de-recomposicao-orcamentaria-para-a-rede-federal-de-educacao-profissional-cientifica-e-tecnologica. Acesso em: 27 jul. 2025.

BRASIL. Portal da Transparência. Execução da despesa por programa/ação. Ação nº 2994. 2025a. Disponível em: https://portaldatransparencia.gov.br/despesas/programa-e-acao?acao=2994&ordenarPor=programa&direcao=asc. Acesso em: 26 jun. 2025.

BRASIL. Lei nº 14.914, de 3 de julho de 2024. Institui a Política Nacional de Assistência Estudantil (PNAES). 2024c. Disponível em: https://agenciagov.ebc.com.br/noticias/202407/nova-lei-da-pnaes-fortalece-assistencia-e-combate-a-evasao. Acesso em: 10 jun. 2025.

BRASIL. Lei nº 14.818, de 16 de janeiro de 2024. Institui o Programa Pé-de-Meia, incentivo financeiro-educacional para estudantes do ensino médio público. 2024d. Disponível em: https://www.gov.br/mec. Acesso em: 10 jun. 2025.

BREIMAN, L. Random forests. Machine Learning, v. 45, n. 1, p. 5–32, 2001.

BRANCO, P.; TORGOSO, L.; RIBEIRO, R. A Survey of Predictive Modelling under Imbalanced Distributions. ACM Computing Surveys, v. 49, n. 2, p. 1–50, 2016.

BROWNLEE, J. Imbalanced Classification with Python: Better Metrics, Balance Skewed Classes, Cost-Sensitive Learning. Machine Learning Mastery, 2020.

BURGES, Christopher J. C. A tutorial on support vector machines for pattern recognition. Data Mining and Knowledge Discovery, v. 2, n. 2, p. 121–167, 1998.

CANO, A.; LEONARD, J. D. Interpretable multiview early warning system adapted to underrepresented student populations. IEEE Transactions on Learning Technologies, v. 12, n. 2, p. 198–211, abr./jun. 2019. Disponível em: https://doi.org/10.1109/TLT.2019.2911079. Acesso em: 23 jul. 2025.

CATBOOST documentation. Categorical features | CatBoost. Disponível em: https://catboost.ai/docs/en/features/categorical-features. Acesso em: 10 jun. 2025.

CHAWLA, Nitesh V.; BOWYER, Kevin W.; HALL, Lawrence O.; KEGELMEYER, W. Philip. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research, v. 16, p. 321–357, 2002. Disponível em: https://www.jair.org/index.php/jair/article/view/10302. Acesso em: 24 out. 2025.

CHEN, T.; GUESTRIN, C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016.

CHEN, Katharine Y.; TORO-MORENO, Maria; SUBRAMANIAM, Arvind Rasi. GitHub enables collaborative and reproducible laboratory research. [S.l.], 2024. Disponível em: https://pmc.ncbi.nlm.nih.gov/articles/PMC11828340/. Acesso em: 22 jun. 2025.

CHO, C. H.; YU, Y. W.; KIM, H. G. A study on dropout prediction for university students using machine learning. Applied Sciences, v. 13, n. 21, p. 12004, 2023. DOI: https://doi.org/10.3390/app132112004

DAVIS, J.; GOADRICH, M. The Relationship Between Precision-Recall and ROC Curves. In: ICML '06: Proceedings of the 23rd International Conference on Machine Learning. New York: Association for Computing Machinery (ACM), 2006. p. 233–240. DOI: https://doi.org/10.1145/1143844.1143874.

DOMINGOS, P.; PAZZANI, M. On the optimality of the simple Bayesian classifier under zero-one loss. Machine Learning, v. 29, n. 2–3, p. 103–130, 1997.

DOMINGUES, R.; BUENO, T. M.; LENGLER, L.; SANTOS, R. M.; FERREIRA, A. C.; OLIVEIRA, M. R. Data preprocessing techniques for machine learning. Journal of Big Data, v. 9, n. 1, p. 1–26, 2022.

DOROGUSH, Anna Veronika; ERSHOV, Vasily; GULIN, Andrey. CatBoost: gradient boosting with categorical features support. arXiv preprint arXiv:1810.11363 [cs.LG], 2018. Disponível em: https://arxiv.org/abs/1810.11363.

FAWCETT, T. An introduction to ROC analysis. Pattern Recognition Letters, v. 27, n. 8, p. 861–874, 2006.

FRANCESCHI, Luca; DONINI, Michele; PERRONE, Valerio; KLEIN, Aaron; ARCHAMBEAU, Cédric; SEEGER, Matthias; PONTIL, Massimiliano; FRASCONI, Paolo. Hyperparameter optimization in machine learning. arXiv preprint arXiv:2410.22854, 2025. DOI: https://doi.org/10.48550/arXiv.2410.22854.

FRIEDMAN, J. H. Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, v. 29, n. 5, p. 1189–1232, 2001.

GÉRON, A. Hands-On: Machine Learning with Scikit-Learn, Keras, and TensorFlow. 2nd ed. Rio de Janeiro: Alta Books, 2023.

GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. Deep Learning. Cambridge, MA: MIT Press, 2016.

HAN, J.; KAMBER, M.; PEI, J. Data Mining: Concepts and Techniques. 4th ed. Amsterdam: Elsevier, 2022.

HANA, Ahmed; LOFSTEAD, Jay. Managing randomness to enable reproducible machine learning. In: ACM SYMPOSIUM ON PRINCIPLES OF DISTRIBUTED COMPUTING, 41., 2022, [S.l.]. Proceedings [...]. New York: ACM, 2022. Disponível em: <https://dl.acm.org/doi/10.1145/3526062.3536353\>. Acesso em: 4 ago. 2025.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The elements of statistical learning: data mining, inference, and prediction. 2. ed. New York: Springer, 2009.

HETTIARACHCHI, G. W.; HARSHANATH, K. D. A Predictive Analytics Framework for Student Dropout Prevention in Vocational Education. International Journal of Educational Technology in Higher Education, v. 22, n. 1, p. 1-18, 2025.

HUTTER, F.; HOOS, H. H.; LEYTON-BROWN, K. Automated Machine Learning: Methods, Systems, Challenges. Cham: Springer, 2019. Disponível em: ttps://doi.org/10.1007/978-3-030-05318-5. Acesso em: 27 maio 2025.

IBM. Introdução ao CRISP-DM. IBM Documentation, 2023. Disponível em: https://www.ibm.com/docs/pt-br/spss-modeler/18.4.0?topic=guide-introduction-crisp-dm. Acesso em: 15 jun. 2025.

JAMES, Gareth; WITTEN, Daniela; HASTIE, Trevor; TIBSHIRANI, Robert. An introduction to statistical learning: with applications in R. 2. ed. New York, NY: Springer, 2021. ISBN 978-1-0716-1418-1. Disponível em: https://doi.org/10.1007/978-1-0716-1418-1.

KE, Guolin; MENG, Qi; FINLEY, Thomas; WANG, Taifeng; CHEN, Wei; MA, Weidong; YE, Qiwei; LIU, Tie-Yan. LightGBM: a highly efficient gradient boosting decision tree. In: CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NeurIPS), 31., 2017, Long Beach. Proceedings. Long Beach: NeurIPS, 2017. Disponível em: https://proceedings.neurips.cc/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf. Acesso em: 23 jul. 2025.

KHORASANI, Mohammad; ABDOU, Mohamed; FERNÁNDEZ, Javier Hernández. Web Application Development with Streamlit: develop and deploy secure and scalable web applications to the cloud using a pure Python framework. Berkeley, CA: Apress, 2022. DOI: https://doi.org/10.1007/978-1-4842-8111-6.

KOCSIS, Ádám; MOLNÁR, Gyöngyvér. Factors influencing academic performance and dropout rates in higher education. Oxford Review of Education, v. 51, n. 3, p. 1–19, 2024. DOI: 10.1080/03054985.2024.2316616.

KOTSIANTIS, S. B. Handling Imbalanced Datasets: A Review. GESTS International Transactions on Computer Science and Engineering, v. 30, n. 1, p. 25–36, 2013.

KORKMAZ, Ö.; AYDIN, M. N. Detection of early school dropout in vocational and technical high schools in Turkey. Sage Open, v. 15, n. 3, 2025. DOI: https://doi.org/10.1177/21582440251370443.

KUHN, M.; JOHNSON, K. Applied Predictive Modeling. New York: Springer, 2013.

LUNDBERG, Scott M.; LEE, Su-In. A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, v. 30, 2017. Disponível em: https://proceedings.neurips.cc/paper/2017/file/8a20a8621978632d76c43dfd28b67767-Paper.pdf. Acesso em: 21 maio 2025.

MARBÁN, Óscar; MARISCAL, Gonzalo; SEGOVIA, Javier. A data mining & knowledge discovery process model. In: PONCE, Julio; KARAHOCA, Adem (Ed.). Data mining and knowledge discovery in real life applications. Rijeka: InTech, 2009. ISBN 978-3-902613-53-

0. Disponível em: https://cdn.intechopen.com/pdfs/5937/InTech-A_data_mining_amp_knowledge_discovery_process_model.pdf. Acesso em: 23 out. 2025.

MARTINEZ, W. L. Computational statistics handbook with MATLAB. 3. ed. Boca Raton: Chapman and Hall/CRC, 2019.

MDUMA, N.; KALEGELE, K.; MACHUVE, D. A survey of machine learning approaches and techniques for student dropout prediction. Data Science Journal, v. 8, p. 14, p. 1–10, 2019. DOI: https://doi.org/10.5334/dsj-2019-014.

MITCHELL, T. M. Machine learning. New York: McGraw-Hill, 1997.

MURPHY, K. P. Machine Learning: A Probabilistic Perspective. Cambridge: MIT Press, 2012.

MURTHY, S. K. Automatic construction of decision trees from data: a multi-disciplinary survey. Data Mining and Knowledge Discovery, v. 2, n. 4, p. 345–389, 1998.

MUSTOFA, Sumaya; EMON, Yousuf Rayhan; MAMUN, Sajib Bin; AKHY, Shabnur Anonna; AHAD, Md Taimur. A novel AI-driven model for student dropout risk analysis with explainable AI insights. Computers and Education: Artificial Intelligence, v. 8, p. 100352, 2025. DOI: https://doi.org/10.1016/j.caeai.2024.100352.

NAUJOKAITIENĖ, J.; TAMOLIŪNĖ, G.; VOLUNGEVIČIENĖ, A. et al. Using learning analytics to engage students: improving teaching practices through informed interactions. Journal of New Approaches in Educational Research, v. 9, p. 231–244, 2020. DOI: https://doi.org/10.7821/naer.2020.7.561

NIYOGISUBIZO, J.; LIAO, L.; NZIYUMVA, E.; MURWANASHYAKA, E.; NSHIMYUMUKIZA, P. C. Predicting student's dropout in university classes using two-layer ensemble machine learning approach: a novel stacked generalization. Computers and Education: Artificial Intelligence, v. 3, 100066, 2022. ISSN 2666-920X. DOI: https://doi.org/10.1016/j.caeai.2022.100066.

OECD. Education Policy Outlook: Brazil. [S.l.], 2020. Disponível em: https://www.oecd.org/content/dam/oecd/en/about/projects/edu/education-policy-outlook/country-profile-Brazil-2021-INT-PT.pdf. Acesso em: 22 jul. 2025.

PANG, Guansong; SHEN, Chunhua; CAO, Longbing; VAN DEN HENGEL, Anton. Deep learning for anomaly detection: a review. ACM Computing Surveys (CSUR), v. 54, n. 2, Art. 38, p. 1–38, 2021. Disponível em: https://doi.org/10.1145/3439950. Acesso em: 23 jul. 2025.

PEDREGOSA, Fabian; VAROQUAUX, Gaël; GRAMFORT, Alexandre; MICHEL, Vincent; THIRION, Bertrand; GRISEL, Olivier; BLONDEL, Mathieu; PRETTENHOFER, Peter; WEISS, Ron; DUBOURG, Vincent; et al. Scikit-learn: machine learning in Python. The Journal of Machine Learning Research, v. 12, p. 2825–2830, 2011. Disponível em: http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf, Acesso em: 23 jul. 2025.

PETNUCHOVA, J.; HRMO, R.; HORNAKOVA, V.; PODARIL, M.; STUR, M.; RIDZONOVA, Z.; NOVOTA, M. Vocational education and training in OECD countries. In:

2012 15th International Conference on Interactive Collaborative Learning (ICL), 2012. p. 1–6. DOI: 10.1109/ICL.2012.6402131.

PROKHORENKOVA, Liudmila; GUSEV, Gleb; VOROBEV, Aleksandr; DOROGUSH, Anna Veronika; GULIN, Andrey. CatBoost: unbiased boosting with categorical features. In: BENGIO, S. et al. (ed.). Advances in Neural Information Processing Systems 31. Red Hook, NY: Curran Associates, Inc., 2018. Disponível em: https://proceedings.neurips.cc/paper_files/paper/2018/file/14491b756b3a51daac41c24863285549-Paper.pdf. Acesso em: 23 jun. 2025.

PYLE, Dorian. Data preparation for data mining. San Francisco: Morgan Kaufmann, 1999.

PYTHON SOFTWARE FOUNDATION. pickle — serialização de objetos Python. Documentação Python 3.13.9, 2025. Disponível em: https://docs.python.org/pt-br/3.13/library/pickle.html. Acesso em: 25 jul. 2025.

QUAYE, S. J.; HARPER, S. R.; PENDAKUR, S. L. (eds.). Engajamento estudantil no ensino superior: perspectivas teóricas e abordagens práticas para populações diversas. 3. ed. Nova York: Routledge, 2019.

RASCHKA, S.; MIRJALILI, V. Python Machine Learning: Machine Learning and Deep Learning with Python, scikit-learn, and TensorFlow 2. 4th ed. Birmingham: Packt Publishing, 2022.

RASTROLLO-GUERRERO, Juan L.; GÓMEZ-PULIDO, Juan A.; DURÁN-DOMÍNGUEZ, Arturo. Analyzing and predicting students' performance by means of machine learning: a review. Applied Sciences, [S. l.], v. 10, n. 3, p. 1042, 2020. Disponível em: https://doi.org/10.3390/app10031042.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. arXiv preprint arXiv:2402.07956, 2024. Disponível em: https://arxiv.org/abs/2402.07956. Acesso em: 17 jun. 2025.

RUMBERGER, R. W. 2018. Why Students Drop Out of School. In:. Dropping Out: Why Students Drop Out of High School and What Can Be Done About It. Cambridge, MA: Harvard University Press, 2018. p. 1-25.

SCHRÖER, Christoph; KRUSE, Felix; GÓMEZ, Jorge Marx. A systematic literature review on applying CRISP-DM process model. Procedia Computer Science, v. 181, p. 526–534, 2021. DOI: 10.1016/j.procs.2021.01.199.

SANDOVAL-PALIS, I.; NARANJO, D.; VIDAL, J.; GILAR-CORBI, R. Modelo de previsão de abandono precoce: um estudo de caso com alunos de cursos de nivelamento universitário. Sustainability, v. 12, n. 22, p. 9314, 2020. Disponível em: https://doi.org/10.3390/su12229314. Acesso em: 23 jul. 2025.

SARANGPURE, N.; DHAMDE, V.; ROGE, A.; DOYE, J.; PATLE, S.; TAMBOLI, S. Automatizando o processo de aprendizado de máquina usando PyCaret e Streamlit. In: INTERNATIONAL CONFERENCE FOR INNOVATION IN TECHNOLOGY (INOCON), 2., 2023, Bangalore, Índia. Anais... [S.l.: s.n.], 2023. p. 1–5. DOI: 10.1109/INOCON57975.2023.10101357.

SCIKIT-LEARN. sklearn.metrics.precision_score. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.precision_score.html. Acesso em: 20 maio 2025.

SEO, Eui-Yeong; YANG, Jaemo; LEE, Ji-Eun; SO, Geunju. Predictive modelling of student dropout risk: Practical insights from a South Korean distance university. Heliyon, v. 10, n. 11, p. e30960, 2024. DOI: https://doi.org/10.1016/j.heliyon.2024.e30960.

SILVA, Samoel Rodrigues da; FILHO, Samuel Brasileiro; FERNANDES, Natal Lânia Roque. Evasão e permanência no ensino técnico ofertado na Rede Federal: análise dos estudos da Pós-graduação stricto sensu brasileira. Revista Brasileira da Educação Profissional e Tecnológica, [S. l.], v. 3, n. 24, p. e13205, 2024. DOI: 10.15628/rbept.2024.13205. Disponível em: https://www2.ifrn.edu.br/ojs/index.php/RBEPT/article/view/13205. Acesso em: 24 jun. 2025.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks. Information Processing & Management, v. 45, n. 4, p. 427–437, 2009.

STREAMLIT COMMUNITY. Streamlit documentation. Disponível em: https://docs.streamlit.io/. Acesso em: 26 jun. 2025.

TINTO, Vincent. Dropout from higher education: a theoretical synthesis of recent research. Review of Educational Research, v. 45, n. 1, p. 89–125, 1975. DOI: 10.3102/00346543045001089.

THOKALA, Vasudhar Sai. Integrating Machine Learning into Web Applications for Personalized Content Delivery using Python. International Journal of Current Engineering and Technology, v. 11, n. 6, p. 652–660, nov./dez. 2021. DOI: https://doi.org/10.14741/ijcet/v.11.6.9.

TOMLINSON, M.; WALKER, L. Student Engagement and Retention in Higher and Technical Education: Global Challenges and Local Responses. London: Routledge, 2022.

UNESCO. Policy Review of Vocational Education and Training. Paris: UNESCO, 2018.

UNESCO. Reimagining Vocational Education for the Future of Work. Paris: UNESCO Publishing, 2023.

VAARMA, M. Predicting student dropouts with machine learning. ScienceDirect, 2024.

VAARMA, M.; LI, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. Technology in Society, v. 76, p. 102474, 2024. DOI: https://doi.org/10.1016/j.techsoc.2024.102474.

YI, H.; ALLEN, J.; SOLGA, H. Firm-Provided Training and the Career Prospects of Young Workers: Training Volumes, Signaling Effects, and Labor Market Outcomes. European Sociological Review, v. 31, n. 5, p. 525–540, 2015. Disponível em: https://doi.org/10.1093/esr/jcv062. Acesso em: 29 maio 2025.

YI, Hongmei; ZHANG, Linxiu; YAO, Yezhou; WANG, Aiqin; MA, Yue; SHI, Yaojiang; CHU, James; LOYALKA, Prashant; ROZELLE, Scott. Exploring the dropout rates and causes of dropout in upper-secondary technical and vocational education and training (TVET)

schools in China. International Journal of Educational Development, v. 42, p. 115-123, 2015. DOI: https://doi.org/10.1016/j.ijedudev.2015.04.009.

7 CONSIDERAÇÕES FINAIS DA TESE

A evasão escolar é um problema que afeta todos os níveis de ensino, mas, no caso da educação profissional e tecnológica, em cursos técnicos, é ainda mais grave, pois a conclusão desses cursos não é um requisito para ingressar em outros níveis educacionais.

Quando a evasão escolar ocorre na Rede Federal de Educação Profissional, Científica e Tecnológica, algumas variáveis se intensificam. Isso ocorre porque, em sua missão de expandir a oferta de educação profissional e tecnológica para regiões que anteriormente não tinham acesso a essa formação, fatores familiares do estudante, infraestrutura, transporte, acesso à internet, entre outros, tornam-se desafios ainda mais significativos para o abandono dos estudos.

Para combater a evasão escolar, é necessário entender de forma unificada as razões que a causam. É evidente que os fatores que contribuem para o abandono escolar geralmente incluem aspectos relacionados ao próprio indivíduo, ao ambiente institucional e ao contexto socioeconômico em que ele está inserido.

A inteligência artificial tem se tornado cada vez mais popular e sua aplicação prática já pode ser vista no cotidiano de milhões de pessoas. Ela se destaca como uma ferramenta estratégica de alto potencial, considerando a existência de bases de dados institucionais acessíveis ao público, que permite a utilização de modelos de aprendizado de máquina capazes de executar análises preditivas confiáveis.

As técnicas de aprendizado de máquina supervisionado permitem classificar os estudantes como evadidos ou não evadidos com base em variáveis categóricas e contínuas. A identificação precoce de alunos em situação de risco oferece aos gestores uma ferramenta eficiente para tomar decisões proativas e implementar intervenções pedagógicas personalizadas.

Este trabalho alcançou os seus objetivos ao desenvolver e implementar técnicas de IA para a predição da evasão escolar em cursos técnicos da RFEPCT. A metodologia fundamentada em pesquisa bibliográfica e documental seguiu as etapas da metodologia CRISP-DM para análise de dados e modelagem computacional. A pesquisa percorreu um ciclo completo, desde o entendimento do negócio e dos dados até a preparação, modelagem, avaliação e, finalmente, implementação dos resultados em um protótipo de ferramenta disponível na web para uso por gestores e técnicos envolvidos na gestão educacional e na formulação de políticas públicas.

A primeira etapa do projeto, detalhadamente descrita no primeiro artigo, relacionada ao entendimento e à preparação dos dados, mostrou-se fundamental para o sucesso das fases

seguintes. O tratamento de um grande volume de informações exigiu um processamento criterioso para assegurar a qualidade e a consistência do conjunto de dados, que se tornou a referência para todos os modelos. Nesta etapa, foi a base que possibilitou a transição da teoria para a aplicabilidade em um projeto real de aprendizado de máquina. Durante a etapa de modelagem, a utilização de algoritmos supervisionados, com destaque para as técnicas de *boosting* (CatBoost, XGBoost e LightGBM), demonstrou a eficiência desses métodos em lidar com o desbalanceamento, um desafio que motivou o uso de técnicas como o SMOTE. O modelo CatBoost demonstrou superioridade, atestada por métricas como precisão, recall, FI-score e, principalmente, a área sob a curva ROC-AUC. Esse modelo se consolidou como o ideal para distinguir com alta acurácia entre alunos propensos à evasão e aqueles que provavelmente permanecerão no curso.

No entanto, desenvolver um modelo preditivo com bom desempenho que não seja explicável pode se tornar insuficiente se suas decisões não forem compreensíveis para os que forem utilizá-lo. Na etapa de avaliação, o estudo deu visibilidade ao implementar o método SHAP (SHapley Additive exPlanations) para explicar as predições do modelo CatBoost. Essa abordagem, apresentada no segundo artigo, identificou as variáveis determinantes e quantificou sua influência na probabilidade de evasão. Ao confirmar que as previsões do modelo estão em concordância com o padrão teórico esperado para a evasão, o SHAP proporciona transparência e credibilidade ao modelo apresentado. Como resultado, a interpretabilidade passou a ser uma característica própria do modelo, possibilitando não só identificar quem está em risco, mas também entender o motivo baseado nas variáveis mais determinantes.

Por fim, a aplicação prática deste projeto foi contemplada na etapa de implantação do CRISP-DM, com o desenvolvimento da plataforma PrevIA. Ao integrar o modelo CatBoost em uma interface interativa e de fácil acesso, o estudo materializou seu objetivo principal. A plataforma possibilita a simulação de forma individualizada do risco de evasão de um estudante, considerando suas características sociodemográficas, individuais e do curso em andamento ou pretendido. A efetivação da modelagem computacional em uma ferramenta operacional representa a ligação entre a pesquisa acadêmica e a gestão educacional prática, pois viabiliza a intervenção antecipada e personalizada, que é fundamental para reverter o cenário de evasão.

A contribuição deste trabalho se dá em dois aspectos. No campo teórico- científico, ele contribui para a formulação de políticas institucionais e pesquisas abrangentes, fortalecendo a cultura de gestão orientada a dados, a aplicação de técnicas de aprendizado de máquina e a discussão sobre os motivos da evasão em cursos técnicos na RFEPCT. Na prática, fornece uma metodologia consistente e replicável, desde a preparação de dados até o desenvolvimento de

uma ferramenta tecnológica, que possibilita a identificação e a explicação do risco de evasão de maneira individualizada, otimizando recursos e direcionando esforços de forma assertiva. Além disso, a plataforma PrevIA está disponível para o público, hospedada na nuvem e com código-fonte acessível, promovendo a transparência, reprodutibilidade e potencial reutilização.

Como limitações do estudo, destaca-se a utilização de um conjunto de dados que inclui apenas matrículas finalizadas de um único ano (2023) e a quantidade restrita de variáveis disponíveis. Além disso, a evasão é um processo com muitos fatores e dinâmico, e a ausência de informações em tempo real, como frequência, desempenho acadêmico e participação em projetos, representa uma oportunidade a ser investigada para melhorar a precisão e a profundidade da análise.

Portanto, este estudo não se encerra aqui, trabalhos futuros podem continuar a pesquisa aprimorando os modelos preditivos com a utilização de outras técnicas, algoritmos de aprendizado de máquina e métodos explicativos para comparação com as saídas explicáveis do método SHAP e novos parâmetros capazes de otimizar os resultados das métricas analisadas. Do ponto de vista dos dados, a inclusão de anos adicionais para criar uma série histórica pode verificar se as probabilidades preditas permanecem. Para a plataforma PrevIA, sugere-se a implementação de funcionalidades como a simulação em lote por curso, a integração com sistemas acadêmicos locais que monitoram a trajetória do estudante e a coleta de *feedback* dos usuários para aprimorar a ferramenta.

A inteligência artificial e os sistemas preditivos, como o desenvolvido neste projeto, são ferramentas estratégicas na gestão educacional. A evasão em cursos técnicos configura-se como um complexo processo social cuja superação demanda um esforço contínuo de diagnóstico, ações pedagógicas e cooperação entre setores da instituição e do governo. A plataforma PrevIA e o modelo implementado ampliam a capacidade de intervenção ao fornecer clareza sobre os padrões que potencializam o risco de evasão. A tecnologia tem o propósito de apoiar a ação humana, oferecendo evidências para que as políticas de permanência e êxito estudantil sejam mais bem direcionadas. A utilização constante da plataforma, com ajustes regulares do modelo com dados atualizados, pode torná-la cada vez mais precisa e eficaz para a tomada de decisões estratégicas nas instituições da Rede Federal de Educação Profissional, Científica e Tecnológica.

REFERÊNCIAS

BARROS, R. P.; FRANCO, S.; MACHADO, L. M.; ZANON, D.; ROCHA, G. Consequências da violação do direito à educação. 1. ed. Rio de Janeiro: Insper, 2021. 148 p. Disponível em: https://www.frm.org.br/a88b2f1a-1c02-4d55-b74a-1653b3309be3. Acesso em: 27 maio 2025.

BAKER, R. S.; SIEMENS, G. Educational data mining and learning analytics. In: SAWYER, R. K. (ed.). The Cambridge Handbook of the Learning Sciences. 2. ed. Cambridge: Cambridge University Press, 2014. p. 253-274.

BESSEY, D.; BACKES-GELLNER, U. Drop-out and occupational outcomes in vocational education and training. Empirical Research in Vocational Education and Training, v. 7, n. 1, p. 1-20, 2015.

BÖHN, S.; DEUTSCHER, V. Determinants of dropout in vocational education and training: a systematic review. Empirical Research in Vocational Education and Training, v. 14, n. 3, p. 1-25, 2022.

BRASIL. Centro de Gestão e Estudos Estratégicos (CGEE). A educação profissional e tecnológica no Brasil: análise e subsídios para políticas públicas. Brasília, DF: CGEE, 2023. Disponível em: https://www.cgee.org.br. Acesso em: 27 maio. 2025.

BRASIL Tribunal de Contas da União. Auditoria operacional na Rede Federal de Educação Profissional, Científica e Tecnológica: Relatório de Auditoria. Brasília: TCU, 2024.

CHAPMAN, P. et al. CRISP-DM 1.0: Step-by-step data mining guide. SPSS Inc., 2000.

CERVO, A. L.; SILVA, R.; BERVIAN, P. A. Metodologia científica. 6. ed. São Paulo: Pearson Prentice Hall, 2007.

CRESWELL, J. W. Research design: qualitative, quantitative, and mixed methods approaches. 3. ed. Thousand Oaks: SAGE, 2010.

MACHADO, J.; FERREIRA, A.; COSTA, F. Aplicações de aprendizado de máquina na previsão da evasão escolar: uma revisão sistemática. Revista Brasileira de Informática na Educação, v. 29, p. 1-20, 2021.

MATZ, Sandra C. et al. Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics. Scientific Reports, [S. l.], v. 13, n. 1, 5705, 2023. Disponível em: https://doi.org/10.1038/s41598-023-32593-6.

PEDDITZI M. L. School Satisfaction and Self-Efficacy in Adolescents and Intention to Drop Out of School. International Journal of Environmental Research and Public Health. 2024; 21(1):111. https://doi.org/10.3390/ijerph21010111.

RASTROLLO-GUERRERO, Juan L.; GÓMEZ-PULIDO, Juan A.; DURÁN-DOMÍNGUEZ, Arturo. Analyzing and predicting students' performance by means of machine learning: a review. Applied Sciences, [S. l.], v. 10, n. 3, p. 1042, 2020. Disponível em: https://doi.org/10.3390/app10031042.

YI, Hongmei; ZHANG, Linxiu; YAO, Yezhou; WANG, Aiqin; MA, Yue; SHI, Yaojiang; CHU, James; LOYALKA, Prashant; ROZELLE, Scott. Exploring the dropout rates and causes of dropout in upper-secondary technical and vocational education and training (TVET) schools in China. International Journal of Educational Development, v. 42, p. 115-123, 2015. DOI: https://doi.org/10.1016/j.ijedudev.2015.04.009.

YIN, R. K. Estudo de caso: planejamento e métodos. 5. ed. Porto Alegre: Bookman, 2015.

APÊNDICE A – PROJETO DISPONIBILIZADO NO GITHUB

A tela do projeto onde constam as bases de dados usadas e processadas na modelagem computacional, além dos notebooks e do arquivo com as bibliotecas utilizadas (requirements.txt).

PrevIA-Predicao-Evasao-Rede-Federal-Copia

jabsoncd Update Home_Profissional.py 2404cfd ·		1 hour ago
.devcontainer	Added Dev Container Folder	last week
artifacts	hoje1	5 months ago
images	previav2	5 months ago
Input	first commit	5 months ago
notebooks	beta2	last week
src	first commit	5 months ago
static/images	first commit	5 months ago
templates	Update Home_Profissional.py	1 hour ago
.gitattributes	first commit	5 months ago
.gitignore	first commit	5 months ago
BR_UF_2024.geojson	ibge	5 months ago
requirements.txt	Update requirements.txt	last week
runtime.txt	req8	5 months ago