

**INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA  
CAMPUS DE CAJAZEIRAS  
CURSO SUPERIOR DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE  
SISTEMAS**

**GERALDO MENDES BATISTA NETO**

**ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR NO ENSINO MÉDIO EM  
INSTITUIÇÕES FEDERAIS BRASILEIRAS**

Cajazeiras  
2024

GERALDO MENDES BATISTA NETO

ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR NO ENSINO MÉDIO EM  
INSTITUIÇÕES FEDERAIS BRASILEIRAS

Trabalho de Conclusão de Curso apresentado junto ao Curso Superior de Tecnologia em Análise e Desenvolvimento de Sistemas do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba - *Campus* Cajazeiras, como requisito a obtenção do título de Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador: Prof. Dr. Fábio Gomes de Andrade.

Cajazeiras  
2024

IFPB / Campus Cajazeiras  
Coordenação de Biblioteca  
Biblioteca Prof. Ribamar da Silva  
Catalogação na fonte: Cícero Luciano Félix CRB-15/750

B333a	<p>Batista Neto, Geraldo Mendes. Análise e previsão da evasão escolar no ensino médio em instituições federais brasileiras / Geraldo Mendes Batista Neto. – 2024.</p> <p>65f. : il.</p> <p>Trabalho de Conclusão de Curso (Tecnólogo em Análise e Desenvolvimento de Sistemas) - Instituto Federal de Educação, Ciência e Tecnologia da Paraíba, Cajazeiras, 2024.</p> <p>Orientador(a): Prof. Dr. Fabio Gomes de Andrade.</p> <p>1. Desenvolvimento de sistemas. 2. Ciência de dados. 3. Evasão escolar. 4. Modelo preditivo. I. Instituto Federal de Educação, Ciência e Tecnologia da Paraíba. II. Título.</p>
-------	-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------



MINISTÉRIO DA EDUCAÇÃO  
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA  
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

GERALDO MENDES BATISTA NETO

**ANÁLISE E PREVISÃO DA EVASÃO ESCOLAR EM INSTITUIÇÕES FEDERAIS DE ENSINO  
MÉDIO NO BRASIL**

Trabalho de Conclusão de Curso apresentado junto ao  
Curso Superior de Tecnologia em Análise e  
Desenvolvimento de Sistemas do Instituto Federal de  
Educação, Ciência e Tecnologia da Paraíba - Campus  
Cajazeiras, como requisito à obtenção do título de  
Tecnólogo em Análise e Desenvolvimento de Sistemas.

Orientador

Prof. Dr. Fabio Gomes de Andrade

Aprovada em: **17 de Outubro de 2024.**

Prof. Dr. Fabio Gomes de Andrade - Orientador

Prof. Me. Janderson Ferreira Dutra - Avaliador

IFPB - Campus Cajazeiras

Prof. Me. Afonso Serafim Jacinto

IFPB - Campus Cajazeiras

Documento assinado eletronicamente por:

- **Fabio Gomes de Andrade**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/10/2024 18:02:06.
- **Janderson Ferreira Dutra**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/10/2024 21:05:14.
- **Afonso Serafim Jacinto**, PROFESSOR ENS BASICO TECN TECNOLOGICO, em 18/10/2024 22:42:38.

Este documento foi emitido pelo SUAP em 18/10/2024. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifpb.edu.br/autenticar-documento/> e forneça os dados abaixo:

Código 621765  
Verificador: 0c8f1cf976  
Código de Autenticação:



Rua José Antônio da Silva, 300, Jardim Oásis, CAJAZEIRAS / PB, CEP 58.900-000  
<http://ifpb.edu.br> - (83) 3532-4100

## **AGRADECIMENTOS**

Primeiramente, agradeço a Deus, por me conceder a vida, a força necessária para superar os momentos mais desafiadores e por sempre me guiar pelos melhores caminhos.

Aos meus pais, Maria Efigênia e Francisco Sitônio, por todo amor, apoio incondicional e pelos sacrifícios feitos para que este sonho se tornasse realidade. Minha gratidão será eterna.

Aos meus irmãos, Nattan Mendes e Vitória Mendes, pelo companheirismo, carinho e por sempre estarem ao meu lado, compartilhando os desafios e as conquistas desta jornada.

Ao meu orientador, Prof. Dr. Fabio Gomes de Andrade, pela orientação precisa, pelas valiosas revisões, sugestões e correções, e por ter aceitado conduzir este trabalho com tanto profissionalismo.

Ao amigo Matheus Valença, pela grande ajuda no desenvolvimento deste trabalho.

Aos meus amigos, pelo apoio constante, palavras de incentivo e pela amizade sincera que sempre foi uma fonte de motivação.

Aos amigos que fiz ao longo desta caminhada, especialmente Leticia Estrela, José Ferreira, Bruno Vasconcelos, Richard Freitas, Elivelton Pereira e Fulgêncio Thierry pelo companheirismo e apoio durante toda a trajetória.

Aos meus colegas de graduação e aos demais amigos que, de diversas formas, contribuíram para o meu crescimento pessoal e acadêmico.

Aos professores do Instituto Federal da Paraíba (IFPB), Campus Cajazeiras, por oferecerem um ensino gratuito e de qualidade, e por serem verdadeiros exemplos de humanidade e dedicação ao ensino.

E a todos aqueles que, mesmo não citados nominalmente, de alguma maneira contribuíram para este trabalho e para a pessoa que sou hoje. A todos, meu mais sincero agradecimento.

## RESUMO

Diante dos desafios enfrentados pelo sistema educacional, a evasão escolar se destaca como um fenômeno crítico, impactando estudantes de diversas realidades socioeconômicas. Com base nisso, este trabalho propõe a construção de um modelo preditivo, fundamentado em técnicas de Aprendizado de Máquina, para prever a evasão no ensino médio brasileiro, utilizando dados do censo escolar dos anos de 2019 a 2021. A pesquisa explora a complexidade da evasão, considerando fatores como idade, turno, renda familiar e carga horária. Foram aplicados modelos de aprendizado de máquina, como KNN, Regressão Logística, Árvore de Decisão e *Random Forest*, sendo o último o que obteve maior acurácia, com 97% nos cursos de nível médio, enquanto a Árvore de Decisão se destacou nos cursos técnicos, apresentando 86% de acurácia. Os resultados visam facilitar intervenções personalizadas e apoiar alunos em situação de vulnerabilidade, oferecendo uma abordagem na prevenção da evasão escolar e contribuindo para o campo educacional de forma prática.

**Palavras-chave:** Evasão escolar. Modelo preditivo. Ciência de Dados. Ensino médio.

## **ABSTRACT**

In light of the challenges faced by the educational system, school dropout stands out as a critical phenomenon, impacting students from diverse socioeconomic backgrounds. This work proposes the construction of a predictive model, grounded in Data Science techniques, to forecast dropout rates in Brazilian high schools, utilizing data from the school census from 2019 to 2021. The research explores the complexity of dropout, considering factors such as age, shift, family income, and workload. Machine learning models were applied, including KNN, Logistic Regression, Decision Tree, and Random Forest, with the latter achieving the highest accuracy of 97% in high school courses, while the Decision Tree excelled in technical courses, presenting an accuracy of 86%. The results aim to facilitate personalized interventions and support students in vulnerable situations, offering an innovative approach to preventing school dropout and contributing practically to the educational field.

**Keywords:** School dropout. Predictive model. Data Science. High school.



## LISTA DE FIGURAS

Figura 1 - Evolução da taxa de evasão escolar no ensino médio – Brasil.....	23
Figura 2 - Taxa de evasão no ensino médio – Brasil, rede pública, 2017/2018. ....	24
Figura 3 - Principal motivo declarado por adolescentes e jovens de 15 a 19 anos para terem saído da escola – Brasil, 2019.....	26
Figura 4 - Interdisciplinaridade da Ciência de Dados. ....	29
Figura 5 - Fases do modelo do CRISP-EDM.....	29
Figura 6 - Exemplo de classificação do KNN com dois rótulos de classe e $k = 7$ .....	35
Figura 7 - Árvore de Decisão. ....	37
Figura 8 - Representação de uma Floresta Aleatória.....	38
Figura 9 - Regressão Logística - Função Sigmoides. ....	39
Figura 10 - Evasão por turno no ensino médio.....	43
Figura 11 - Evasão por faixa etária ensino médio. ....	43
Figura 12 - Evasão por idade no ensino médio. ....	44
Figura 13 - Evasão por Cor/Raça no ensino médio.....	45
Figura 14 - Evasão por Renda Familiar no ensino médio. ....	46
Figura 15 - Evadido por UF no ensino médio. ....	46
Figura 16 - Evadidos x Concluintes por carga horária no ensino médio.....	47
Figura 17 - Evasão por turno no ensino médio técnico. ....	47
Figura 18 - Evasão por faixa etária no ensino médio técnico.....	48
Figura 19 - Evasão por idade no ensino médio técnico.....	49
Figura 20 - Evasão por Cor/Raça no ensino médio técnico. ....	49
Figura 21 - Evasão por renda familiar no ensino médio técnico. ....	50
Figura 22 - Evasão por região no ensino médio técnico. ....	51
Figura 23 - Evadidos x Concluintes por carga horária o ensino médio técnico. ....	51

## LISTA DE TABELAS

Tabela 1 - Estatísticas dos modelos do ensino médio. ....	54
Tabela 2 - Estatísticas dos modelos do ensino médio técnico. ....	55

## LISTA DE QUADROS

Quadro 1 - Dicionário das principais variáveis.....	41
-----------------------------------------------------	----

## LISTA DE ABREVIATURAS E SIGLAS

AED	Análise Exploratória de Dados
AM	Aprendizado de Máquina
CD	Ciência de Dados
CRISP-DM	<i>CRoss-Industry Standard Process for Data Mining</i>
CRISP-EDM	<i>CRoss Industry Standard Process for Educational Data Mining</i>
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
MEC	Ministério da Educação
KNN	K-Nearest Neighbors

## SUMÁRIO

<b>1. INTRODUÇÃO .....</b>	<b>15</b>
<b>1.1. Motivação.....</b>	<b>15</b>
<b>1.2. Objetivos.....</b>	<b>16</b>
1.2.1. Objetivo Geral.....	16
1.2.2. Objetivos Específicos .....	16
<b>1.3. Trabalhos Relacionados .....</b>	<b>17</b>
<b>1.4. Metodologia .....</b>	<b>18</b>
<b>1.5. Organização do Documento .....</b>	<b>19</b>
<b>2. FUNDAMENTAÇÃO TEÓRICA.....</b>	<b>21</b>
<b>2.1. Evasão escolar.....</b>	<b>21</b>
2.1.1. Panorama atual da evasão escolar no Brasil.....	22
2.1.2. Causas da evasão escolar no ensino médio .....	25
<b>2.2. Ciência de Dados .....</b>	<b>27</b>
2.2.1. Análise exploratória de dados .....	31
2.2.2. Pré-processamento de dados .....	32
2.2.3. Aprendizado de Máquina.....	33
<b>2.3. Algoritmos de classificação .....</b>	<b>34</b>
2.3.1. <i>K-Nearest Neighbors</i> .....	35
2.3.2. Árvore de Decisão .....	36
2.3.3. Floresta Aleatória.....	37
2.3.4. Regressão Logística.....	38
<b>3. ANÁLISE EXPLORATÓRIA DOS DADOS .....</b>	<b>40</b>
<b>3.1. Entendimento de dados .....</b>	<b>40</b>
<b>3.2. Etapas da análise exploratória de dados .....</b>	<b>41</b>
3.2.1. Limpeza de dados e tratamento de valores ausentes .....	42
3.2.2. Análise de distribuição das variáveis do ensino médio.....	42

3.2.3. Análise de distribuição das variáveis do ensino médio técnico .....	47
<b>4. MODELAGEM PREDITIVA.....</b>	<b>52</b>
<b>4.1. Preparação dos dados para modelagem .....</b>	<b>52</b>
<b>4.2. Treinamento dos modelos .....</b>	<b>53</b>
<b>4.3. Avaliação de desempenho dos modelos .....</b>	<b>54</b>
<b>5. CONCLUSÃO.....</b>	<b>57</b>
<b>5.1. Trabalhos futuros .....</b>	<b>58</b>
<b>REFERÊNCIAS.....</b>	<b>60</b>

## **1. INTRODUÇÃO**

A persistente e complexa problemática da evasão escolar nas instituições de ensino no Brasil representa um desafio significativo para o sistema educacional, impactando diretamente o desenvolvimento acadêmico dos estudantes (OLIVEIRA; NÓBREGA, 2021). Segundo Queiroz (2002), a evasão escolar é um tema histórico nos debates sobre a educação pública brasileira, ocupando, até os dias atuais, um espaço de relevância no cenário das políticas educacionais. Essa questão, que há décadas permeia as discussões sobre a educação no Brasil, evidencia a necessidade urgente de ações efetivas e políticas inovadoras para enfrentar e superar os desafios associados.

Diante dos obstáculos enfrentados pelo sistema educacional, a evasão escolar desempenha um papel crucial ao afetar alunos de diferentes realidades socioeconômicas (OLIVEIRA; NÓBREGA, 2021). Conforme destacado por Rodrigues (2020), essa problemática transcende a esfera educacional, tornando-se uma questão social que demanda soluções proativas.

Com o intuito de abordar a evasão escolar, este Trabalho de Conclusão de Curso propõe o desenvolvimento de um modelo preditivo para identificar características de risco de evasão, utilizando dados do Censo Escolar da Educação Básica no Brasil. Além disso, foi aplicada uma abordagem de Aprendizado de Máquina para prever a evasão escolar de maneira eficaz. Reconhecendo as complexidades desse fenômeno e as limitações inerentes aos dados e métodos empregados, a pesquisa realizada neste trabalho busca oferecer uma contribuição significativa para a compreensão e mitigação desse desafio persistente.

### **1.1. Motivação**

A motivação para o desenvolvimento deste trabalho é impulsionada pela necessidade de enfrentar o desafio persistente da evasão escolar. Conforme abordado por Oliveira e Nóbrega (2021), esse problema representa não apenas uma perda significativa de recursos e investimentos educacionais, mas também impacta diretamente o potencial de desenvolvimento e sucesso dos estudantes.

O problema central a ser abordado reside na complexidade da compreensão e previsão dos fatores que contribuem para a evasão escolar, especialmente em instituições federais de ensino médio. A ausência de estratégias proativas, como a utilização de modelos preditivos para a antecipação de casos de evasão, limita a

capacidade das instituições de intervir precocemente. Isso prejudica a eficácia na implementação de medidas preventivas, como o reforço acadêmico, apoio psicológico ou ajustes nos currículos, que poderiam ser aplicados para alunos em risco de abandono escolar.

Essa abordagem tem como objetivo equipar as instituições federais de ensino médio com uma ferramenta que lhes permita atuar de forma proativa, identificando padrões comportamentais ou demográficos associados à evasão e implementando estratégias de intervenção personalizada, como acompanhamento pedagógico ou suporte social. Ao direcionar a atenção para a previsão da evasão, este trabalho busca contribuir para a eficácia do sistema educacional, promovendo a manutenção dos estudantes nas instituições de ensino e, conseqüentemente, ampliando suas oportunidades de sucesso acadêmico e profissional.

## **1.2. Objetivos**

Esta seção descreve os objetivos alcançados por meio do desenvolvimento deste trabalho.

### **1.2.1. Objetivo Geral**

Analisar e compreender os fatores que influenciam a evasão escolar nos cursos de ensino médio e ensino médio técnico das instituições federais de ensino no Brasil, além de desenvolver um modelo preditivo que permita prever casos de evasão em ambos os níveis de ensino.

### **1.2.2. Objetivos Específicos**

O trabalho tem ainda os seguintes objetivos específicos:

- identificar e analisar os fatores acadêmicos, socioeconômicos e demográficos relacionados à evasão escolar nas instituições federais de ensino médio e ensino médio técnico;
- compreender, com base na literatura, os principais métodos e técnicas de Aprendizado de Máquina aplicados à previsão de evasão escolar;
- desenvolver um modelo preditivo que permita identificar antecipadamente os alunos com risco de evasão nos cursos do ensino médio e técnico.
- fornecer percepções para a implementação de estratégias preventivas, visando o aprimoramento do sistema educacional para o ensino médio e ensino médio técnico.



### 1.3. Trabalhos Relacionados

Diversos estudos correlatos abordam a problemática da evasão escolar. A utilização de técnicas de Aprendizado de Máquina, conforme abordado por Mitchell (1997), tem ganhado destaque como uma ferramenta eficaz no enfrentamento da evasão escolar nos últimos anos, com uma ênfase significativa nos níveis de ensino superior.

Silva (2023) realizou uma pesquisa que teve como objetivo principal desenvolver modelos de aprendizado de máquina que fossem capazes de identificar alunos que apresentam maior chance de abandonar o curso técnico de nível médio. Para isso, o autor utilizou dados dos alunos obtidos através do cadastro do aluno na ficha de matrícula e incluiu informações socioeconômicas dos alunos e seu desempenho no processo seletivo. No estudo foram utilizados cinco algoritmos de aprendizado de máquina para prever a evasão escolar em cursos técnicos de nível médio. Foram eles: *Naive Bayes* (NB) (FRIEDMAN *et al.*, 1997) Método do Vizinho mais próximo (KNN) (LAAKSONEN e OJA, 1996), Máquina de Vetor de Suporte (SVM) (NOBLE, 2006), Árvore de Decisão (*Random Forest*) (KOTSIANTIS, 2013) e Redes Neurais Artificiais (MÜLLER *et al.*, 1995). A avaliação do desempenho dos algoritmos foi conduzida por meio de uma abordagem de validação cruzada, e a métrica de acurácia média foi empregada como indicador-chave. O trabalho apresenta uma abordagem aplicada e evidencia a importância da utilização de técnicas de aprendizado de máquina no contexto escolar para a prevenção da evasão escolar em cursos técnicos de nível médio.

O trabalho desenvolvido por Colpani, (2018) apresenta um estudo voltado para a análise e compreensão do problema da evasão escolar no ensino médio brasileiro. Nessa pesquisa o autor se utilizou de técnicas de correlação e regressão linear, visando identificar as principais variáveis educacionais que estão relacionadas à evasão dos estudantes do ensino médio.

Outro estudo que buscou prever a evasão dos alunos do ensino médio por meio de dados educacionais públicos foi apresentado por Machado (2019), que teve como objetivo principal apresentar modelos de aprendizado de máquina capazes de identificar alunos que apresentavam maior chance de abandonar o ensino médio. Para isso, o autor utilizou dados do Censo Escolar e de geolocalização de escolas e comunidades fluminenses. No estudo foram testados modelos de regressão e *random forest* multiníveis para prever a evasão escolar.

Souza (2016) realizou um estudo de caso visando analisar os fatores que influenciam a permanência ou evasão dos alunos nos cursos técnicos subsequentes, com ênfase no Curso Técnico Subsequente de Redes de Computadores, ofertado pelo Instituto Federal de Educação, Ciência e Tecnologia do Rio Grande do Norte – IFRN, Campus São Gonçalo do Amarante. A pesquisa foi realizada por meio de uma abordagem essencialmente qualitativa, utilizando pesquisa bibliográfica, pesquisa documental, entrevista e questionário como instrumentos de coleta de dados.

Ao contrário dos trabalhos supracitados, este trabalho tem o objetivo de utilizar uma abordagem abrangente e focalizada nas instituições federais de ensino médio e ensino médio técnico do país. O estudo visa explorar, de maneira aprofundada, a problemática da evasão escolar nesse contexto específico, utilizando uma ampla gama de dados e técnicas analíticas. Embora muitos estudos se concentram em regiões geográficas ou níveis específicos de ensino mais amplos, este trabalho busca preencher uma lacuna ao dedicar-se exclusivamente às instituições federais de ensino médio e técnico em todo o Brasil. A análise proposta pretende fornecer contribuições aplicáveis para políticas públicas, gestores educacionais e pesquisadores interessados na compreensão e mitigação da evasão escolar nesse contexto particular.

#### **1.4. Metodologia**

Para a realização deste trabalho, inicialmente, foi realizado um estudo bibliográfico com o objetivo de compreender os conceitos e contextos relacionados à evasão escolar e às técnicas de análise de dados. Diversas fontes, como livros, artigos científicos, dissertações e teses, foram incorporadas para estabelecer uma base teórica sólida sobre o tema. Essa revisão bibliográfica proporcionou informações fundamentais para a condução da pesquisa. Gil (2008) afirma que a pesquisa bibliográfica possibilita ao pesquisador compreender inúmeros fenômenos, em comparação com a pesquisa direta.

A segunda etapa consistiu na identificação e obtenção dos dados utilizados na pesquisa. Foram selecionados conjuntos de dados públicos fornecidos pelo Ministério da Educação (MEC)<sup>1</sup>, que contêm informações sobre o ensino médio e o ensino médio técnico, abrangendo aspectos acadêmicos, socioeconômicos e demográficos

---

<sup>1</sup> <https://dadosabertos.mec.gov.br/pnp>

dos estudantes. Essa fase foi essencial para garantir a qualidade e a representatividade dos dados utilizados na análise.

Na terceira fase, foi realizada a análise exploratória dos dados, com o intuito de identificar padrões e tendências relacionados à evasão escolar e ao perfil socioeconômico dos estudantes. Essa análise preliminar possibilitou a extração de variáveis relevantes para o desenvolvimento do modelo.

A quarta etapa envolveu a aplicação de técnicas de Aprendizado de Máquina no contexto educacional, visando desenvolver um modelo preditivo capaz de prever a evasão escolar com base nos dados disponíveis. Por fim, a quinta etapa consistiu na interpretação dos resultados do modelo, identificando as variáveis mais influentes e fornecendo sugestões para intervenções práticas.

Seguindo essas etapas, a pesquisa avançou de maneira estruturada e direcionada, culminando na construção de um modelo preditivo para a evasão escolar. Os resultados obtidos forneceram informações valiosas para orientar a tomada de decisões e implementar ações preventivas no contexto das instituições federais de ensino médio no Brasil.

Quanto à natureza da pesquisa, esta é caracterizada como aplicada, visando fornecer compreensões práticas para a prevenção da evasão escolar. A pesquisa possui uma abordagem predominantemente quantitativa, pois envolve a análise de dados estatísticos disponíveis, embora a abordagem qualitativa tenha sido incorporada na interpretação dos resultados e na contextualização das variáveis envolvidas.

Classificada como exploratória, a pesquisa busca elucidar fatores e padrões subjacentes à evasão escolar, contribuindo para um entendimento mais aprofundado desse fenômeno. Essa abordagem metodológica integrada visa proporcionar uma análise abrangente da evasão escolar, resultando na criação de modelos preditivos relevantes para a previsão e prevenção desse fenômeno nas instituições federais de ensino médio no Brasil.

### **1.5. Organização do Documento**

Os próximos capítulos deste documento estão organizados da seguinte forma: o Capítulo 2 fornece a base teórica essencial para a compreensão deste trabalho, explorando o panorama da evasão escolar no ensino médio e apresentando conceitos e técnicas de Ciência de Dados. No Capítulo 3, é realizada a análise exploratória dos

dados. O Capítulo 4 aborda a modelagem preditiva, enquanto o Capítulo 5 apresenta as conclusões e considerações finais do trabalho.

## 2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, são apresentados os fundamentos teóricos essenciais para a compreensão do tema abordado nesta pesquisa. A fundamentação está dividida em dois tópicos, que compreendem a Evasão escolar no ensino médio do Brasil (2.1) e a Ciência de Dados (2.2).

### 2.1. Evasão escolar

A evasão escolar é um desafio persistente e complexo que impacta diretamente o desenvolvimento educacional do país, afetando não apenas o progresso acadêmico dos estudantes, mas também gerando consequências sociais e econômicas significativas ao longo de muitos anos.

Segundo o Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)<sup>2</sup>, há uma diferença conceitual no termo técnico entre evasão e abandono escolar. “Abandono quer dizer que o aluno deixa a escola num ano, mas retorna no ano seguinte” (INEP, 2010). Já a evasão refere-se aos alunos que deixam de frequentar o ambiente escolar e não voltam mais para o sistema. “A evasão, de forma clássica, consiste no ato ou processo de evadir, de fugir, de escapar ou esquivar-se dos compromissos assumidos ou por vir a assumir” (ANUTO, 2013, p. 19).

Segundo Ferreira e Oliveira (2020), a evasão escolar é um fenômeno complexo com múltiplos fatores, capaz de ocasionar danos significativos não apenas ao aluno, mas também à sociedade como um todo. Isso ocorre porque o estudante que abandona a escola pode se tornar um membro à margem, excluído de uma sociedade letrada e, conseqüentemente, distante da compreensão plena da realidade.

A evasão escolar não se limita às questões internas da escola; grande parte das causas desse fenômeno reside fora do ambiente escolar. Frequentemente, essas razões estão relacionadas a problemas econômicos que envolvem o Estado ou a desafios estruturais na dinâmica familiar (FERREIRA; OLIVEIRA, 2020). Neste contexto, o debate sobre o papel da família e da escola tem se tornado central, na busca de soluções viáveis para a plena jornada escolar do aluno. A legislação brasileira compreende que a responsabilidade do percurso socioeducacional do aluno

---

<sup>2</sup> <https://www.gov.br/inep/pt-br>

é dos pais e do Estado. A Lei de Diretrizes e Bases da Educação - LDB é bastante clara a esse respeito:

“Art. 2º A educação, dever da família e do Estado, inspirada nos princípios de liberdade e nos ideais de solidariedade humana, tem por finalidade o pleno desenvolvimento do educando, seu preparo para o exercício da cidadania e sua qualificação para o trabalho”. (BRASIL, 1996. P. 01)

O que se percebe é o não cumprimento desses princípios, principalmente quando são elencados e levantados os motivos para a evasão escolar no Brasil. Observa-se que a evasão escolar tem ganhado cada vez mais destaque nas conversas e reflexões conduzidas pelo Estado e pela sociedade civil. Isso é especialmente evidente nas discussões promovidas por organizações e movimentos ligados à educação, tanto no contexto da pesquisa científica quanto no âmbito das políticas públicas.

#### 2.1.1. Panorama atual da evasão escolar no Brasil

O Congresso Brasileiro, em 2009, promoveu uma alteração significativa na Constituição Federal por meio da Emenda Constitucional Nº 59 (EC 59). Essa emenda reformulou o artigo 208 da Constituição Brasileira, estabelecendo a obrigatoriedade do ensino no país para a faixa etária dos 4 aos 17 anos. Tal modificação garantiu, de maneira jurídica, o acesso universal à educação infantil e ao ensino para os jovens de até 17 anos (PEREIRA, 2022).

O Plano Nacional de Educação (PNE), fundamentado no artigo 208 da Constituição Federal, estabeleceu:

“Universalizar, até 2016, o atendimento escolar para toda a população de quinze a dezessete anos e elevar, até o final do período de vigência deste PNE, a taxa líquida de matrículas no ensino médio para oitenta e cinco por cento”. (BRASIL, 2014. p. 33)

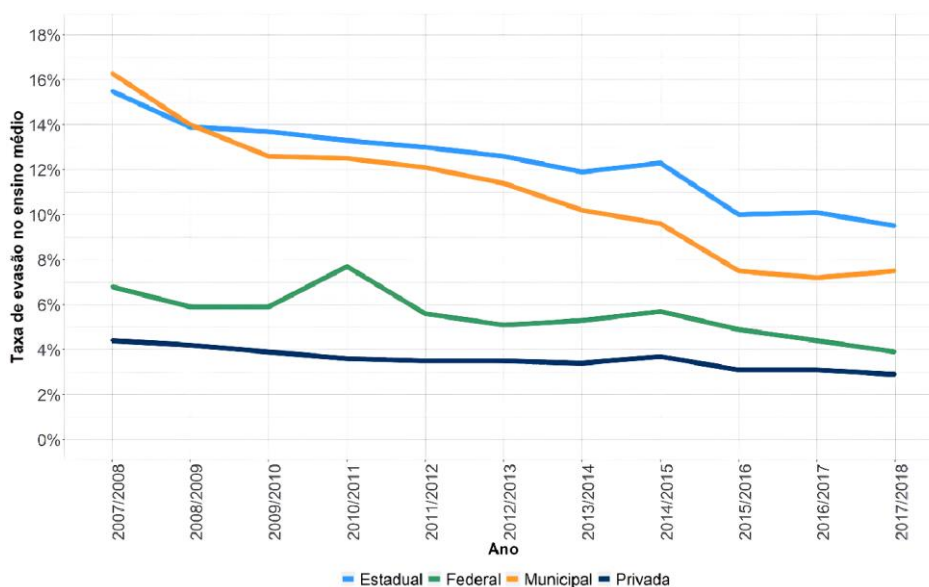
O PNE não apenas propôs assegurar a permanência dos jovens na escola, mas também visou corrigir a defasagem idade-série, buscando elevar a proporção de estudantes de 15 a 17 anos que frequentam a etapa adequada para sua faixa etária, ou seja, o ensino médio (PEREIRA, 2022).

No entanto, uma década após a aprovação da Emenda Constitucional, o Brasil ainda não conseguiu atingir a universalização do ensino médio ou a obrigatoriedade da educação até os 17 anos. A Emenda Constitucional Nº 59 mostrou-se relativamente eficaz em ampliar o acesso à pré-escola para crianças de 4 e 5 anos.

Conforme os dados da Pesquisa Nacional por Amostra de Domicílios Contínua (PNADC) de 2019, 94,1% das crianças nessa faixa etária estavam matriculadas na pré-escola, em comparação com 75% em 2009. Contudo, entre os jovens, o progresso tem sido notavelmente mais lento. Em 2019, um em cada dez jovens de 15 a 17 anos estava fora da escola. A proporção de jovens de 15 anos fora da escola era de 4,6%, aos 16 anos era de 9%, e aos 17 anos atingia 13,4%. A taxa líquida de matrícula no ensino médio em 2020 era de 75,4%, distante da meta estabelecida pelo Plano Nacional de Educação (PNE) de 85% até o ano de 2024 (PEREIRA, 2022).

Os resultados apresentados no trabalho de Pereira (2022) indicam uma necessidade urgente de se abordar a evasão escolar, especialmente nas instituições federais de ensino médio. De acordo com os dados mostrados na Figura 2, as taxas de evasão no ensino federal apresentaram uma redução, embora positiva, caindo de 7% para 5% entre 2007/2008 e 2017/2018. Essa diminuição, destaca a importância de continuar a investigação sobre os fatores que ainda contribuem para a evasão escolar nessas instituições.

Figura 1 - Evolução da taxa de evasão escolar no ensino médio – Brasil.



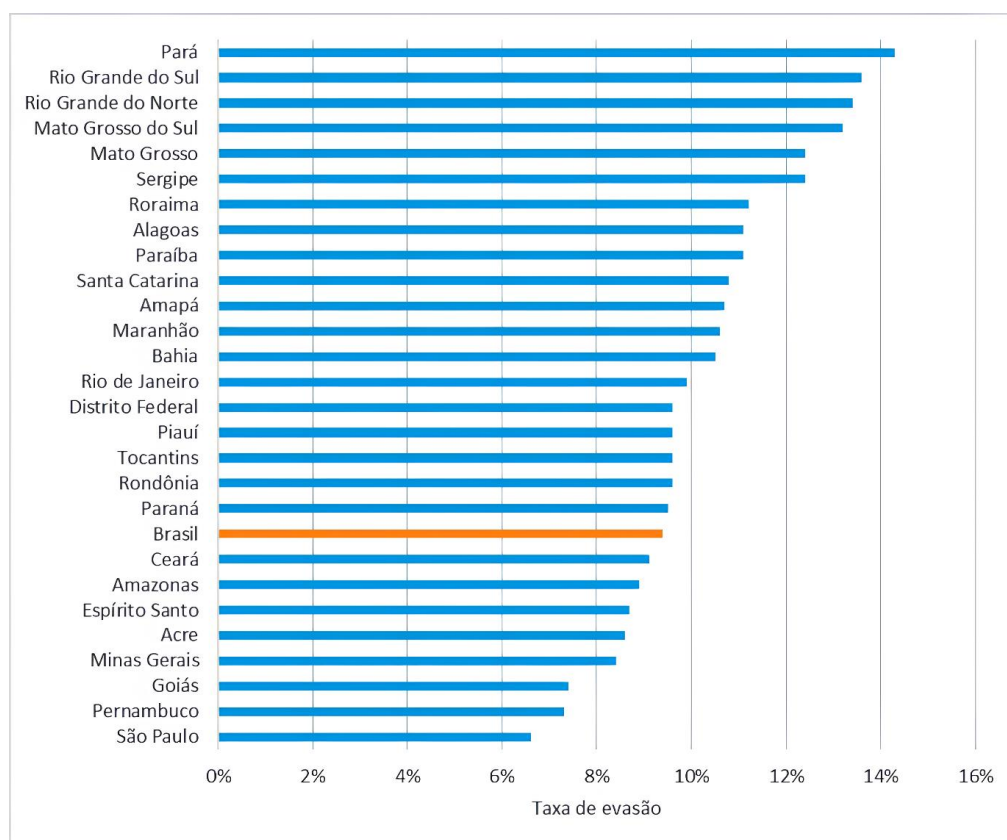
Fonte: Pereira, 2022.

Apesar da taxa de evasão continuar alta, é notável a diferença entre as taxas de evasão nas redes pública e privada. Isso aponta para desafios significativos na mobilidade educacional, sugerindo indiretamente que crianças de ambientes familiares mais desfavorecidos são mais propensas a abandonar e evadir-se do sistema educacional (FERREIRA, S. G.; RIBEIRO, G. e TAFNER, P., 2022).

As taxas de evasão escolar no ensino médio variam significativamente entre os estados brasileiros. Entre 2017 e 2018, os estados do Pará e Rio Grande do Sul registraram as maiores médias de evasão, com índices superiores a 14% ao longo das três séries do ensino médio. Em seguida, estados como Rio Grande do Norte, Mato Grosso do Sul, Mato Grosso e Sergipe também apresentaram taxas acima da média nacional. Por outro lado, São Paulo, Pernambuco e Goiás se destacaram por suas taxas de evasão mais baixas, todas inferiores a 8%. São Paulo, em especial, teve a menor média de evasão, com apenas 6,6% nesse período.

Apesar dessas variações, é importante ressaltar que mais estados apresentam índices de evasão superiores à média nacional do que abaixo dela, evidenciando um desafio significativo para a educação pública. Não há um padrão regional claro: por exemplo, no Nordeste, enquanto estados como Sergipe e Alagoas superam a média nacional, outros, como Ceará e Pernambuco, ficam abaixo. Isso demonstra que as causas da evasão escolar são diversas e podem variar dentro de uma mesma região. Esses dados ressaltam a necessidade de políticas públicas localizadas e adaptadas às particularidades de cada estado.

Figura 2 - Taxa de evasão no ensino médio – Brasil, rede pública, 2017/2018.



Fonte: Pereira, 2022.



### 2.1.2. Causas da evasão escolar no ensino médio

O Ensino Médio é a última etapa da Educação Básica, desempenhando um papel crucial na formação integral dos estudantes e na preparação para os desafios acadêmicos e profissionais. No entanto, as disparidades persistem nesse nível de ensino, refletindo-se nas taxas de abandono e evasão, especialmente entre as redes pública e privada.

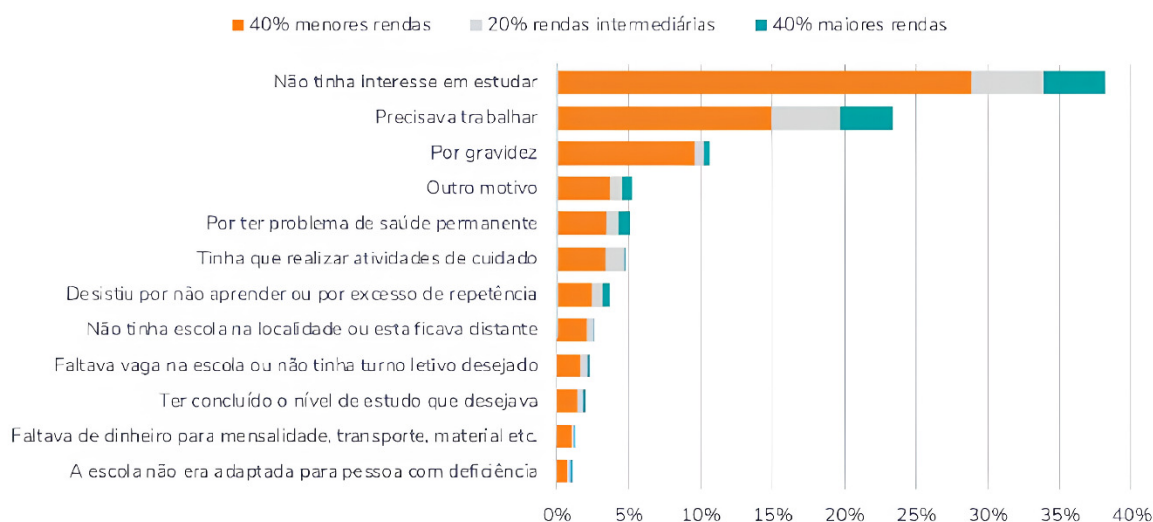
A interrupção permanente do vínculo com a escola é um fenômeno complexo, cumulativo e frequentemente desencadeado por diversos fatores. No entanto, é possível monitorar indicadores associados à evasão escolar, tanto aqueles que precedem, como o desempenho acadêmico, quanto os que ocorrem simultaneamente, como o atraso escolar, proporcionando uma maneira de rastrear os potenciais riscos (FERREIRA, S. G.; RIBEIRO, G.; TAFNER, P., 2022).

Diversas evidências destacam a importância crucial da educação em várias dimensões econômicas e sociais no Brasil. Estudos indicam que um nível educacional mais elevado está associado a salários mais altos, uma menor propensão ao envolvimento em atividades criminosas, melhorias na saúde e uma redução nas taxas de desemprego. Além disso, em termos nacionais, uma população mais educada contribui para um crescimento econômico mais robusto, impulsiona a produtividade das empresas e amplifica os efeitos da globalização (SAMPAIO, 1991).

Moraes e Linhares (1982) demonstram que a evasão escolar está relacionada a diversos fatores, como a repetência, renda familiar, gravidez, falta de incentivo da família, necessidade dos alunos de trabalharem, baixa autoestima devido à dificuldade no aprendizado, desinteresse geral, desestruturação familiar e outros motivos.

Assim, reconhecer a complexidade da evasão escolar é crucial, especialmente do ponto de vista do gestor público que se preocupa em mitigar o problema. Um aspecto relevante nesse contexto é compreender a motivação principal citada pelos jovens que optam pela evasão. Esse entendimento pode fornecer percepções valiosas para identificar desafios específicos e ajudar na busca por soluções adequadas. Entre os jovens que abandonaram os estudos e têm entre 15 e 19 anos, a principal razão mencionada para a evasão é a falta de interesse pela escola, conforme relatado por 38,1% dos entrevistados (Figura 4). Em segundo lugar, os entrevistados apontam a necessidade de ingressar no mercado de trabalho, com aproximadamente 23% (FERREIRA; RIBEIRO; TAFNER, 2022).

Figura 3 - Principal motivo declarado por adolescentes e jovens de 15 a 19 anos para terem saído da escola – Brasil, 2019.



Fonte: Instituto Mobilidade e Desenvolvimento Social (2022).

É importante notar, entretanto, um significativo contraste na hierarquização dos motivos. Primeiramente, a grande maioria dos jovens entre 15 e 19 anos que abandonaram a escola sem completar o ensino básico residem em domicílios que pertencem à faixa dos 40% com as menores rendas familiares. Em segundo lugar, destaca-se que a falta de interesse em estudar é mais preponderante, do ponto de vista relativo, do que a necessidade de trabalhar, quando comparado aos demais jovens. Em terceiro lugar, a gravidez é o terceiro motivo mais citado para jovens provenientes dos domicílios mais pobres, enquanto não figura entre os cinco motivos mais importantes para os demais jovens (FERREIRA, S. G.; RIBEIRO, G. e TAFNER, P., 2022).

Diversos estudos ressaltam outros fatores sociais que desempenham um papel significativo na evasão escolar, tais como desestruturação familiar, as políticas governamentais, o desemprego, a desnutrição, a dinâmica escolar e até mesmo as características individuais dos discentes.

Dessa forma, entender o panorama da evasão escolar é uma ação inicial essencial para a formulação de estratégias educativas mais direcionadas e eficazes. Ao examinar os elementos que desempenham um papel na decisão de evasão, torna-se viável identificar áreas que requerem intervenção. Isto permite a criação de medidas preventivas e corretivas que visam diminuir a frequência de alunos evadirem a escola antes da formatura.

Intervenções eficazes e políticas públicas voltadas para a educação são cruciais para reverter essa tendência preocupante. Ao abordar esses pontos, é possível construir uma narrativa mais completa e contextualizada sobre a evasão escolar no ensino médio brasileiro, fornecendo análises importantes para a compreensão do problema e o desenvolvimento de estratégias de enfrentamento.

## **2.2. Ciência de Dados**

A Ciência de Dados (CD) refere-se a uma área emergente de trabalho que aborda a coleta, preparação, análise, visualização, gerenciamento e preservação de grandes conjuntos de informações (AALST, 2016). Embora o termo pareça associado a áreas como bancos de dados e ciência da computação, é importante ressaltar que são necessárias diversas habilidades, incluindo aquelas que não são estritamente matemáticas (STANTON, 2012).

Soares (2020) destaca que a CD é um campo de estudo que lida tanto com dados estruturados (tabelas) quanto com dados não estruturados (textos, imagens e sons). Isso inclui processos relacionados à limpeza, preparação e análise final dos dados. De forma simplificada, pode-se considerar a Ciência de Dados como uma coleção de várias técnicas, métodos e modelos computacionais e estatísticos utilizados para extrair informações e percepções dos dados, visando auxiliar na tomada de decisão. A CD é responsável por extrair informações úteis de vastas bases de dados complexas, dinâmicas, heterogêneas e distribuídas (BUGNION; MANIVANNAN; NICOLAS, 2017).

A CD é uma disciplina dedicada à extração de conhecimento e compreensões significativas de conjuntos de dados complexos. Sua abordagem abrange diversas técnicas e métodos, combinando elementos de estatística, matemática, programação e conhecimento específico do domínio. O principal objetivo da Ciência de Dados é transformar dados brutos em informações úteis para a tomada de decisões.

No contexto das habilidades necessárias para atuar na área de Ciência de Dados, Stanton (2012) destaca competências não matemáticas que desempenham um papel crucial. Comunicação, análise de dados e raciocínio são habilidades fundamentais para profissionais nesse campo dinâmico. Amaral (2016, p. 5) complementa, afirmando que a Ciência de Dados é "composta por várias outras ciências, modelos, tecnologias, processos e procedimentos relacionados ao dado", criando relações interdisciplinares na área.

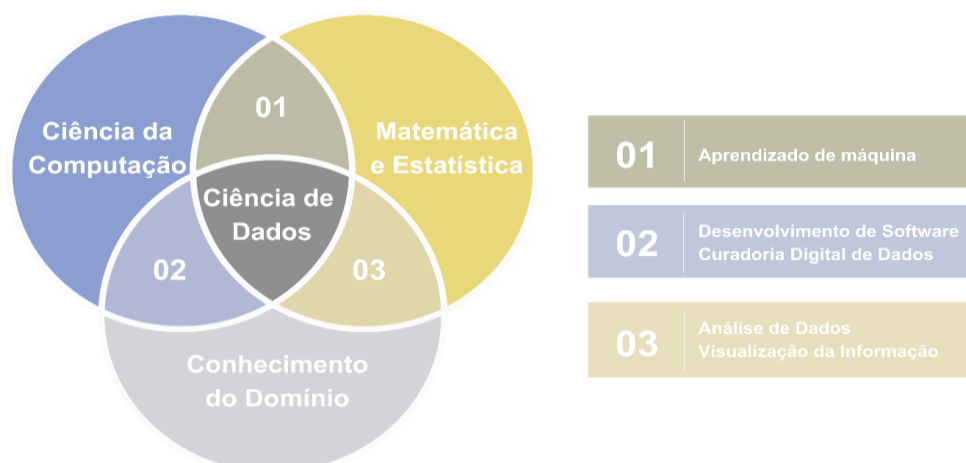
Os domínios do conhecimento na Ciência de Dados, conforme abordado por Rautenberg e Carmo (2019), são:

- **ciência da computação:** os especialistas devem apresentar habilidades em programação, manipulação de dados e desenvolvimento de software, dado o caráter computacional das operações. Além disso, os dados são centralmente armazenados, manipulados e transmitidos por meio de computadores. Nesse cenário, ambientes computacionais destinados ao desenvolvimento de software surgem como ferramentas indispensáveis para implementar algoritmos de aprendizado de máquina e criar interfaces para visualização da informação. A habilidade de se utilizar eficazmente essas tecnologias é imperativa para acessar e transformar dados, possibilitando abstrair e representar informações úteis;
- **matemática e estatística:** é indispensável para compreender algoritmos de aprendizado de máquina e realizar análises de dados de maneira estatisticamente robusta;
- **conhecimento do domínio:** a compreensão aprofundada do contexto do problema é crucial, demandando familiaridade com os cenários específicos. Essa habilidade permite a formulação de hipóteses relevantes e a tradução eficaz dos resultados em soluções aplicáveis, desempenhando um papel fundamental no processo de tomada de decisão.

A harmonia entre esses domínios capacita o cientista de dados a abordar desafios desde a manipulação técnica dos dados até a interpretação significativa no contexto de sua aplicação.

A Figura 5, ilustra a interação dos três domínios de conhecimento essenciais para atuar na Ciência de Dados: Ciência da Computação, Matemática e Estatística e Conhecimento do Domínio.

Figura 4 - Interdisciplinaridade da Ciência de Dados.

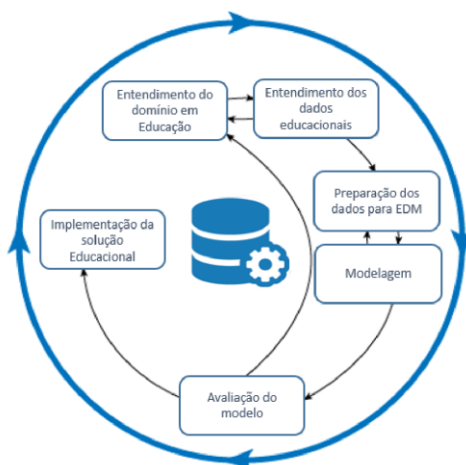


Fonte: Adaptado de (Rautenberg e Carmo, 2019).

Para a efetiva aplicação da Ciência de Dados, é crucial compreender como se desenrola o fluxo de trabalho em um projeto dessa natureza. A Figura 5 ilustra as etapas desse fluxo, voltado para a área educacional, que podem ser executadas de forma iterativa. A seguir, cada uma dessas etapas é detalhada, conforme abordado por Ramon (2020).

O método CRISP-EDM (acrônimo de *Cross Industry Standard Process for Educational Data Mining*) (RAMOS et al., 2020), uma versão adaptada do método CRISP-DM (*Cross-Industry Standard Process for Data Mining*) (WIRTH; JOCHEN, 2000) para uso em contextos educacionais, contém técnicas e métodos adequados à área educacional analisada. Algumas etapas podem ser divididas em fases menores para um melhor desenvolvimento.

Figura 5 - Fases do modelo do CRISP-EDM.



Fonte: adaptado de Shearer (2000 *apud* RAMOS et al., 2020)

As etapas do CRISP-EDM são brevemente explicadas a seguir (RAMOS *et al.*, 2020):

**1. compreensão do domínio educacional:** a primeira etapa envolve entender o problema educacional e o contexto em que ele se insere. Isso inclui a identificação clara do problema ou questão a ser investigada, os objetivos a serem alcançados e os desafios educacionais que se deseja enfrentar. Um entendimento profundo do domínio facilita a escolha adequada das variáveis e dos métodos a serem aplicados.

**2. compreensão dos dados educacionais:** nesta fase, realiza-se a coleta, análise e entendimento dos dados educacionais disponíveis. Esses dados podem variar conforme a área de estudo e podem incluir informações sobre desempenho acadêmico, engajamento estudantil, participação nas aulas ou características demográficas. O conhecimento da origem e da estrutura dos dados é fundamental para direcionar o projeto de maneira eficiente.

**3. preparação dos dados:** a preparação dos dados abrange a organização e limpeza dos dados coletados. Isso inclui a remoção de valores faltantes ou inconsistentes, a transformação dos dados em formatos adequados e a criação de novos atributos, se necessário. A qualidade e adequação dos dados preparados impactam diretamente nos resultados dos modelos.

**4. modelagem:** nesta fase, diferentes algoritmos de machine learning são aplicados aos dados educacionais. São testados métodos como regressão, árvores de decisão, redes neurais, entre outros, para identificar padrões e prever resultados, como o risco de evasão ou o desempenho futuro de alunos. A escolha do algoritmo depende da natureza do problema e dos dados.

**5. avaliação do modelo:** após a modelagem, os resultados obtidos são revisados e avaliados quanto à sua relevância e aplicabilidade no contexto educacional. A avaliação inclui a verificação de como os resultados respondem às questões originalmente propostas e o impacto que esses resultados podem ter nas práticas educacionais ou políticas institucionais.

**6. implementação das soluções:** a última etapa envolve a implementação dos resultados em um ambiente educacional real. Isso pode significar a integração de sistemas preditivos em plataformas de gestão escolar, relatórios para gestores e professores, ou a criação de dashboards para monitorar o progresso de alunos. O objetivo é que os insights gerados pelos modelos possam ser aplicados para melhorar o processo de tomada de decisão na educação.

### 2.2.1. Análise exploratória de dados

A Análise Exploratória de Dados (AED) desempenha um papel crucial na Ciência de Dados, consistindo na exploração inicial e visualização de dados para identificar padrões, tendências, *outliers* e relações entre variáveis (TUKEY, 1977). Iniciada por John Tukey, um renomado estatístico, em 1977, a AED busca ampliar o entendimento do pesquisador sobre uma população com base em uma amostra (LOPES *et al.*, 2019).

Ainda segundo o autor, a AED é um conjunto de métodos para coletar, explorar, descrever e interpretar conjuntos de dados numéricos. Esses métodos visam obter a maior quantidade possível de informação dos dados, indicando modelos plausíveis para análise posterior, como a inferência estatística (MEDRI, 2011).

Wickham *et al.* (2023) definem a AED como uma metodologia que envolve a visualização, transformação e modelagem de dados para identificar padrões de variação e covariação entre as variáveis. Essa abordagem típica engloba a caracterização de diversas variáveis, comparando suas propriedades ou comportamentos (PEARSON, 2018).

A Análise Exploratória de Dados (AED) envolve conceitos-chave que incluem revelação (visualização de dados), resíduos (diferença entre valores observados e esperados), re-expressão (transformações para aprimorar a representatividade dos dados) e resistência (capacidade de evitar influências indevidas de valores discrepantes).

A sequência recomendada por Pearson (2018) para explorar um novo conjunto de dados envolve:

- avaliar as características gerais do conjunto de dados, como o número de registros e variáveis, os nomes das variáveis e os tipos de variáveis;
- examinar as estatísticas descritivas para cada variável;
- onde possível, examinar visualizações exploratórias, como gráficos e histogramas;
- aplicar procedimentos para procurar por anomalias;
- analisar as relações entre as variáveis com técnicas multivariadas;
- resumir os resultados em um dicionário de dados, incluindo as informações sobre o conjunto de dados e suas características relevantes.

De modo geral, a AED busca identificar padrões, estrutura, *outliers* ou relações inesperadas entre variáveis, utilizando ferramentas visuais e estatísticas descritivas.

### 2.2.2. Pré-processamento de dados

O pré-processamento de dados desempenha um papel fundamental na Ciência de Dados, envolvendo a limpeza e preparação dos dados para análises subsequentes. Ela representa uma etapa essencial que antecede a implementação de técnicas mais avançadas, visando garantir a qualidade e consistência dos dados, preparando-os adequadamente.

Os dados do mundo real frequentemente apresentam características como incompletude, inconsistência, ruído e grande volume. Para enfrentar esses desafios, diversas técnicas de pré-processamento são aplicadas, incluindo limpeza, integração, redução e transformação de dados (HAN *et al.*, 2011).

Um pré-processamento eficaz contribui significativamente para a qualidade e confiabilidade dos resultados obtidos durante a análise de dados. As etapas delineadas pelos autores compreendem:

- **limpeza de dados:** essa etapa visa lidar com dados inconsistentes ou errôneos. A limpeza de dados envolve geralmente a identificação e retirada de dados duplicados, o preenchimento de valores ausentes, a remoção de ruídos e *outliers*. Essas atividades podem ser realizadas por meio de técnicas como suavização de dados, agregação de dados e detecção de valores discrepantes. Essa etapa pode ajudar a melhorar a qualidade dos dados e aumentar a confiabilidade e a precisão da análise de dados. Ela também pode ajudar a reduzir custos e tempo de processamento, tornando a análise de dados mais eficaz e eficiente;
- **integração de dados:** a integração de dados é o processo de combinar dados de várias fontes em um único conjunto de dados que possa ser usado para análise. O objetivo da integração de dados é fornecer uma visão completa e consistente dos dados para dar suporte à tomada de decisões em uma organização. Uma técnica comum de integração de dados é a construção de um data *warehouse*, que é um repositório centralizado de dados integrados de várias fontes que suporta a análise de dados eficiente e eficaz. A falta de



integração de dados pode levar a redundâncias de dados, inconsistências de dados e dificuldades em acessar e analisar dados;

- **redução de dados:** o objetivo da redução de dados é reduzir o tamanho do conjunto de dados sem perder informações importantes. Existem várias técnicas de redução de dados disponíveis, incluindo a eliminação de dados redundantes, a amostragem e a agregação. A redução dos dados pode ajudar a economizar tempo e recursos de processamento de dados;
- **transformação de dados:** é o processo de modificação dos dados para torná-los adequados para a análise. Isso pode envolver, por exemplo, a normalização de dados para garantir que os dados estejam na mesma escala. Também pode envolver o mapeamento de dados para reduzir o número de atributos ou agrupar dados em categorias. Em geral, o objetivo da transformação de dados é melhorar a qualidade dos dados e prepará-los para a análise, bem como para atender ao formato de dados de entrada para os algoritmos de AM. Essa etapa é crucial para garantir que os dados de entrada sejam precisos e adequados para análises subsequentes;

A importância do pré-processamento de dados é evidente na melhoria da qualidade, confiabilidade e eficácia da análise de dados. Em geral, realizar um pré-processamento eficaz é fundamental para obter resultados precisos e de alta qualidade na análise de dados, estabelecendo uma base sólida para compreensões valiosas (HAN; KAMBER; PEI, 2011).

### 2.2.3. Aprendizado de Máquina

O Aprendizado de Máquina (AM) é um campo que envolve o desenvolvimento de algoritmos e técnicas que permitem que sistemas computacionais aprendam a partir de dados, sem necessitar de programação explícita (MITCHELL, 1997). Esse processo de aprendizado automático é aplicado em diversas áreas, como reconhecimento de padrões, análise preditiva e decisões automatizadas (BISHOP, 2006).

O aprendizado de máquina, subcampo da inteligência artificial, tem registrado um crescimento notável nos últimos anos, impulsionado por avanços significativos na capacidade computacional, no desenvolvimento de algoritmos sofisticados e na disponibilidade de conjuntos de dados variados.

Bishop (2006) descreve quatro tipos fundamentais de aprendizado de máquina:

- **aprendizado supervisionado:** neste tipo de aprendizado, o modelo recebe dados rotulados como entrada e aprende a mapear esses dados para suas respectivas saídas. O objetivo é aprender uma função que possa generalizar para novos dados não vistos;
- **aprendizado não supervisionado:** neste tipo de aprendizado, o modelo recebe dados não rotulados como entrada e tenta encontrar padrões e estruturas ocultas nos dados. O objetivo é aprender sobre a estrutura dos dados e, possivelmente, encontrar agrupamentos ou reduzir a dimensionalidade do espaço de variáveis;
- **aprendizado semi-supervisionado:** neste tipo de aprendizado, o modelo recebe tanto dados rotulados quanto não rotulados como entrada e tenta aprender uma função que possa generalizar para novos dados não vistos, aproveitando as informações contidas nos dados não rotulados. O objetivo é melhorar o desempenho do modelo em comparação com o aprendizado supervisionado usando apenas dados rotulados;
- **aprendizado por reforço:** neste tipo de aprendizado, o modelo aprende a tomar ações em um ambiente para maximizar uma recompensa. O objetivo é aprender uma política que maximize a recompensa ao longo do tempo.

Esses quatro tipos de aprendizado de máquina são essenciais para compreender e aplicar técnicas em diversos problemas em várias áreas do conhecimento.

No contexto da Ciência de Dados, o AM é uma ferramenta fundamental para desenvolver modelos preditivos, classificatórios ou de agrupamento, dependendo dos objetivos da análise. Neste estudo, aplicou-se o aprendizado supervisionado para classificar casos de evasão escolar. Essa escolha é justificada pela necessidade de se prever novas ocorrências com base em dados anteriores, permitindo que o modelo aprendesse a mapear dados rotulados para prever a evasão escolar.

### 2.3. Algoritmos de classificação

O Aprendizado Supervisionado, como já mencionado, oferece uma variedade de algoritmos capazes de realizar a classificação em conjuntos de dados. Nesta

seção, serão apresentados os principais algoritmos aplicados neste trabalho, entre os muitos disponíveis para essa finalidade.

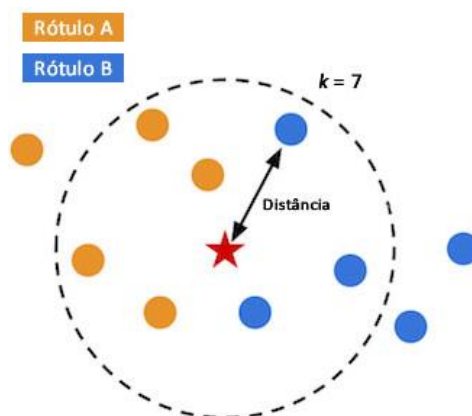
### 2.3.1. *K-Nearest Neighbors*

O *K-Nearest Neighbors* (KNN), ou K-Vizinhos mais próximos, é um algoritmo de aprendizado de máquina amplamente utilizado em problemas de classificação e regressão. Sua simplicidade o destaca, pois ele não exige pressupostos matemáticos complexos (COOMANS; MASSART, 1982). O KNN funciona armazenando o conjunto de dados de treinamento, e quando novos dados são inseridos, o algoritmo calcula a distância entre esses dados e os exemplos de treino. A classe dos novos dados é atribuída com base nos vizinhos mais próximos, ou seja, nas amostras que apresentam a menor distância em relação à nova entrada (ALTMAN, 1992).

O cálculo da distância é feito utilizando métricas como a Distância Euclidiana, Manhattan ou Minkowski (ESCOVEDO; KOSHIYAMA, 2020). O valor de  $k$ , que representa o número de vizinhos considerados, é definido pelo usuário. Em problemas de classificação, a nova amostra é atribuída à classe mais comum entre os  $k$  vizinhos, enquanto em problemas de regressão, é feita a média dos valores dos vizinhos.

A ideia principal do KNN é determinar o rótulo de classificação de uma amostra com base em amostras vizinhas de um conjunto de treinamento. A Figura abaixo exemplifica esse processo, onde uma nova amostra (representada por uma estrela) tem sua classe determinada a partir das sete amostras mais próximas ( $k = 7$ ), sendo 4 pertencentes à classe A (azul) e 3 à classe B (amarela). Assim, a nova amostra é classificada como pertencente à classe A (PACHECO, 2017).

Figura 6 - Exemplo de classificação do KNN com dois rótulos de classe e  $k = 7$ .



Fonte: Pacheco, 2017.

O KNN é considerado um algoritmo de treinamento baseado em instâncias, onde o processo de aprendizagem envolve apenas o armazenamento dos dados, e o processamento é feito no momento da classificação, sendo chamado de "preguiçoso" por adiar o cálculo até a fase de consulta (MITCHELL et al., 1997). Embora seja intuitivo e eficiente em tarefas de classificação, especialmente em cenários binários e sistemas de recomendação, ele pode ser computacionalmente custoso para grandes bases de dados devido ao cálculo repetido das distâncias.

### 2.3.2. Árvore de Decisão

As Árvores de Decisão são algoritmos de aprendizado de máquina amplamente utilizados em problemas de classificação e regressão, caracterizados por sua estrutura hierárquica composta por nós de decisão, ramos e folhas. Segundo Alpaydin (2020), a raiz da árvore se encontra no topo, onde cada nó representa uma decisão que direciona o fluxo do algoritmo, e os nós folha correspondem às classes finais atribuídas às entradas. O processo de decisão é baseado em testes aplicados em cada nó, que determinam o caminho a seguir até que um nó folha seja alcançado.

Esses modelos utilizam a estratégia "dividir para conquistar", decompondo problemas complexos em subproblemas mais simples (Gama, 2004). O critério para as partições é geralmente baseado no ganho de informação, sendo os atributos escolhidos aqueles que proporcionam maior discriminação (Onoda e Ebecken, 2001). Métodos comuns para determinar esses critérios incluem a entropia e o índice Gini, que ajudam a definir a utilidade de cada atributo para a classificação.

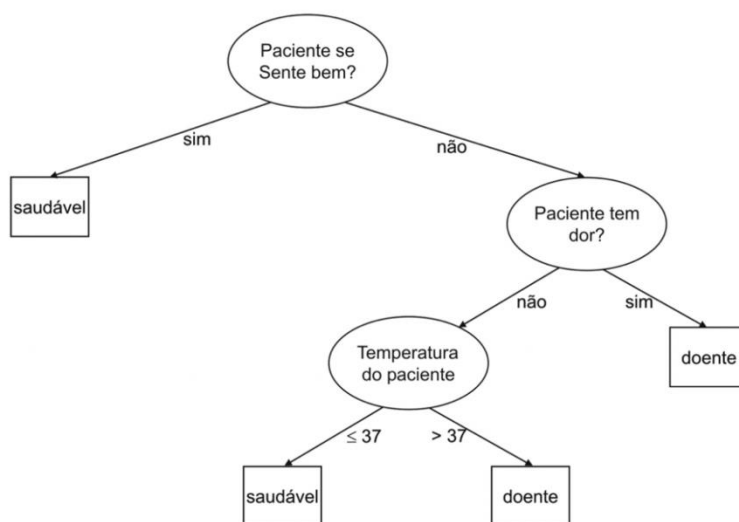
As Árvores de Decisão são intuitivas e de fácil interpretação, permitindo a extração de regras do tipo "se-então", que são compreensíveis para usuários não técnicos (Mitchell, 1997). Cada percurso da árvore, desde a raiz até uma folha, representa uma regra de classificação, onde cada folha corresponde a uma região do espaço de atributos, formando hiper-retângulos que não se sobrepõem (Gama, 2004). Além disso, a apresentação hierárquica dos atributos mais relevantes torna esse modelo uma ferramenta eficaz para a tomada de decisões, facilitando a identificação dos fatores que influenciam diretamente as escolhas (Garcia, 2000).

Segundo Monard e Baranauskas (2003, p.60), uma árvore de decisão é uma estrutura de dados recursiva, onde cada nó de decisão contém um teste sobre algum atributo. O resultado do teste direciona a uma nova subárvore, que segue a mesma estrutura. Crepaldi et al. (2010, p.3) complementam afirmando que o atributo mais

importante é apresentado como o primeiro nó, e os atributos menos relevantes aparecem nos nós subsequentes.

A principal vantagem dessa técnica é sua capacidade de selecionar os atributos mais relevantes para a tomada de decisão. Ao apresentar os atributos em ordem de importância, as Árvores de Decisão ajudam os usuários a identificar rapidamente os principais fatores que influenciam os resultados. A apresenta um exemplo de árvore de decisão.

Figura 7 - Árvore de Decisão.



Fonte: (Monard e Baranauskas, 2003).

### 2.3.3. Floresta Aleatória

A Floresta Aleatória (*Random Forest*) é um algoritmo de aprendizado de máquina desenvolvido por Breiman (2001). Ele combina diversas Árvores de Decisão, criadas a partir de amostras aleatórias de dados e subconjuntos de atributos, com o objetivo de melhorar a precisão das previsões e reduzir problemas como o *overfitting*<sup>3</sup>.

Cada árvore na floresta realiza uma classificação ou predição com base em seus dados específicos, e a decisão final do modelo é tomada por meio de uma votação majoritária entre as árvores (no caso de classificação) ou pela média das previsões (para regressão) (GÉRON, 2017). A aleatoriedade na escolha dos dados e

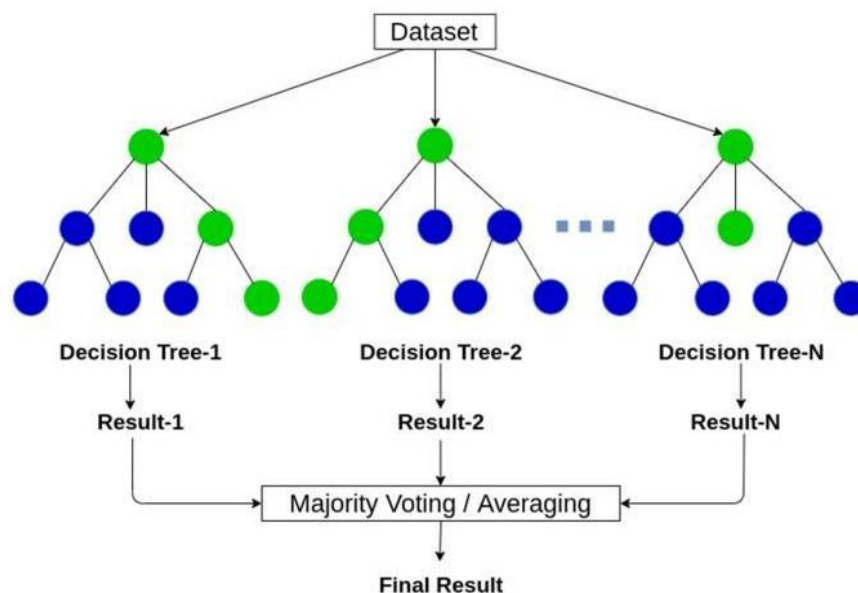
<sup>3</sup> Overfitting é um fenômeno que ocorre quando um modelo de machine learning se ajusta excessivamente aos dados de treinamento, capturando não apenas os padrões reais, mas também o ruído ou variações aleatórias. Isso resulta em um modelo com bom desempenho nos dados de treinamento, mas com baixa capacidade de generalização para novos dados, prejudicando sua eficácia em previsões reais.

atributos aumenta a diversidade entre as árvores, resultando em um modelo mais robusto e menos suscetível a ruídos e variações nos dados.

Uma das principais vantagens da Floresta Aleatória é sua capacidade de lidar com grandes volumes de dados e com a presença de atributos numéricos e categóricos. Além disso, o modelo se destaca por ser menos propenso ao sobreajuste, uma limitação comum das Árvores de Decisão individuais. Contudo, essa abordagem envolve maior complexidade computacional, já que várias árvores precisam ser geradas e avaliadas (HO, 1995). A Floresta Aleatória também facilita a análise de importância dos atributos, ajudando a identificar quais variáveis têm maior influência no resultado (GÉRON, 2017).

A Figura 7 exemplifica o funcionamento de uma Floresta Aleatória, onde várias árvores de decisão são criadas a partir de subconjuntos de dados e realizam a votação para determinar a classe final.

Figura 8 - Representação de uma Floresta Aleatória.



Fonte: (KHAN et al., 2021).

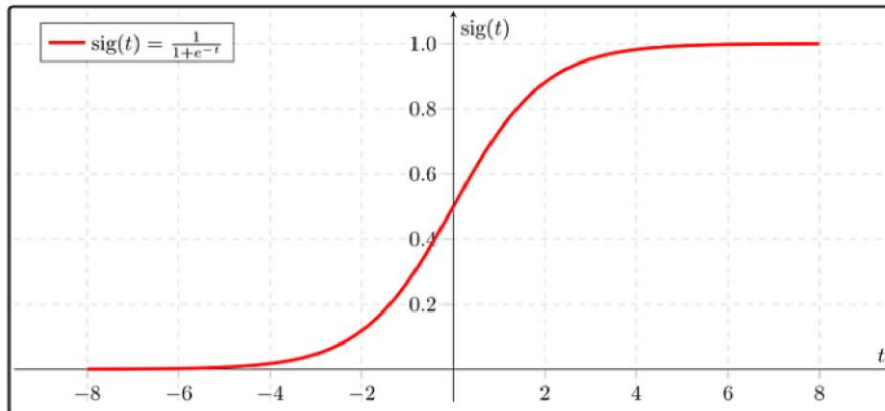
#### 2.3.4. Regressão Logística

A Regressão Logística (*Logistic Regression*) é uma técnica estatística amplamente utilizada para problemas de classificação binária, onde a variável dependente assume valores dicotômicos, como sucesso (1) e falha (0) (Hilbe, 2011).

O modelo de Regressão Logística utiliza a função *sigmóide* para converter as previsões em probabilidades, permitindo que as saídas do modelo sejam

interpretadas em um intervalo de 0 a 1. A partir dessa conversão, estabelece-se um limiar, comumente de 0,5, que determina a classificação: valores acima desse limite são atribuídos à classe 1 e valores abaixo à classe 0.

Figura 9 - Regressão Logística - Função Sigmoide.



Fonte: (SWAMINATHAN, 2018).

Dentre as suas aplicações, a Regressão Logística é frequentemente empregada em modelos de *Credit Scoring*, que classificam clientes como bons ou maus pagadores (Crook et al., 2007; Gouvêa, 2015). Uma das principais vantagens da RL é que ela não exige que as variáveis preditoras sigam uma distribuição normal, o que a torna mais robusta em situações onde a independência entre variáveis não é respeitada (Hair et al., 2009; Gouvêa, 2015). Além disso, permite a inclusão de variáveis categóricas (desde que transformadas em formato binário) e contínuas, ampliando a flexibilidade do modelo na análise de dados.

### 3. ANÁLISE EXPLORATÓRIA DOS DADOS

Este capítulo mostra a análise exploratória dos dados coletados para identificar variáveis mais importantes possam influenciar a evasão escolar nos cursos de ensino médio e técnico. São descritas as variáveis utilizadas e apresentadas representações gráficas que evidenciam as relações entre as principais variáveis em relação à evasão. Essa análise fornece uma base para a modelagem preditiva, destacando características que ajudam a identificar grupos vulneráveis e sugerir intervenções eficazes.

#### 3.1. Entendimento de dados

O conjunto de dados adotado para esta pesquisa foi extraído do censo escolar da educação básica, disponibilizado pelo MEC por meio da Plataforma Nilo Peçanha<sup>4</sup>. O repositório inclui informações abrangentes sobre os estudantes, como desempenho acadêmico, perfil socioeconômico e características demográficas, permitindo uma análise detalhada da evasão escolar.

O foco da análise realizada neste trabalho foi nas matrículas dos anos de 2019<sup>5</sup>, 2020<sup>6</sup> e 2021<sup>7</sup>, que abrangem 56 variáveis, sendo 23 qualitativas e 33 quantitativas. Esses anos foram selecionados devido à padronização das variáveis, que permitiu uma análise mais consistente, uma vez que anos anteriores não apresentavam as mesmas informações detalhadas, o que comprometia a comparabilidade. Além disso, os dados dos anos mais recentes, como 2022 e 2023, não estavam disponíveis no momento da pesquisa, sendo que o censo de 2022 foi incluído apenas em setembro de 2024. Essas variáveis foram categorizadas em três grandes grupos: características socioeconômicas, acadêmicas e demográficas dos estudantes. A análise permitiu uma visão detalhada dos fatores que podem influenciar a evasão escolar e serviu como base para a modelagem preditiva. Para uma melhor compreensão das variáveis utilizadas, um dicionário das principais variáveis é apresentado no Quadro 1.

---

<sup>4</sup> <https://dadosabertos.mec.gov.br/pnp>

<sup>5</sup> [https://dadosabertos.mec.gov.br/images/conteudo/pnp/2020/microdados\\_matriculas\\_2020.zip](https://dadosabertos.mec.gov.br/images/conteudo/pnp/2020/microdados_matriculas_2020.zip)

<sup>6</sup> [https://dadosabertos.mec.gov.br/images/conteudo/pnp/2021/microdados\\_matriculas\\_2021.zip](https://dadosabertos.mec.gov.br/images/conteudo/pnp/2021/microdados_matriculas_2021.zip)

<sup>7</sup> [https://dadosabertos.mec.gov.br/images/conteudo/pnp/2022/microdados\\_matriculas\\_2022.zip](https://dadosabertos.mec.gov.br/images/conteudo/pnp/2022/microdados_matriculas_2022.zip)



Quadro 1 - Dicionário das principais variáveis.

Variável	Descrição	Tipo	Subtipo
Carga horária	Carga horária do ciclo de matrícula.	Quantitativa	Discreta
Categoria da situação	Situações de matrícula: concluintes, em curso e evadidos.	Qualitativa	Nominal
Código da matrícula	Código da matrícula.	Quantitativa	Discreta
Cor/Raça	Cor/Raça do aluno.	Qualitativa	Nominal
Data de início do ciclo	Data de início do ciclo de matrícula.	Quantitativa	Discreta
Data de fim previsto do ciclo	Data prevista para o final do ciclo de matrícula.	Quantitativa	Discreta
Faixa etária	Agrupamento baseado na idade dos estudantes.	Qualitativa	Ordinal
Fator esforço curso	Ajusta a contagem de matrículas-equivalentes para cursos que demandem, para o desenvolvimento de suas atividades, uma menor relação aluno por professor.	Quantitativa	Contínua
Idade	Idade do estudante.	Quantitativa	Discreta
Instituição	Sigla da Instituição.	Qualitativa	Nominal
Mês de ocorrência da situação	Mês/Ano em que a situação da matrícula efetivamente mudou.	Quantitativa	Discreta
Região	Região Geográfica do país onde está instalada a instituição.	Qualitativa	Nominal
Renda familiar	Faixa de renda per capita familiar do aluno.	Qualitativa	Ordinal
Sexo	Informa o sexo do estudante.	Qualitativa	Nominal
Turno	Período de tempo determinado em que o aluno cursa a maior parte das aulas.	Qualitativa	Nominal
UF	Unidade da Federação onde está instalada a instituição.	Qualitativa	Nominal
Unidade de ensino	Nome da unidade de ensino a qual a matrícula está vinculada.	Qualitativa	Nominal

Fonte: Autor, 2024.

### 3.2. Etapas da análise exploratória de dados

Nesta seção, são apresentadas as etapas da análise exploratória, visando garantir a qualidade das informações coletadas. O processo inclui a limpeza de dados e o tratamento de valores ausentes, assegurando uniformidade nos conjuntos de 2019 a 2021. A análise enfoca a distribuição de variáveis como idade, situação de matrícula e renda familiar, identificando padrões e vulnerabilidades que fundamentam a modelagem preditiva e as intervenções direcionadas.

### 3.2.1. Limpeza de dados e tratamento de valores ausentes

Para garantir a qualidade e consistência dos dados utilizados na análise, foi realizado um processo de limpeza de dados e tratamento de valores ausentes. Foram trabalhados três conjuntos de dados referentes aos anos de 2019, 2020 e 2021, que, embora compartilhassem variáveis semelhantes, apresentavam inconsistências de nomenclatura, valores duplicados e dados ausentes que necessitavam de ajustes, tais como:

- **Padronização de colunas:** os nomes e categorias de colunas, como "Cor/Raça" e "Renda Familiar", foram uniformizados para garantir a consistência entre os anos de 2019, 2020 e 2021;
- **Remoção de dados irrelevantes:** colunas que não eram necessárias para a análise, como "Código da Unidade de Ensino" e "Fonte de Financiamento", foram excluídas;
- **Tratamento de valores ausentes:** valores ausentes em colunas categóricas foram substituídos por "Não declarada", ou registros foram removidos em situações que comprometiam a análise;
- **Remoção de duplicatas:** registros duplicados foram eliminados com base no "Código da Matrícula";
- **Ajustes de formato de data:** o formato das variáveis de data foi padronizado para uma estrutura compatível.

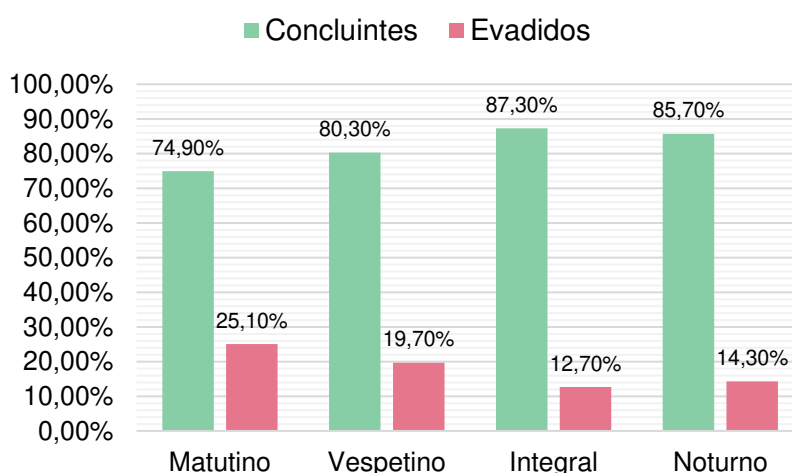
Esse processo foi aplicado tanto aos dados do ensino médio quanto aos do ensino médio técnico, assegurando a consistência e a integridade dos dados para a modelagem preditiva.

### 3.2.2. Análise de distribuição das variáveis do ensino médio

Na análise das variáveis do ensino médio, foram considerados aspectos como idade, situação de matrícula, turno do curso, renda familiar e cor/raça. A Figura 10 analisa a distribuição das taxas de evasão escolar de acordo com o turno de matrícula dos alunos. O turno matutino concentra o maior percentual de evasão. Essa concentração pode ser atribuída a fatores como a necessidade de conciliar estudos com atividades extracurriculares ou familiares, que tendem a ser mais desafiadoras durante o período da manhã. O turno vespertino apresenta uma evasão menos acentuada, enquanto o turno integral possui a menor taxa de abandono, superando

até mesmo o noturno. Esse dado sugere que os alunos matriculados em período integral podem estar mais engajados com a rotina escolar, o que pode contribuir para sua maior permanência. Políticas de flexibilização de horários e suporte escolar no turno matutino podem ser eficazes na redução da evasão, mas é importante também analisar os benefícios que a estrutura do turno integral oferece em termos de retenção.

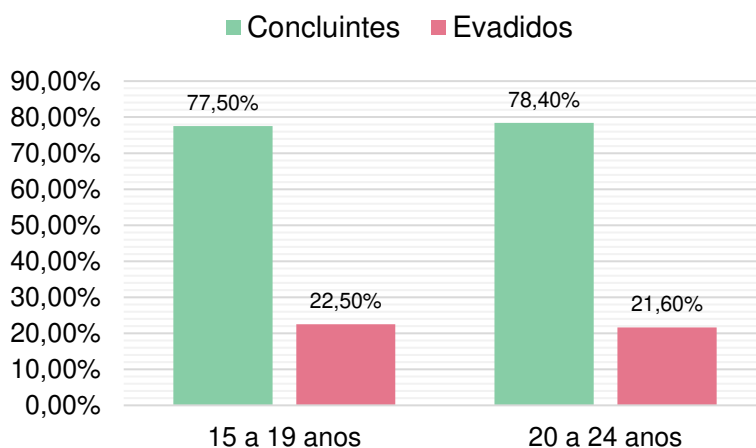
Figura 10 - Evasão por turno no ensino médio.



Fonte: Autor, 2024.

A Figura 11 apresenta a relação entre a idade dos alunos e as taxas de evasão escolar no ensino médio. Observa-se que as maiores taxas de evasão estão concentradas entre os 15 e 16 anos, fase inicial do ensino médio, sugerindo que os alunos mais jovens enfrentam maiores desafios de adaptação ao ambiente escolar.

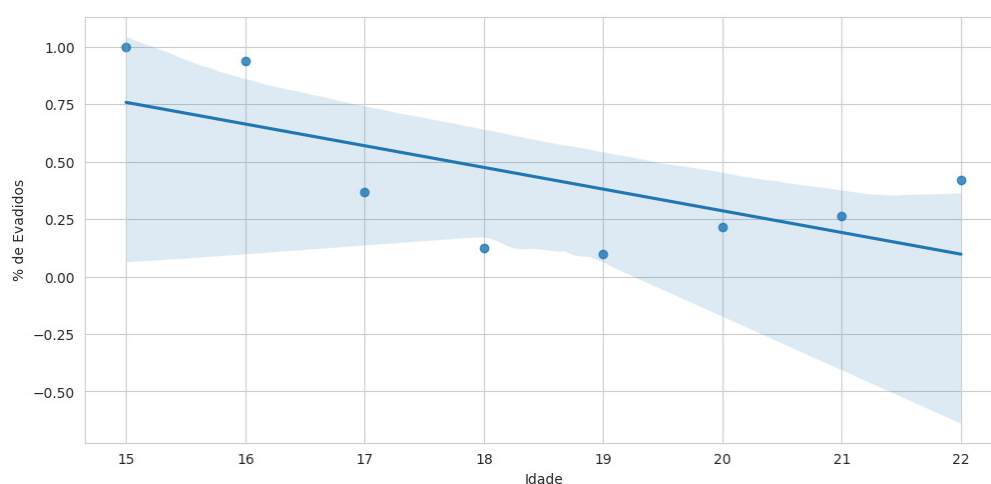
Figura 11 - Evasão por faixa etária ensino médio.



Fonte: Autor, 2024.

A Figura 12 ilustra uma tendência acentuada de redução das taxas de evasão à medida que a idade avança, evidenciada pela inclinação negativa da linha de tendência. As maiores taxas de evasão são observadas nos 15 e 16 anos, período correspondente ao início do ensino médio, o que pode indicar desafios de adaptação. A partir dos 17 anos, a evasão diminui significativamente, sugerindo que aqueles que permanecem até essa idade têm maior probabilidade de concluir o curso. Contudo, a variação dos dados sugere que a idade não é o único fator determinante para a evasão, exigindo a consideração de outras variáveis.

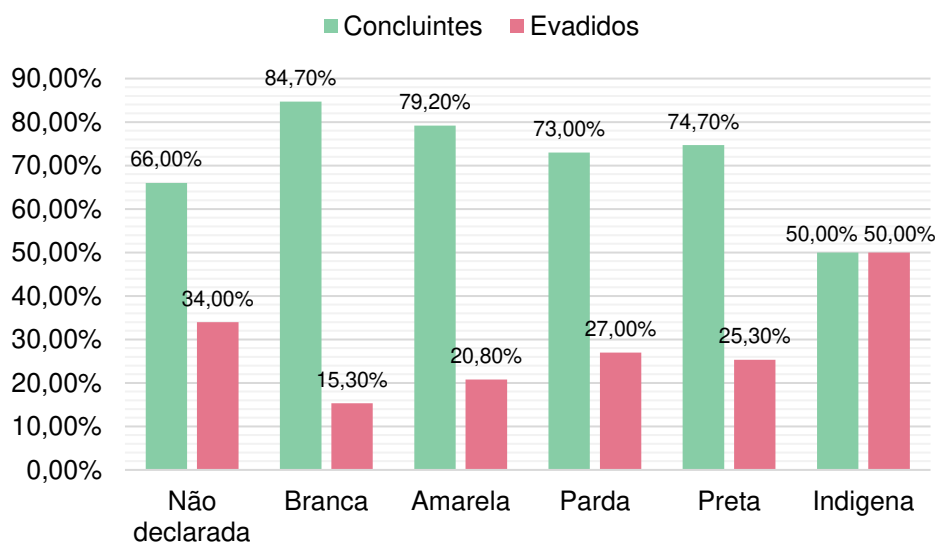
Figura 12 - Evasão por idade no ensino médio.



Fonte: Autor, 2024.

Ao se analisar a variável cor/raça, observou-se que estudantes que se autodeclararam pretos, pardos e indígenas apresentam taxas de evasão mais elevadas, enquanto alunos brancos demonstram uma maior taxa de conclusão, com 15,3% de evasão entre os brancos, em comparação a 25,3% entre os pretos, 27% entre os pardos e 50% entre os indígenas.

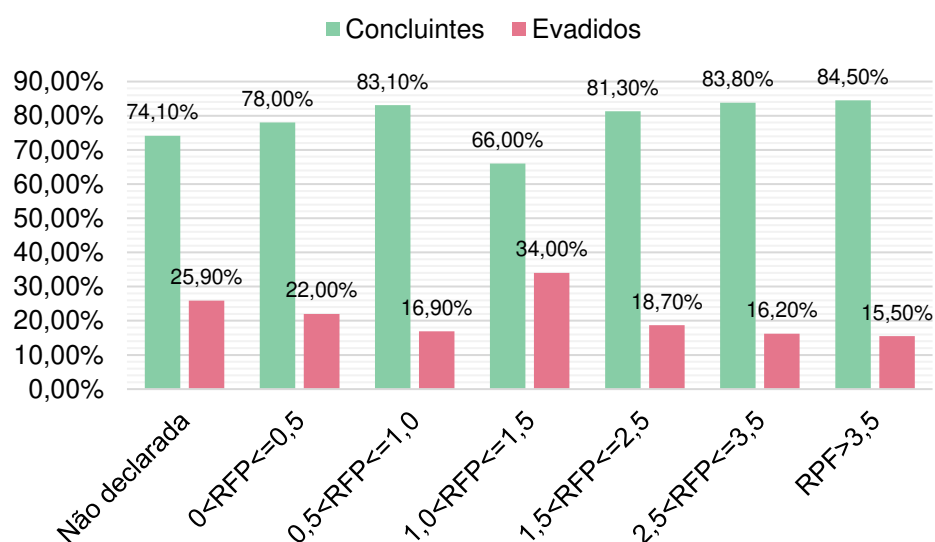
Figura 13 - Evasão por Cor/Raça no ensino médio.



Fonte: Autor, 2024.

Quanto à renda familiar, a maior taxa de evasão foi observada entre os alunos com renda entre 1,0 e 1,5 salários mínimos (34%), seguidos pelos que não declararam renda (25,9%) e pelos alunos com renda per capita de até 0,5 salário mínimo (22%). Estudantes com renda entre 1,5 e 2,5 salários mínimos, assim como aqueles com renda entre 0,5 e 1,0 salário mínimo, apresentaram taxas de evasão mais baixas, de 18,7% e 16,9%, respectivamente. Em contraste, os alunos com renda per capita entre 2,5 e 3,5 salários mínimos e os que possuem renda superior a 3,5 salários mínimos tiveram as menores taxas de evasão, com 16,2% e 15,5%, respectivamente. Esses dados indicam que a evasão é mais acentuada entre estudantes com menor renda ou renda não declarada, evidenciando a necessidade de políticas de apoio financeiro para garantir a permanência desses alunos.

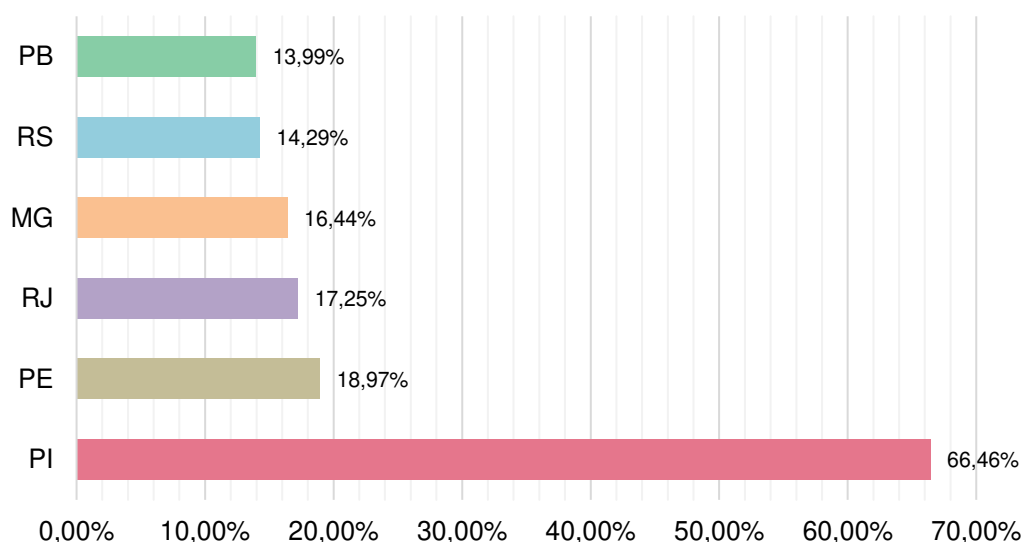
Figura 14 - Evasão por Renda Familiar no ensino médio.



Fonte: Autor, 2024.

A Figura 15 apresenta à análise por estado, o Piauí (PI) registrou a maior taxa de evasão, com 66,46%, seguido por Pernambuco (PE) com 18,97% e Rio de Janeiro (RJ) com 17,25%. Esses estados apresentaram índices de evasão mais elevados em comparação a outros estados, indicando a necessidade de estratégias específicas voltadas para a retenção e o apoio aos alunos nessas localidades.

Figura 15 - Evadido por UF no ensino médio.

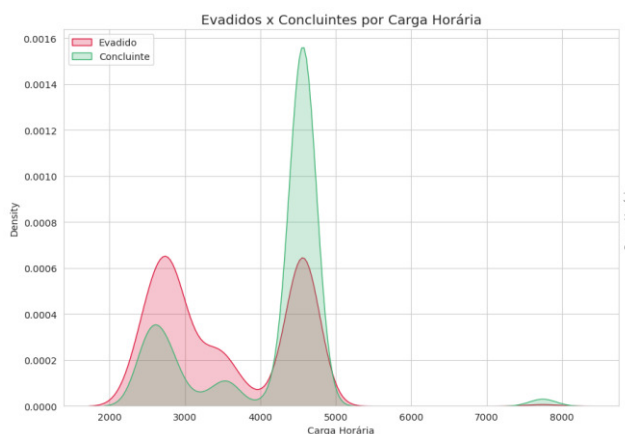


Fonte: Autor, 2024.

Por fim, ao se examinar a relação entre carga horária e evasão, foi observado que alunos de cursos com cargas horárias menores (em torno de 3.000 horas) têm

maior probabilidade de evasão. No entanto, a análise das distribuições de evadidos e concluintes revela uma sobreposição significativa, sugerindo que a carga horária, por si só, não é um fator decisivo para a evasão escolar. Isso indica que outros fatores, como contexto socioeconômico e apoio institucional, devem ser considerados para uma compreensão abrangente do fenômeno.

Figura 16 - Evadidos x Concluintes por carga horária no ensino médio.

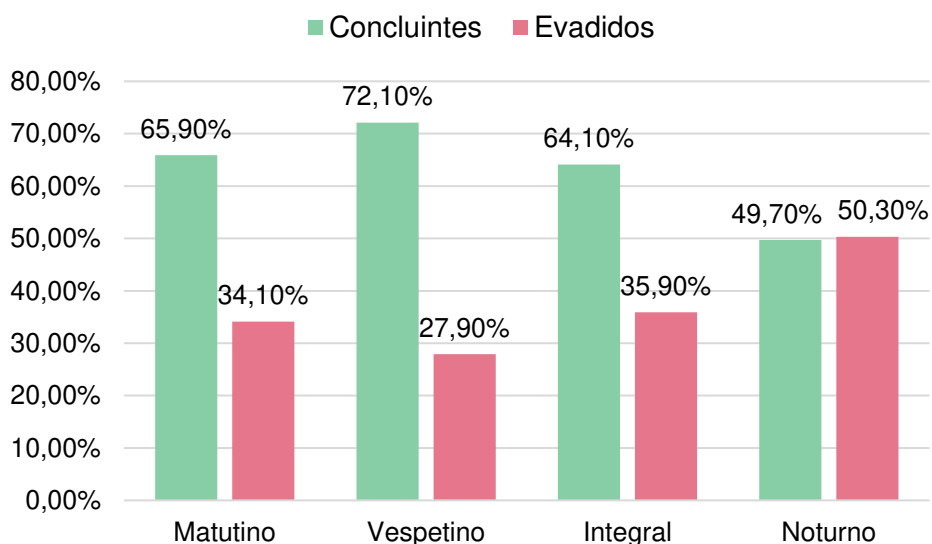


Fonte: Autor, 2024.

### 3.2.3. Análise de distribuição das variáveis do ensino médio técnico

Na análise do ensino médio técnico, foram avaliadas as mesmas variáveis escolhidas no ensino médio. Na Figura 17 observa-se a distribuição das taxas de evasão escolar de acordo com o turno de matrícula dos alunos. O turno noturno concentra o maior percentual de casos de evasão.

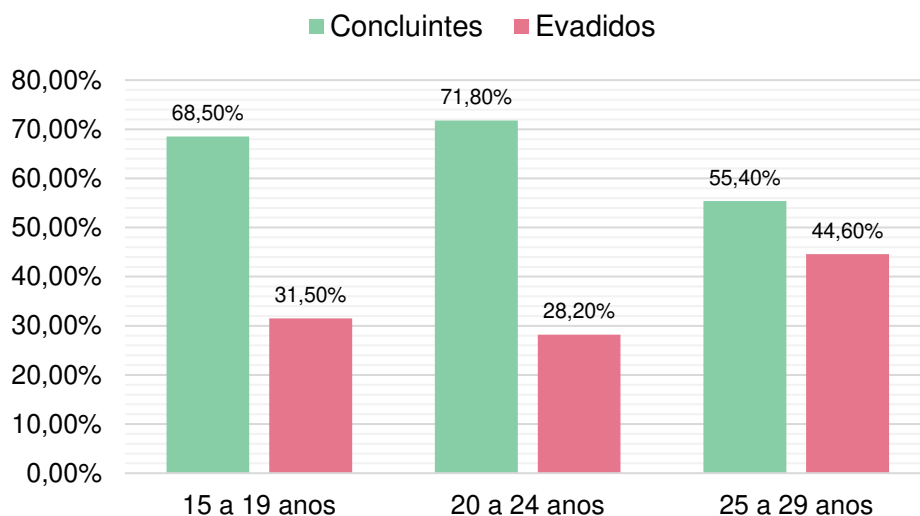
Figura 17 - Evasão por turno no ensino médio técnico.



Fonte: Autor, 2024.

A Figura 18 apresenta a relação entre a idade dos alunos e as taxas de evasão. Observa-se que as maiores taxas de evasão estão concentradas entre os 25 e 29 anos indicando uma maior vulnerabilidade nesses grupos.

Figura 18 - Evasão por faixa etária no ensino médio técnico.

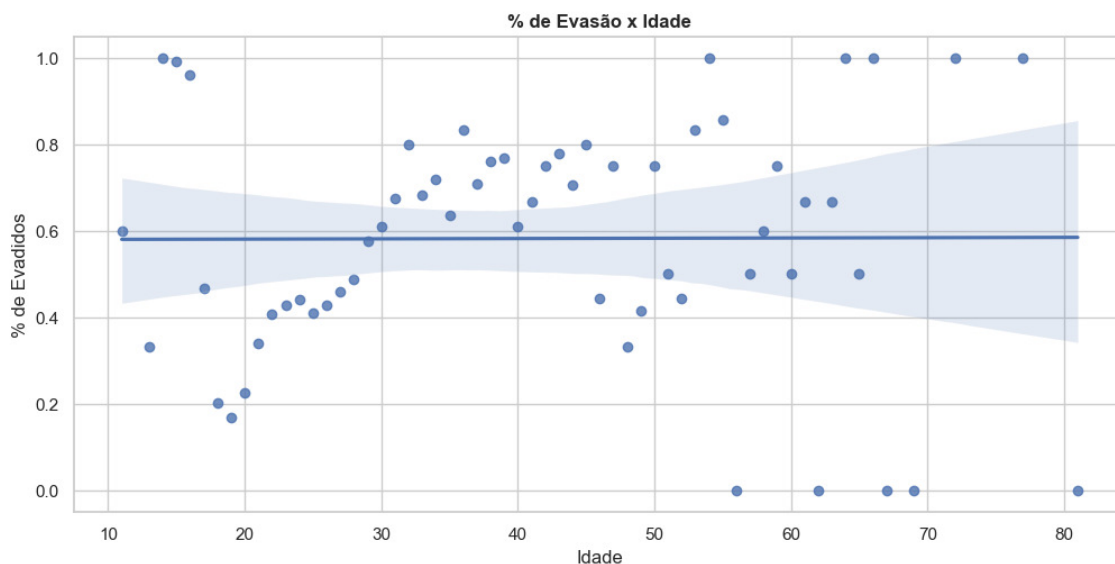


Fonte: Autor, 2024.

A Figura 19 apresenta uma correlação fraca entre a idade dos estudantes e o percentual de evasão, evidenciada por uma linha de regressão quase horizontal, o que sugere que a idade não exerce um impacto significativo na taxa de desistência escolar. No entanto, é notável uma maior concentração de evasão entre alunos mais jovens, especialmente aqueles na faixa de 10 a 20 anos, o que indica uma vulnerabilidade potencial nesse grupo etário. Além disso, embora os dados mostrem alta dispersão nas faixas etárias mais jovens e médias, a evasão também varia consideravelmente entre os alunos acima de 60 anos, sugerindo que outros fatores devem ser considerados além da idade. O intervalo de confiança, que se amplia nas extremidades, revela maior incerteza nas previsões para idades extremas, sejam muito jovens ou muito avançadas. Apesar da baixa correlação observada, a alta taxa de evasão entre os alunos mais jovens requer atenção, pois pode indicar desafios específicos relacionados a essa fase de transição, que necessitam de abordagens mais focadas nas políticas educacionais.



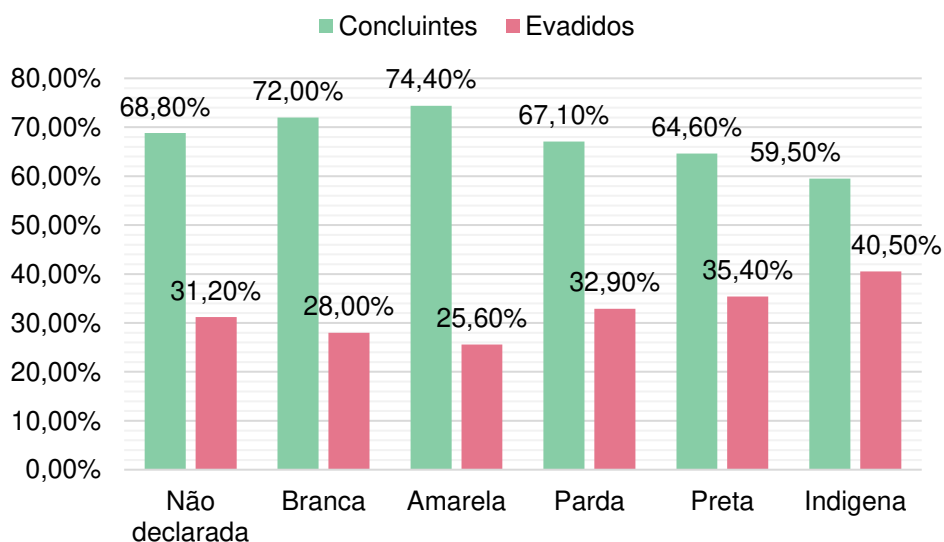
Figura 19 - Evasão por idade no ensino médio técnico.



Fonte: Autor, 2024.

Na análise da variável cor/raça, também foi identificado que alunos que se autodeclararam pretos, pardos e indígenas apresentam taxas de evasão superiores, enquanto os alunos brancos tendem a ter uma taxa de conclusão mais elevada.

Figura 20 - Evasão por Cor/Raça no ensino médio técnico.

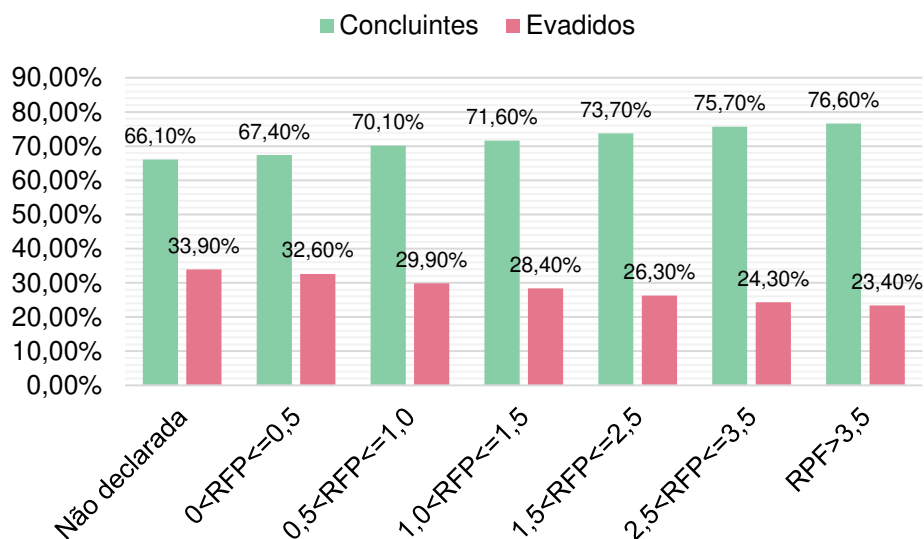


Fonte: Autor, 2024.

No que diz respeito à renda familiar, a maior taxa de evasão foi registrada entre alunos com renda não declarada (33,9%), seguidos por aqueles com renda per capita de até 0,5 salário mínimo (32,6%). Por outro lado, os alunos com renda entre 1,5 e

2,5 salários mínimos, entre 2,5 e 3,5 salários mínimos, e aqueles com mais de 3,5 salários mínimos apresentaram as menores taxas de evasão, que foram de 18,7%, 16,2% e 15,5%, respectivamente. Esses resultados indicam que a evasão tende a ser maior entre aqueles com menor renda ou renda não declarada, reforçando a necessidade de implementar políticas de apoio financeiro para garantir a permanência dos alunos.

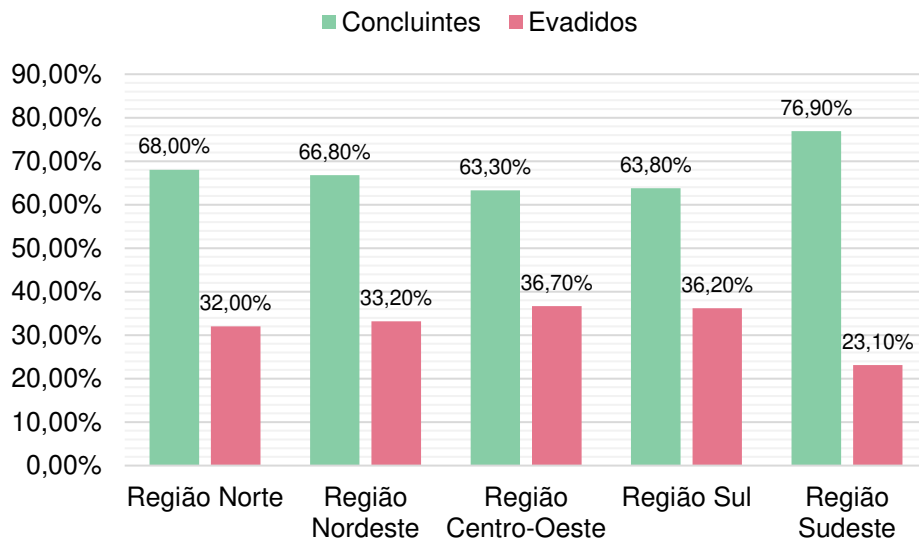
Figura 21 - Evasão por renda familiar no ensino médio técnico.



Fonte: Autor, 2024.

Na análise regional, constatou-se que a Região Centro-Oeste apresentou a maior taxa de evasão, com 36,7%, seguida pela Região Sul, com 36,2%. Esses índices elevados em comparação com outras regiões sinalizam a necessidade de desenvolver estratégias focadas na retenção e apoio aos alunos nessas áreas.

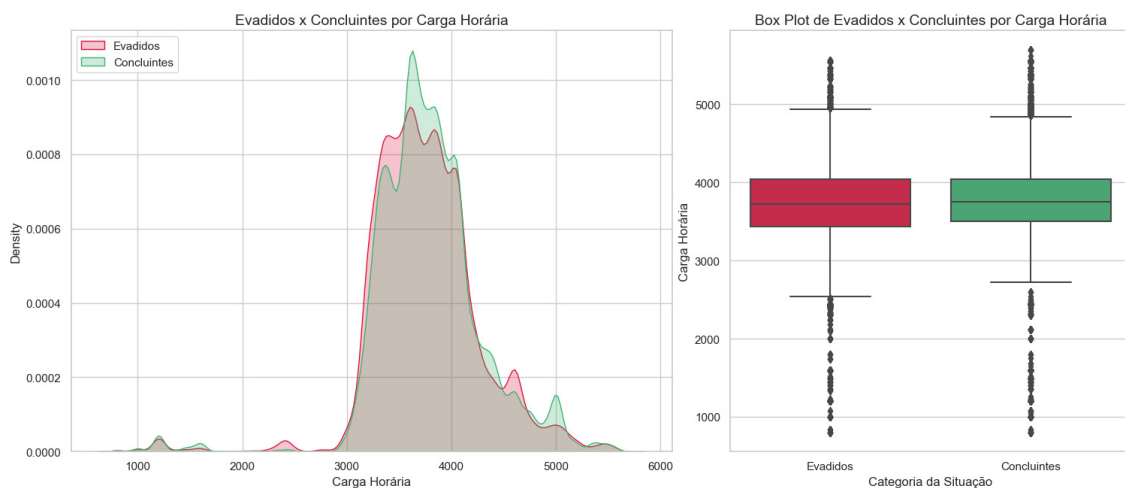
Figura 22 - Evasão por região no ensino médio técnico.



Fonte: Autor, 2024.

Por fim, ao investigar a relação entre carga horária e evasão, observou-se que alunos de cursos com cargas horárias superiores a 3.000 horas apresentam uma maior probabilidade de evasão. Contudo, as distribuições de evadidos e concluintes revelam uma sobreposição significativa, sugerindo que a carga horária, por si só, não é um determinante claro da evasão.

Figura 23 - Evadidos x Concluintes por carga horária o ensino médio técnico.



Fonte: Autor, 2024.

## 4. MODELAGEM PREDITIVA

Este capítulo apresenta a modelagem preditiva com o objetivo de prever a probabilidade de evasão escolar entre estudantes do ensino médio e técnico. Com base nas variáveis relevantes identificadas na análise exploratória, foram desenvolvidos dois processos de modelagem, um para cada tipo de curso. Nessas modelagens, aplicaram-se técnicas de Machine Learning, como Árvore de Decisão, KNN, Regressão Logística e Random Forest. Para melhorar a eficácia dos modelos, foram utilizados métodos de balanceamento de classes, assegurando uma avaliação mais precisa do desempenho preditivo.

### 4.1. Preparação dos dados para modelagem

Para a construção dos modelos de predição, inicialmente, as variáveis independentes (*features*) foram separadas da variável dependente (*target*). As *features* foram extraídas da base de dados, excluindo-se a coluna referente à situação da matrícula dos alunos (concluído ou evadido), que constitui o alvo da predição.

Antes de gerar os modelos, foi necessário realizar uma preparação adequada dos dados. Primeiramente, foi feita uma seleção cuidadosa das variáveis mais relevantes para cada tipo de curso, com base na análise exploratória. Para ambos os cursos, as variáveis renda familiar, idade, sexo, região, carga horária e turno foram selecionadas como preditores.

As variáveis categóricas foram transformadas utilizando o método *OneHotEncoder*<sup>8</sup>, enquanto as variáveis numéricas foram normalizadas com o *MinMaxScaler*<sup>9</sup>. Esse processo foi encapsulado em um pipeline, facilitando a automação das etapas de preparação e garantindo que todas as transformações fossem aplicadas de maneira consistente em cada iteração do modelo.

Durante a análise exploratória dos dados do ensino médio, foi identificado um significativo desbalanceamento nas classes, com 77% (3279) dos alunos concluintes e aproximadamente 23% (971) alunos evadidos. Esse desbalanceamento poderia introduzir vieses nos modelos, prejudicando sua capacidade de generalizar

---

<sup>8</sup> *OneHotEncoder* é uma técnica utilizada para transformar variáveis categóricas em uma representação numérica, criando uma nova coluna para cada categoria única. Essa técnica é importante para que os modelos de Machine Learning possam processar adequadamente os dados categóricos.

<sup>9</sup> *MinMaxScaler* é uma técnica de normalização que ajusta os valores numéricos para um intervalo específico, geralmente entre 0 e 1. Isso evita que variáveis com grandes diferenças de escala influenciem de forma desproporcional os modelos de Machine Learning.

corretamente. Para mitigar esse problema, foram aplicadas técnicas de *oversampling*<sup>10</sup> e *undersampling*<sup>11</sup>, resultando em uma distribuição equilibrada de 2623 amostras em cada classe.

Nos dados do ensino médio técnico, a distribuição era mais equilibrada, com 69% (114.416) dos alunos que se formaram e 31% (51.724) alunos que evadiram, o que permitiu um treinamento mais direto, sem a necessidade de balanceamento das classes.

## 4.2. Treinamento dos modelos

Para o treinamento dos modelos, os dados foram divididos em conjuntos de treinamento e teste, utilizando-se a função “train\_test\_split”, com 80% dos dados alocados para o treinamento e 20% para o teste. Essa proporção foi escolhida para garantir uma quantidade suficiente de dados para treinar os modelos, preservando um conjunto de teste robusto para avaliar a capacidade de generalização.

Foram implementados diferentes algoritmos de aprendizado de máquina, de acordo com as particularidades de cada conjunto de dados:

- **ensino médio:** foram utilizados os algoritmos Regressão Logística, KNN e Random Forest. O algoritmo Random Forest foi escolhido por sua capacidade de capturar a complexidade dos dados, especialmente com a alta variabilidade dos fatores socioeconômicos.
- **ensino médio técnico:** neste caso, optou-se por usar Regressão Logística, KNN e Árvore de Decisão. Devido ao grande volume de dados e à menor complexidade, a Árvore de Decisão mostrou-se mais eficiente que o Random Forest, tendo em vista que a máquina teve dificuldades de processamento e não conseguiu concluir a execução do modelo.

O ambiente computacional utilizado para o treinamento dos modelos incluiu um sistema com um processador Intel Core i7 da 10ª geração, 16 GB de memória RAM, e sistema operacional *Windows 11 Home Single Language*, equipado com uma placa de vídeo NVIDIA GeForce RTX 2060. Esta infraestrutura permitiu o processamento

---

<sup>10</sup> *Oversampling* é uma técnica usada para aumentar a quantidade de exemplos da classe minoritária em um conjunto de dados desbalanceado. Isso é feito duplicando ou criando novas amostras da classe minoritária, equilibrando a proporção entre as classes.

<sup>11</sup> *Undersampling* é uma técnica utilizada para reduzir a quantidade de exemplos da classe majoritária em um conjunto de dados desbalanceado, removendo algumas amostras dessa classe para equilibrar a proporção entre as classes.

de grandes volumes de dados e a execução eficiente de algoritmos mais complexos, como Random Forest e KNN.

Para otimizar o desempenho dos modelos, os hiperparâmetros de cada algoritmo foram ajustados utilizando a técnica de validação cruzada por meio do método *RandomizedSearchCV*<sup>12</sup> da biblioteca Scikit-learn. A seguir, os principais hiperparâmetros ajustados para cada algoritmo:

- **KNN:** foram testados diferentes valores para `n_neighbors` (de 1 a 19) e duas opções para o parâmetro `weights`: "uniform" e "distance".
- **Random Forest:** foram ajustados os hiperparâmetros `criterion` (com opções "gini" e "entropy"), `n_estimators` (100, 200 e 300), `max_depth` (sem limite, 10, 20 e 30), `min_samples_split` (2, 5 e 10) e `min_samples_leaf` (1, 2 e 4).
- **Regressão Logística:** os ajustes incluíram a `penalty` ("l1" e "l2"), o parâmetro `C` (valores de 0.1 a 1.2) e o `class_weight` ("balanced" e None).
- **Árvore de Decisão:** foram testados o `criterion` (com "gini" e "entropy"), `max_depth` (10, 20, 30 e 50), `min_samples_split` (2, 5 e 10) e `min_samples_leaf` (1, 2 e 4).

Esse processo de ajuste foi fundamental para melhorar a capacidade de predição dos modelos e otimizar seu desempenho nas tarefas de detecção da evasão escolar.

### 4.3. Avaliação de desempenho dos modelos

Para avaliar o desempenho dos modelos, considerou métricas como acurácia, precisão, recall e F1-score, permitindo uma análise detalhada do desempenho de cada modelo em relação à predição da evasão escolar.

As estatísticas dos modelos para o nível médio após a validação cruzada e ajuste de parâmetros encontram-se na Tabela 1:

Tabela 1 - Estatísticas dos modelos do ensino médio.

<b>Modelo</b>	<b>Accuracy</b>	<b>F1-Score</b>	<b>Precision</b>	<b>Recall</b>
KNN	0.87 ± 0.02	0.72 ± 0.03	0.71 ± 0.06	0.74 ± 0.03
<i>Random Forest</i>	0.97 ± 0.00	0.92 ± 0.00	0.95 ± 0.00	0.89 ± 0.01
<i>Logistic Regression</i>	0.87 ± 0.02	0.73 ± 0.03	0.68 ± 0.03	0.80 ± 0.02

Fonte: Autor, 2024.

<sup>12</sup> *RandomizedSearchCV* é uma técnica de otimização de hiperparâmetros para modelos de Machine Learning que realiza a busca aleatória em uma grade de parâmetros predefinida.

O modelo Random Forest demonstrou um desempenho significativamente superior, com acurácia de 97%. Essa alta acurácia pode ser atribuída à capacidade do Random Forest de capturar a complexidade dos dados do ensino médio, sendo um ensemble de múltiplas árvores de decisão que se mostrou mais robusto para esse conjunto de dados. Adicionalmente, o recall de 89% indica uma boa capacidade do modelo em identificar corretamente os alunos que evadiram.

Já o KNN e a Regressão Logística apresentaram acurácias de 87%, mas com um desempenho inferior em termos de precisão e recall em comparação com o Random Forest. O KNN apresentou um recall de 74%, o que é aceitável, porém sua precisão de 71% sugere que ele teve mais dificuldade em identificar corretamente os alunos evadidos.

As estatísticas dos modelos para os cursos médio técnicos após a validação cruzada e ajuste de parâmetros encontram-se na Tabela 2:

Tabela 2 - Estatísticas dos modelos do ensino médio técnico.

<b>Modelo</b>	<b>Accuracy</b>	<b>F1-Score</b>	<b>Precision</b>	<b>Recall</b>
<i>Decision Tree</i>	0.86 ± 0.03	0.76 ± 0.07	0.88 ± 0.12	0.69 ± 0.15
KNN	0.71 ± 0.06	0.48 ± 0.07	0.59 ± 0.15	0.42 ± 0.06
<i>Logistic Regression</i>	0.84 ± 0.02	0.70 ± 0.09	0.87 ± 0.08	0.62 ± 0.17

Fonte: Autor, 2024.

No contexto dos cursos técnicos, a Árvore de Decisão foi o modelo que mais se destacou, com acurácia de 86%. Entretanto, o recall de 69% indica que o modelo teve dificuldades para identificar com precisão todos os alunos evadidos. A simplicidade do algoritmo em comparação com o Random Forest pode explicar essa diferença de desempenho, especialmente em um conjunto de dados mais amplo e complexo.

O KNN teve um desempenho inferior nos dados do ensino médio técnico, com acurácia de 71% e um F1-score de apenas 0.48, o que sugere que o modelo teve dificuldades com a generalização no conjunto de dados técnicos. O baixo valor de recall (42%) é indicativo da incapacidade do KNN de identificar adequadamente os alunos evadidos neste cenário.

A Regressão Logística também apresentou resultados medianos, com acurácia de 84%, mas um recall de 62%, o que demonstra a dificuldade do modelo em capturar corretamente todos os casos de evasão.

De maneira geral, os resultados observados podem ser atribuídos a diferentes fatores. Primeiramente, o balanceamento dos dados no ensino médio foi essencial para melhorar o desempenho dos modelos. Além disso, o tamanho da amostra nos dados técnicos, sendo significativamente maior que no ensino médio, contribuiu para uma melhor generalização da Árvore de Decisão.

Em termos de desempenho, a estrutura do Random Forest foi capaz de lidar melhor com a complexidade dos dados do ensino médio, explicando seu desempenho superior. Em contraste, a Árvore de Decisão mostrou-se mais eficiente nos dados técnicos, mas com uma margem de acurácia inferior ao Random Forest, possivelmente devido à natureza dos dados em si.



## 5. CONCLUSÃO

A evasão escolar no ensino médio e no ensino médio técnico é uma questão complexa que afeta diretamente o desenvolvimento educacional de jovens no Brasil. Identificar os fatores que contribuem para essa evasão é essencial para a formulação de políticas públicas e intervenções voltadas à permanência escolar. Diversos fatores, como a condição socioeconômica dos alunos, têm sido apontados como influências cruciais nesse fenômeno. No contexto da rede federal de ensino, a análise dos dados referentes à evasão escolar permite um entendimento mais aprofundado sobre essas questões, além de possibilitar a aplicação de modelos preditivos que auxiliam na identificação precoce de alunos em risco de evasão.

Com base nisso, este estudo propôs a aplicação de técnicas de modelagem preditiva para auxiliar na detecção de padrões e fatores que indicam maior probabilidade de abandono escolar, contribuindo assim para o desenvolvimento de soluções mais direcionadas e eficazes. Os dados utilizados foram cuidadosamente preparados e organizados, permitindo a aplicação de técnicas de modelagem preditiva com o intuito de identificar alunos em risco de evasão. A análise exploratória revelou padrões e tendências significativas, destacando a relevância de variáveis como renda familiar, idade, sexo e turno. A compreensão desses fatores foi fundamental para a construção de modelos preditivos, os quais incluíram algoritmos como *Árvore de Decisão*, *Regressão Logística*, *Random Forest* e *K-Nearest Neighbors* (KNN).

Os resultados indicaram que o modelo de *Random Forest* se destacou nos cursos de ensino médio, apresentando uma acurácia de 97%, enquanto a *Regressão Logística* demonstrou maior eficácia nos cursos técnicos, com uma acurácia de 84%. Esses resultados evidenciam a capacidade dos modelos em prever a evasão escolar, ressaltando a importância de intervenções direcionadas. O conjunto de dados gerado, utilizado para a modelagem e análise, está disponível para novas investigações e pode ser acessado no repositório do GitHub<sup>13</sup>.

Além disso, este trabalho sublinha o papel crucial da Ciência de Dados na tomada de decisões informadas, possibilitando a criação de intervenções personalizadas que visem construir ambientes educacionais mais resilientes e propícios ao engajamento dos estudantes. As percepções obtidas durante a análise

---

<sup>13</sup> <https://github.com/Geraldomendes/SchoolDropoutPrediction>

exploratória oferecem subsídios para a elaboração de políticas de retenção mais eficazes, especialmente voltadas para alunos em situação de vulnerabilidade socioeconômica.

No entanto, é importante reconhecer algumas limitações que impactaram este estudo. Primeiramente, a falta de acesso a dados mais abrangentes, como o histórico escolar detalhado dos alunos, informações sobre frequência às aulas e desempenho acadêmico contínuo, restringiu a profundidade da análise. A inclusão dessas variáveis poderia ter contribuído para uma compreensão mais holística dos fatores que influenciam a evasão e, potencialmente, aumentado a precisão dos modelos preditivos. Outra limitação relevante foi a falta de variáveis relacionadas a aspectos psicossociais, como o nível de apoio familiar e a motivação dos alunos, que também podem exercer grande influência nas decisões de permanência ou abandono escolar.

Além disso, o desbalanceamento das classes nos dados de evasão, especialmente nos cursos de ensino médio, embora tratado com técnicas de *oversampling* e *undersampling*, ainda pode ter introduzido vieses nos resultados. Esse fenômeno é comum em estudos de evasão escolar, mas suas implicações para a modelagem preditiva precisam ser cuidadosamente monitoradas, pois modelos treinados em dados desbalanceados podem superestimar a capacidade de retenção e subestimar os fatores reais de evasão.

Em suma, os objetivos deste trabalho foram alcançados por meio da aplicação de métodos de Machine Learning e da análise detalhada de dados abertos da rede federal de ensino. Os resultados obtidos oferecem uma base sólida para a implementação de medidas proativas de retenção e melhoria do desempenho acadêmico, contribuindo para a criação de um ambiente educacional mais inclusivo e sustentável. Contudo, futuros estudos poderiam expandir o escopo da análise com a inclusão de dados mais detalhados e variáveis adicionais, ampliando o potencial de identificar padrões mais complexos e interações não exploradas entre os fatores de evasão escolar.

### **5.1. Trabalhos futuros**

Para estudos futuros, recomenda-se a inclusão de dados mais abrangentes, como histórico escolar detalhado, informações sobre frequência e desempenho acadêmico, com o objetivo de enriquecer a análise e aprimorar os modelos preditivos. A adoção de algoritmos avançados, como redes neurais artificiais e métodos

ensemble, pode melhorar significativamente as previsões, permitindo uma captura mais precisa de padrões complexos.

Além disso, é fundamental mensurar a efetividade das políticas de intervenção implementadas com base nos resultados desta pesquisa. Esse acompanhamento permitirá avaliar se houve uma redução significativa nas taxas de evasão, contribuindo para o desenvolvimento de soluções inovadoras que enfrentem os desafios persistentes da educação.

Uma aplicação prática que poderia ser desenvolvida a partir dos modelos de Machine Learning elaborados neste estudo é um sistema de monitoramento da evasão escolar nas instituições da rede federal. Essa plataforma utilizaria os modelos preditivos para identificar em tempo real os alunos com maior risco de evasão, permitindo intervenções proativas pela administração escolar. O sistema integraria dados sobre desempenho acadêmico, frequência e fatores socioeconômicos, gerando alertas para que a equipe pedagógica implemente ações de suporte personalizadas.

Assim, a implementação de um sistema dessa natureza representaria a última etapa do CRISP-EDM, que envolve a aplicação de soluções baseadas em dados para aprimorar a tomada de decisões e apoiar políticas educacionais mais eficazes.

## REFERÊNCIAS

AALST, W. V. D. “*Data Science in Action*”. Em *Process Mining: Data Science in Action*, organizado por Wil van der Aalst, 3–23. Berlin, Heidelberg: Springer, 2016. [https://doi.org/10.1007/978-3-662-49851-4\\_1](https://doi.org/10.1007/978-3-662-49851-4_1).

ALPAYDIN, E. **Introduction to machine learning**. [S.l.]: MIT press, 2020.

ALTMAN, N. An introduction to kernel and nearest-neighbor nonparametric regression. *American Statistician - AMER STATIST*, v. 46, p. 175-185, 1992. <https://doi.org/10.2307/2685209>

AMARAL, F. **Introdução à ciência de dados: mineração de dados e Big Data**. Rio de Janeiro: ALTA Books, 2016.

ANUTO, Thaína Francis. Evasão escolar no ensino médio: possíveis inferências para mudar esse cenário. 2013. 34 f. Trabalho de Conclusão de Curso (Especialização) – Universidade Tecnológica Federal do Paraná, Medianeira, 2013.

BISHOP, C. M. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer, 2006.

BRASIL. Lei 9.394, de 20 de dezembro de 1996. Estabelece as diretrizes e bases da educação nacional. Diário Oficial da República. Disponível em: <[http://portal.mec.gov.br/seesp/arquivos/pdf/lei9394\\_ldbn1.pdf](http://portal.mec.gov.br/seesp/arquivos/pdf/lei9394_ldbn1.pdf)>. Acesso em: 25 de nov. 2023.

Brasil. Lei nº 13.005, de 25 de junho de 2014. Plano Nacional de Educação 2014-2024: – Brasília: Câmara dos Deputados, Edições Câmara, 2014. Disponível em: <<http://www.proec.ufpr.br/download/extensao/2016/creditacao/PNE%202014-2024.pdf>>. Acesso em: 25 de nov. 2023.

BREIMAN, L. Random Forests, *Machine Learning*, Vol. 45, pp. 5 – 32, 2001.

BUGNION, P.; MANIVANNAN, A.; NICOLAS, P.R. *Scala: Guide for Data Science Professionals*. Birmingham: Packt Publishing, 2017, 1077 p.

COLPANI, R. Mineração de Dados Educacionais: um estudo da evasão no ensino médio com base nos indicadores do Censo Escolar. **Informática na educação: teoria & prática**, Porto Alegre, v. 21, n. 3, 2018. DOI: 10.22456/1982-1654.87880. Disponível em: <https://seer.ufrgs.br/index.php/InfEducTeoriaPratica/article/view/87880>. Acesso em: 21 nov. 2023.

CONAMAY, D. The data science venn diagram (2010). Disponível em: <http://drewconway.com/zia/2013/3/26/the-data-science-venndiagram>. Acesso em: 28 nov. 2023.

COOMANS, D.; MASSART, D. Alternative k-nearest neighbour rules in supervised pattern recognition: Part 1. k-nearest neighbour classification by using alternative voting rules. *Analytica Chimica Acta*, v. 136, p. 15-27, 1982. [https://doi.org/10.1016/S0003-2670\(01\)95359-0](https://doi.org/10.1016/S0003-2670(01)95359-0)

CREPALDI, P. G.; AVILA, R. N. P.; OLIVEIRA, J. P. N.; RODRIGUES, P. R.; MARTINS R. L. Um Estudo Sobre a Árvore de Decisão e sua Importância na Habilidade de Aprendizado, 2010.

Crook, J. N., Edelman, D. B., & Thomas, L. C. 2007. Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, 183(3), 1447–1465.

ESCOVEDO, T.; KOSHIYAMA, A. **Introducao a Data Science: Algoritmos de**

FERREIRA, E. C. da S.; OLIVEIRA, N. M. de. EVASÃO ESCOLAR NO ENSINO MÉDIO: causas e consequências. **Scientia Generalis**, [S. l.], v. 1, n. 2, p. 39–48, 2020. Disponível em: <https://scientiageneralis.com.br/index.php/SG/article/view/v1n2a4>. Acesso em: 29 nov. 2023.

FERREIRA, S. G.; RIBEIRO, G.; TAFNER, P. Abandono e evasão escolar no Brasil. Instituto Mobilidade e Desenvolvimento Social, 2022.

FRIEDMAN, N., GEIGER, D. e GOLDSZMIDT, M. *Bayesian Network Classifiers*. *Machine Learning* 29, nº 2 (1º de novembro de 1997): 131–63. <https://doi.org/10.1023/A:1007465528199>.

GAMA, J. a. Functional trees. *Machine Learning*, v. 55, p. 219–250, 2004. 22, 31, 32, 50

GARCIA, S.C. O uso de árvores de decisão na descoberta de conhecimento na área da saúde. In: SEMANA ACADÊMICA, 2000. Rio Grande do Sul: Universidade Federal do Rio Grandedo Sul, 2000.

GÉRON, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow*. United States of America: O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, 2017.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6. ed. São Paulo: Atlas, 2008.  
Gouvêa, M. A. and Gonçalves, E. B. and Mantovani D. M. N. 2015. Análise de Risco de Crédito com Aplicação de Regressão Logística e Redes Neurais. *Contabilidade Vista Revista*, 24(4), 96–123.

Hair, JR. J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. 2009. *Análise multivariada de dados*. first edn. Bookman.

HAN, J.; KAMBER, M.; PEI, J. *Data Mining: Concepts and Techniques*. 3. Ed. Morgan Kaufmann, 2011.

HILBE, J. M. Logistic regression. **International encyclopedia of statistical science**,

HO, T. K. Random decision forests. 3rd International Conference on Document Analysis and Recognition, v. 1, pp. 278-282, 1995. <https://doi.org/10.1109/ICDAR.1995.598994>

KOTSIANTIS, S. B. “Decision Trees: A Recent Overview”. *Artificial Intelligence Review* 39, nº 4 (1º de abril de 2013): 261–83. <https://doi.org/10.1007/s10462-011-9272-4>.

LAAKSONEN, J. e OJA, E. “Classification with learning k-nearest neighbors”. Em *Proceedings of International Conference on Neural Networks (ICNN'96)*, 3:1480–83 vol.3, 1996. <https://doi.org/10.1109/ICNN.1996.549118>.

LEMOS, I. V. de R. **Prevedo a evasão escolar em uma instituição de ensino técnico utilizando mineração de dados educacionais**. 2021. 43 f. Trabalho de Conclusão de Curso (Bacharelado em Ciência da Computação) – Departamento de Computação, Universidade Federal Rural de Pernambuco, Recife, 2021.

LOPES, G. R.; ALMEIDA, A. W. S.; DELBEM, A. C. B.; TOLEDOO, C. F. M. *Introdução à Análise Exploratória de Dados com Python*, 2019

MACHADO, F. S. *Análise e previsão da evasão escolar do ensino médio através de dados públicos*. dez. 2019.

**Machine Learning e metodos de analise**. [S.l.]: Casa do Codigo, 2020.

MEDRI, W. ANÁLISE EXPLORATÓRIA DE DADOS. UEL, 2011. Disponível em: <<https://docs.ufpr.br/~benitoag/apostilamedri.pdf>>. Acesso em: 1 dez. 2023.

MITCHELL, T. M. *Machine Learning*. 1. ed. McGraw-Hill series in computer science. New York: McGraw-Hill Science/Engineering/Math, 1997.

MONARD, Maria Carolina; BARANAUSKAS, José Augusto. *Sistemas Inteligentes-Fundamentos e Aplicações*, v. 1, 2003

MORAES, E. R. P. T.; LINHARES, C. *Evasão escolar*. Analecta, Guarapuava, 1982.

MÜLLER, B.; JOACHIM R. e MICHAEL T. Strickland. *Neural Networks: An Introduction*. Springer Science & Business Media, 1995.

NOBLE, W. S. “What Is a Support Vector Machine?” *Nature Biotechnology* 24, nº 12 (dezembro de 2006): 1565–67. <https://doi.org/10.1038/nbt1206-1565>.

OLIVEIRA, F. L. de; NÓBREGA, L. *Evasão escolar: um problema que se perpetua na educação brasileira*. *Revista Educação Pública*, v. 21, nº 19, 25 de maio de 2021. Disponível em: <<https://educacaopublica.cecierj.edu.br/artigos/21/19/evasao-escolar-um-problema-que-se-perpetua-na-educacao-brasileira>> Acesso em: 27 nov. 2023.

ONODA, M.; EBECKEN, N. *Implementação em java de um algoritmo de Árvore de decisão acoplado a um sgbd relacional*. In: . [S.l.: s.n.], 2001. p. 55–64.

PACHECO, André. K vizinhos mais próximos - KNN. 17 mar. 2017. Disponível em: <https://computacaointeligente.com.br/algoritmos/k-vizinhos-mais-proximos/>. Acesso em: 21 out. 2024.

PEARSON, R. K. *Exploratory Data Analysis Using R*. New York: Chapman and Hall/CRC, 2018. <https://doi.org/10.1201/9781315382111>.

PEREIRA, V. *Causas e consequências do abandono e da evasão escolar*. Instituto Mobilidade e Desenvolvimento Social, 2022.

QUEIROZ, L. D. Um Estudo Sobre a Evasão Escolar: Para se Pensar a Inclusão Social. 25ª Reunião anual da Anped, Caxambu, v. 1, n.1, p. 01-01, 2002.

RAMOS, J. L. C.; RODRIGUES, R. L.; SILVA, J. C. S.; OLIVEIRA, P. L. S. CRISP-EDM: uma proposta de adaptação do Modelo CRISP-DM para mineração de dados educacionais. In: SIMPÓSIO BRASILEIRO DE INFORMÁTICA NA EDUCAÇÃO, 31., 2020, Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1092-1101. DOI: <https://doi.org/10.5753/cbie.sbie.2020.1092>.

RAUTENBERG, S.; CARMO, P. R. V. do. Big data e ciência de dados: complementariedade conceitual no processo de tomada de decisão. **Brazilian Journal of Information Science: research trends**, [S. l.], v. 13, n. 1, p. 56–67, 2019. DOI: 10.36311/1981-1640.2019.v13n1.06.p56. Disponível em: <https://revistas.marilia.unesp.br/index.php/bjis/article/view/8315>. Acesso em: 28 nov. 2023.

RODRIGUES, L. P. “EVASÃO ESCOLAR: UMA EXPRESSÃO DA “QUESTÃO SOCIAL””. TCC. Universidade Federal da Paraíba, 1º de abril de 2020. <https://repositorio.ufpb.br>.

SAMPAIO, H. *Evolução do ensino superior brasileiro: 1808 – 1990*. Documento de Trabalho. NUPES, 8/91. Núcleo de Pesquisa sobre Ensino Superior da Universidade de São Paulo, 1991.

SILVA, J. C. L. Definição de modelos de aprendizado de máquina para predição de evasão de alunos do curso técnico. **Refas - Revista Fatec Zona Sul**, [S. l.], v. 9, n. 3, p. 1–12, 2023. DOI: 10.26853/Refas\_ISSN-2359-182X\_v09n03\_01. Disponível em:



<https://www.revistarefas.com.br/RevFATECZS/article/view/438>. Acesso em: 20 nov. 2023.

SOARES, G. F. Ciência de dados aplicada à Auditoria Interna. **Revista da CGU**, v. 12, n. 22, p. 196–208, 30 dez. 2020.


SOUZA, J. A. S. PERMANÊNCIA E EVASÃO ESCOLAR: UM ESTUDO DE CASO EM UMA INSTITUIÇÃO DE ENSINO PROFISSIONAL. **Revista Brasileira da Educação Profissional e Tecnológica**, [S. l.], v. 1, n. 6, p. 19–29, 2016. DOI: 10.15628/rbept.2013.3498. Disponível em: <https://www2.ifrn.edu.br/ojs/index.php/RBEPT/article/view/3498>. Acesso em: 21 nov. 2023.

STANTON, J. INTRODUCTION TO DATA SCIENCE. *School of Information Studies - Faculty Scholarship*, 2012. Disponível em: <<https://surface.syr.edu/istpub/165>>.

TUKEY, J. W. **Exploratory Data Analysis**. [S.l.]: Addison-Wesley, 1977. v. 1, p. 15–32, 2011.

WICKHAM, HADLEY, MINE ÇETINKAYA-RUNDEL, e GARRETT GROLEMUND. *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 2. Ed. Beijing Boston Farnham Sebastopol Tokyo: O'Reilly, 2023.

WIRTH, R.; JOCHEN, H. “*CRISP-DM: Towards a standard process model for data mining*”. Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining, 2000.

	<b>INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA</b>
	Campus Cajazeiras - Código INEP: 25008978
	Rua José Antônio da Silva, 300, Jardim Oásis, CEP 58.900-000, Cajazeiras (PB)
	CNPJ: 10.783.898/0005-07 - Telefone: (83) 3532-4100

## Documento Digitalizado Ostensivo (Público)

### TCC - Versão final

<b>Assunto:</b>	TCC - Versão final
<b>Assinado por:</b>	Geraldo Mendes
<b>Tipo do Documento:</b>	Dissertação
<b>Situação:</b>	Finalizado
<b>Nível de Acesso:</b>	Ostensivo (Público)
<b>Tipo do Conferência:</b>	Cópia Simples

Documento assinado eletronicamente por:

- **Geraldo Mendes Batista Neto, DISCENTE (202122010031) DE TECNOLOGIA EM ANÁLISE E DESENVOLVIMENTO DE SISTEMAS - CAJAZEIRAS**, em 22/10/2024 15:19:54.

Este documento foi armazenado no SUAP em 22/10/2024. Para comprovar sua integridade, faça a leitura do QRCode ao lado ou acesse <https://suap.ifpb.edu.br/verificar-documento-externo/> e forneça os dados abaixo:

Código Verificador: 1287652

Código de Autenticação: 770b1b98e7

