UNIVERSIDADE NOVE DE JULHO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO

FERNANDA PEREIRA GOMES

MINERAÇÃO DE DADOS EDUCACIONAIS NA IDENTIFICAÇÃO DO PERFIL DE ALUNOS COM RISCO DE EVASÃO EM ESCOLA TÉCNICA PÚBLICA COM BASE NA TRAJETÓRIA ACADÊMICA

São Paulo
2026

FERNANDA PEREIRA GOMES

MINERAÇÃO DE DADOS EDUCACIONAIS NA IDENTIFICAÇÃO DO PERFIL DE ALUNOS COM RISCO DE EVASÃO EM ESCOLA TÉCNICA PÚBLICA COM BASE NA TRAJETÓRIA ACADÊMICA

Dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, como requisito parcial para obtenção do título de Mestre em Informática e Gestão do Conhecimento.

Orientador: Prof. Dr. Renato José Sassi
Linha de Pesquisa 3: Gestão da Informação e do Conhecimento

São Paulo
2026

Gomes, Fernanda Pereira.
Mineração de dados educacionais na identificação do perfil de alunos com risco de evasão em escola técnica pública com base na trajetória acadêmica / Fernanda Pereira Gomes. 2026.
116 f.
Dissertação (Mestrado)- Universidade Nove de Julho - UNINOVE, São Paulo, 2026.
Orientador (a): Prof. Dr. Renato José Sassi.

1. Perfil do Aluno; 2. Evasão escolar. 3. ETEC 4. Mineração de Dados Educacionais; 5. Gestão Escolar.
I. Sassi, Renato José.

CDU 004

UNINOVE
Universidade Nove de Julho

PARECER – EXAME DE DEFESA

Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento, ao qual se submeteu a aluna Fernanda Pereira Gomes.

Tendo examinado o trabalho apresentado para a obtenção do título de “Mestre em Informática e Gestão do Conhecimento”, com a Dissertação intitulada "MINERAÇÃO DE DADOS EDUCACIONAIS NA IDENTIFICAÇÃO DO PERFIL DE ALUNOS COM RISCO DE EVASÃO EM ESCOLA TÉCNICA PÚBLICA BASEADA NA TRAJETÓRIA ACADÊMICA”, a Comissão Examinadora considerou o trabalho:

(X) Aprovado **( ) Aprovado condicionalmente**
( ) Reprovado com direito a novo exame **( ) Reprovado**

EXAMINADORES

Prof. Dr. Renato José Sassi

Prof. Dr. Irapuan Glória Júnior

gov.br Documento assinado digitalmente
IRAPUAN GLORIA JUNIOR
Data: 12/06/2026 18:13:24-0300
Verifique em https://validar.iti.gov.br

Prof. Dr. Marcio Romero

São Paulo, 03 de junho de 2026

Dedico este trabalho à minha mãe que sempre me incentivou a continuar.

AGRADECIMENTOS

Aos docentes, pesquisadores e colaboradores do Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGI) da Universidade Nove de Julho (UNINOVE), agradeço pela formação acadêmica proporcionada, pelas contribuições ao longo do curso e pelo ambiente de aprendizado que possibilitou o desenvolvimento deste trabalho.

Ao Centro Paula Souza, agradeço pela disponibilização dos dados utilizados nesta pesquisa e pelo apoio institucional que tornou possível a realização deste trabalho. À equipe pedagógica e administrativa da ETEC Raposo Tavares, registro meus agradecimentos pelo acolhimento, pela colaboração e pelo suporte oferecido ao longo do desenvolvimento desta pesquisa.

À memória de minha mãe, Genovina Pereira Gomes, registro minha gratidão pelos ensinamentos, pelo apoio e pelos valores transmitidos, que permaneceram como referência ao longo de toda a minha trajetória acadêmica e pessoal. Ao meu pai e aos meus irmãos, agradeço pela presença, compreensão e apoio nos momentos em que a dedicação ao trabalho acadêmico se fez necessária.

Ao meu filho, Ricardo Gomes Silva, agradeço pela compreensão, pelo carinho e por ser fonte permanente de motivação. Ao meu marido, Jorge Luis de Souza Rodrigues Junior, registro meu agradecimento pelo apoio, pela paciência e pelo companheirismo ao longo de todo o processo de desenvolvimento desta dissertação. Ao meu orientador, Renato José Sassi, agradeço pela orientação, pelas contribuições técnicas e metodológicas, pela disponibilidade e pelas reflexões que contribuíram de forma significativa para a construção deste trabalho.

Por fim, agradeço ao meu cão, Myers, pela companhia silenciosa e constante, que tornou os momentos de estudo mais leves e acolhedores.

A todos que, de alguma forma, contribuíram para a realização deste trabalho, meus sinceros agradecimentos.

“Quanto mais estudo, mais insaciável se torna meu desejo pelo conhecimento.”

Ada Lovelace

RESUMO

As Escolas Técnicas Públicas do Estado de São Paulo, denominadas ETECs, buscam alinhar sua formação às exigências do mercado de trabalho por meio de cursos que articulam teoria e prática. Essas escolas enfrentam desafios relacionados à evasão escolar, que podem ser analisados a partir de dados acadêmicos institucionais. O caso da ETEC Raposo Tavares ilustra esse cenário, pois somente 43,59% dos alunos que ingressaram no curso Técnico em Química em 2023 concluíram o curso em 2024, percentual inferior à meta institucional de 65%. Nesse contexto, a Mineração de Dados Educacionais (MDE) permite explorar esses dados e identificar padrões associados aos perfis de alunos com risco de evasão. O objetivo deste trabalho é aplicar a Mineração de Dados Educacionais para identificar perfis de alunos com risco de evasão do Curso Técnico em Química da ETEC Raposo Tavares, com base na análise da trajetória acadêmica, para apoiar a gestão escolar no desenvolvimento de estratégias educacionais. A base utilizada nos experimentos computacionais reúne informações de 185 alunos matriculados do primeiro ao terceiro módulo nos anos de 2023 e 2024, extraídas do Novo Sistema Acadêmico (NSA) e complementadas por registros das atas dos conselhos de classe. Os experimentos seguiram as fases do processo de Descoberta de Conhecimento em Bases de Dados: Seleção dos Dados, Pré-processamento dos Dados, Transformação dos Dados, Mineração de Dados Educacionais e Interpretação e Avaliação do Conhecimento. Foram aplicadas as técnicas Regressão Logística, Random Forest e XGBoost. A Random Forest apresentou o melhor desempenho. Com base nos resultados dessa técnica, a equipe de gestão escolar da ETEC identificou três perfis: Alto Desempenho Escolar, associado ao aluno que se dedica ao curso e busca a formação; Médio Desempenho Escolar, associado ao aluno que, apesar das dificuldades, busca concluir o curso; e Baixo Desempenho Escolar, associado ao maior risco de evasão. Esse último perfil pode relacionar-se a desigualdades sociais, ausência de apoio familiar, desalinhamento vocacional ou fragilidade no sentimento de identificação e pertencimento institucional. Entre as medidas selecionadas pela equipe de gestão para reduzir a evasão escolar, destaca-se a institucionalização de um protocolo interno de monitoramento permanente, no qual os indicadores sejam acompanhados ao longo do período letivo. O protocolo deve articular-se às instâncias pedagógicas para que os dados produzidos se convertam em estratégias educacionais concretas de apoio ao aluno. Conclui-se que a MDE constitui um recurso relevante para a ETEC Raposo Tavares e pode apoiar a gestão escolar na definição de estratégias educacionais voltadas à redução do risco de evasão.

Palavras-chave: Perfil do Aluno; Evasão Escolar; ETEC; Mineração de Dados Educacionais; Gestão Escolar.

ABSTRACT

Public Technical Schools in the State of São Paulo, known as ETECs, seek to align their educational programs with labor market demands through courses that combine theory and practice. These institutions face challenges related to school dropout, which can be examined through institutional academic data. The case of ETEC Raposo Tavares illustrates this context: only 43.59% of the students who entered the Technical Course in Chemistry in 2023 completed the program in 2024, a percentage below the institutional target of 65%. In this context, Educational Data Mining (EDM) makes it possible to explore these data and identify patterns associated with students at risk of dropout. This study aims to apply Educational Data Mining to identify profiles of students at risk of dropout in the Technical Course in Chemistry at ETEC Raposo Tavares, based on their academic trajectory, and to support school management in the development of educational strategies. The database used in the computational experiments contains information on 185 students enrolled from the first to the third module in 2023 and 2024. The data were extracted from the New Academic System (NSA) and supplemented with records from class council minutes. The experiments followed the phases of the Knowledge Discovery in Databases process: Data Selection, Data Preprocessing, Data Transformation, Educational Data Mining, and Knowledge Interpretation and Evaluation. Logistic Regression, Random Forest, and XGBoost were applied, and Random Forest achieved the best performance. Based on its results, the school management team identified three profiles: High Academic Performance, associated with students who are committed to the course and seek completion; Medium Academic Performance, associated with students who, despite difficulties, seek to complete the course; and Low Academic Performance, associated with a higher risk of dropout. The last profile may be related to social inequalities, lack of family support, vocational misalignment, or a weak sense of institutional identification and belonging. Among the measures selected by the management team, the institutionalization of an internal continuous monitoring protocol stands out. The protocol should connect the indicators to pedagogical practices so that the data can support concrete educational strategies. The study concludes that EDM is a relevant resource for ETEC Raposo Tavares and can support school management in the development of strategies aimed at reducing the risk of dropout.

Keywords: *Student Profile; School Dropout; ETEC; Educational Data Mining; School Management.*

LISTA DE ILUSTRAÇÕES

LISTA DE ABREVIATURAS E SIGLAS

AUC	**Area Under the Curve**
CPS	Centro Paula Souza
ETEC	Escola Técnica Estadual
FN	Falso Negativo
FP	Falso Positivo
KDD	Knowledge Discovery in Databases
MDE	Mineração de Dados Educacionais
MDI	Mean Decrease in Impurity
NSA	Novo Sistema Acadêmico
RF	Random Forest
RL	Regressão Logística
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XGBoost	Extreme Gradient Boosting

LISTA DE SÍMBOLOS

x	Variável independente (atributo de entrada)
y	Variável dependente (classe real)
$\hat{y}$	Valor predito pelo modelo
n	Número de observações (amostras)
i	Índice da observação
j	Índice do atributo
p	Probabilidade estimada da classe positiva
β_0	Intercepto do modelo
$\beta_1 \ldots \beta_p$	Coeficientes das variáveis independentes
$\log$	Função logarítmica natural
$\hat{y}_i$	Valor predito para a i-ésima instância
K	Número total de árvores
L	Função de perda (loss function)
Ω	Termo de regularização
γ (gamma)	Parâmetro de penalização de complexidade
λ (lambda)	Parâmetro de regularização L2
η (eta)	Taxa de aprendizado
p_k	Proporção de elementos da classe k no nó
Σ	Somatório
$\sum$	Operador de soma
$f(x)$	Função preditiva do modelo
$P(y\|x)$	Probabilidade condicional da classe dado x

SUMÁRIO

1 INTRODUÇÃO

A expansão da educação profissional no Estado de São Paulo acompanha um movimento nacional de valorização da formação técnica como alternativa de qualificação e inserção no mercado de trabalho (Centro Paula Souza, 2024).

Práticas voltadas ao ensino técnico e tecnológico buscam diversificar trajetórias formativas, possibilitando o acesso de jovens e adultos a cursos articulados às demandas econômicas e sociais do país (Machado, Ferreira e Cezar, 2021; Pontes et al., 2022).

Nesse cenário, o Centro Paula Souza (CPS) coordena a oferta pública de cursos técnicos no Estado de São Paulo. As Escolas Técnicas Estaduais (ETECs), inseridas nessa estrutura, são responsáveis pela oferta de educação profissional de nível médio e se expandiram ao longo do estado, ampliando o acesso à formação técnica.

Apesar desse alcance, a diversidade de contextos e perfis atendidos impõe desafios à gestão pedagógica e ao acompanhamento das trajetórias escolares, especialmente em relação à permanência e conclusão dos alunos (Centro Paula Souza, 2024).

O acompanhamento dessas trajetórias depende do uso de dados produzidos no ambiente escolar. Os dados na gestão escolar correspondem a registros das atividades acadêmicas dos alunos, incluindo informações de frequência, notas, avaliações, histórico escolar e movimentações ao longo do curso.

Esses registros são produzidos continuamente a partir das atividades pedagógicas e administrativas, compondo a base informacional da escola e sendo organizados em sistemas acadêmicos que concentram registros educacionais e administrativos (Pereira, 2025).

Nas ETECs, essa base de dados é organizada por meio do Novo Sistema Acadêmico (NSA), desenvolvido no âmbito do Centro Paula Souza com o objetivo de padronizar e apoiar as rotinas acadêmicas, diante da expansão da rede e da necessidade de centralização dos registros escolares.

Inicialmente voltado ao controle administrativo, o sistema é o principal ambiente de registro acadêmico das ETECs, reunindo informações como notas, frequência e histórico escolar em uma plataforma única (Centro Paula Souza, 2024).Entre os dados armazenados no NSA estão aqueles relacionados à evasão escolar, um dos desafios presentes nas ETECs, especialmente em cursos técnicos modulares.

Esse fenômeno está associado à descontinuidade da formação, à perda de vagas ofertadas e a impactos nos indicadores de conclusão (Etec Raposo Tavares, 2023).

O caso específico da ETEC Raposo Tavares ilustra esse cenário, pois apenas 43,59% dos alunos que ingressaram no Curso Técnico em Química em 2023 concluíram os estudos em 2024, percentual inferior à meta institucional de 65% (Centro Paula Souza, 2025).

A literatura tem indicado que a evasão em cursos técnicos decorre de padrões mensuráveis na trajetória acadêmica, especialmente relacionados ao rendimento, assiduidade e recorrência de reprovações, que tendem a se intensificar ao longo dos módulos (Musso, Hernández e Cascallar, 2020; Psyridou et al., 2024).

Musso et al. (2020) demonstram que oscilações persistentes no desempenho podem sinalizar risco de evasão, enquanto Psyridou et al. (2024) mostram que a combinação para análise desses indicadores melhora a capacidade de distinguir trajetórias de continuidade e de abandono. Esses trabalhos mostram que a evasão não se manifesta de forma abrupta, mas como resultado de um conjunto de variáveis que, observadas em série, permitem identificar padrões de risco.

Para que esses dados possam ser analisados, torna-se necessário utilizar técnicas capazes de explorar os registros institucionais e identificar padrões que surgem ao longo da trajetória acadêmica para identificar perfis de alunos e estimar risco de evasão. Nesse contexto, Romero e Ventura (2010) definem a Mineração de Dados Educacionais (MDE), subárea vinculada à Inteligência Artificial, com o uso de algoritmos inteligentes voltados à extração de conhecimento a partir de registros acadêmicos.

A análise da trajetória acadêmica por meio da MDE tem sido apontada como uma alternativa para organizar e interpretar informações registradas ao longo do curso, permitindo observar variações em desempenho, assiduidade e progressão, além de possibilitar a análise de perfis e padrões presentes nos registros escolares (Silva e Pereira, 2021; Moura, 2023).

Ao integrar diferentes dimensões da trajetória escolar, as técnicas permitem reconhecer grupos com características semelhantes, bem como alunos com maior probabilidade de evadir no futuro (Romero e Ventura, 2010; Costa et al., 2012).

O uso de técnicas como *Random Forest* (RF), *XGBoost* (*Extreme Gradient Boosting*) e Regressão Logística (RL) é recorrente em estudos que envolvem predição na área educacional, sobretudo por lidarem adequadamente com variáveis diversas, relações não lineares e bases desbalanceadas (Vadakara e Kumar, 2019; Chen e Guestrin, 2016).

Diante deste contexto, considera-se que o desenvolvimento de um trabalho que aplica técnicas de MDE para identificar perfis de alunos com risco de evasão no Curso Técnico em Química da ETEC Raposo Tavares, a partir da trajetória acadêmica, mostra-se relevante frente à necessidade de as escolas se servirem de ferramentas que auxiliem na redução da evasão escolar e, consequentemente, apoiem a gestão.

1.1 QUESTÃO DE PESQUISA

Como aplicar Mineração de Dados Educacionais para identificar perfis de alunos com risco de evasão do Curso Técnico em Química da ETEC Raposo Tavares, com base na trajetória acadêmica, para apoiar a gestão escolar no desenvolvimento de estratégias educacionais?

1.2 OBJETIVO GERAL E ESPECÍFICOS

1.2.1 *Objetivo Geral*

O objetivo geral deste trabalho é aplicar a Mineração de Dados Educacionais para identificar perfis de alunos com risco de evasão do Curso Técnico em Química da ETEC Raposo Tavares, com base na trajetória acadêmica, para apoiar a gestão escolar no desenvolvimento de estratégias educacionais.

1.2.2 *Objetivos Específicos*

Para atingir o objetivo geral foram estabelecidos os seguintes objetivos específicos:

- Organizar os dados acadêmicos dos alunos do Curso de Química da ETEC Raposo Tavares em uma base de dados;
- Preparar a base de dados para a aplicação das técnicas de MDE selecionadas: Regressão Logística, *Random Forest* e *XGBoost*.
- Aplicar e comparar os resultados das técnicas selecionadas.

1.3 **JUSTIFICATIVA DE PESQUISA**

A escolha da ETEC Raposo Tavares e do Curso Técnico em Química se justifica em função de os índices de conclusão do curso, nos anos de 2024 e 2025, ficarem abaixo das metas estabelecidas no Plano Plurianual de Gestão (2023–2027), indicando que, caso esse comportamento se mantenha, os resultados poderão permanecer abaixo do previsto. Esse cenário aponta a necessidade de examinar a trajetória acadêmica dos alunos (Etec Raposo Tavares, 2023).

O curso Técnico em Química recebe alunos majoritariamente jovens, oriundos de escolas públicas, muitos provenientes de contextos socioeconômicos com restrições de acesso a serviços como saúde, emprego e continuidade dos estudos. Esses alunos apresentam metas profissionais vinculadas tanto à inserção no mercado de trabalho quanto ao ingresso no ensino superior, sendo a formação técnica uma possibilidade de qualificação e ampliação de oportunidades educacionais e profissionais (Centro Paula Souza, 2024).

A evasão, nesse contexto, representa a interrupção da trajetória acadêmica do aluno, com reflexos na continuidade dos estudos e na inserção profissional. Para a família, pode significar a perda de uma oportunidade de formação, enquanto, no âmbito institucional, impacta a ocupação de vagas, os indicadores de conclusão e a organização dos cursos nas ETECs (Tinto, 2017; Centro Paula Souza, 2024).

Diante desses desafios, a identificação dos perfis de alunos com risco de evasão pode apoiar a gestão escolar na definição de estratégias educacionais, especialmente voltadas à prevenção e ao acompanhamento das trajetórias acadêmicas.

Destaca-se que a escolha do Curso Técnico em Química também se justifica pela relevância dessa formação para a região onde está localizada a ETEC, que concentra empresas do setor químico e industrial, como Archote Indústria Química, Midland Química do Brasil, Denver Especialidades Químicas, BioChemicals do Brasil, Avanzi Química e Proquitec, entre outras instaladas em Cotia e municípios próximos.

Essas organizações demandam profissionais qualificados para atividades técnicas e operacionais, o que indica a possibilidade de inserção dos alunos e egressos no mercado de trabalho local, considerando a relação entre a formação ofertada e as demandas produtivas da região.

A justificativa também se baseia na revisão da literatura, que pode ser encontrada no Capítulo 3, na qual se observa que a aplicação da MDE em escolas técnicas, especialmente no contexto paulista é escassa, abrindo espaço para ampla pesquisa.

Boa parte dos trabalhos está concentrada nos estudos sobre ensino superior ou em ambientes virtuais de aprendizagem, como nos estudos de Waheed, Khan e Ali (2020), Musso, Hernández e Cascallar (2020) e Silva, Souza e Pereira (2025), cujos dados e contextos diferem daqueles da educação profissional de nível médio. Destaca-se o estudo de Andrelo (2022) aplicado na ETEC Paulistano, porém na identificação do perfil do egresso.

A MDE foi selecionada devido a sua capacidade de lidar com múltiplas variáveis e com bases desbalanceadas, características presentes em estudos sobre evasão (Waheed, Khan e Ali, 2020).

Nesse cenário, os dados organizados no NSA, que incluem informações de matrícula, frequência, notas e registros da trajetória escolar, constituem uma base compatível com estudos em MDE. Esses registros, estruturados ao longo do tempo, permitem a aplicação de técnicas voltadas à identificação de padrões em contextos educacionais (Romero e Ventura, 2010; Baker e Inventado, 2014).

Estudos em MDE indicam que registros da trajetória escolar podem ser utilizados na análise da permanência dos alunos. Trabalhos como os de Romero e Ventura (2010), Musso et al. (2020) e Silva et al. (2025) utilizam variáveis como desempenho, frequência e progressão escolar para identificar padrões relacionados à evasão.

Nesse sentido, este trabalho aplica essas técnicas de MDE ao contexto da educação profissional de nível médio, com o objetivo de identificar perfis de alunos com risco de evasão a partir de dados institucionais.

Finalmente, a motivação do trabalho decorre da vivência profissional e pessoal da autora, professora da ETEC e do curso selecionado, por meio da observação dos impactos da evasão na trajetória acadêmica dos alunos e nos indicadores institucionais.

1.4 **DELIMITAÇÃO DO TEMA DE PESQUISA**

Este trabalho delimita-se à aplicação da MDE sobre os registros acadêmicos dos alunos regularmente matriculados no Curso Técnico em Química da ETEC Raposo Tavares, abrangendo o período entre o primeiro e o terceiro módulo dos anos de 2023 a 2024.

A análise concentra-se nas variáveis disponíveis no Novo Sistema Acadêmico (NSA), especialmente aquelas relacionadas ao desempenho, à frequência e à progressão escolar, por serem elementos que estruturam a trajetória acadêmica dos alunos. As variáveis sociais, familiares e vocacionais não integram o escopo da análise, pois não estão disponíveis no sistema utilizado.

A adoção da MDE justifica-se pela sua capacidade de lidar com bases de dados institucionais que, embora não sejam extensas em volume, apresentam múltiplas variáveis e distribuição desbalanceada entre as categorias de permanência e evasão. Diferentemente de análises estatísticas tradicionais, a MDE emprega técnicas capazes de reconhecer padrões ocultos, relacionar variáveis ao longo da trajetória acadêmica e estimar a probabilidade de eventos educacionais relevantes, como risco de evasão e oscilações de desempenho (Baker, 2009).

Esse diferencial torna sua aplicação especialmente pertinente no contexto das ETECs, onde identificar a dificuldade dos alunos exige técnicas capazes de analisar diferentes dimensões do registro escolar (Romero e Ventura, 2010).

O campo de análise situa-se no uso de dados institucionais para apoiar práticas de gestão escolar na educação profissional pública. Dessa forma, o trabalho enfatiza a articulação entre MDE, gestão educacional e análise de fatores associados à permanência, evasão e evolução do desempenho ao longo dos módulos.

O nível de análise restringe-se à aplicação de algoritmos de MDE voltados à identificação de perfis acadêmicos e estimativas de risco de evasão, considerando padrões registrados na trajetória acadêmica dos alunos.

A delimitação metodológica compreende a construção e a avaliação das técnicas preditivas, buscando antecipar situações que possam demandar intervenção pedagógica ou administrativa (Fayyad et al., 1996; Romero et al., 2014). A escolha pela MDE fundamenta-se no processo de Descoberta de Conhecimento em Bases de Dados (KDD), que permite extrair, transformar e estruturar informações educacionais de maneira sistemática.

Esse processo favorece análises relacionadas à permanência, ao desempenho e à progressão, contribuindo para a organização dos dados educacionais (Silva, 2016).

A seleção das técnicas Regressão Logística, Random Forest e XGBoost apoia-se na revisão da literatura, que registra sua aplicação em tarefas de classificação de dados educacionais. Essas técnicas foram selecionadas pela capacidade de tratar múltiplas variáveis acadêmicas, distribuição desbalanceada entre as classes e relações não lineares entre atributos (Romero; Ventura, 2010; Musso et al., 2020; Silva et al., 2025).

A Regressão Logística foi incluída por sua interpretabilidade e por constituir uma referência em tarefas de classificação binária. Sua estrutura linear permite identificar a contribuição de cada variável para a estimativa de risco, o que facilita a comunicação dos resultados à equipe de gestão escolar (Romero; Ventura, 2010).

O Random Forest foi selecionado por seu desempenho em cenários com variáveis diversas e classes desbalanceadas, características presentes neste trabalho. A técnica combina múltiplas árvores de decisão treinadas sobre subconjuntos aleatórios dos dados, reduz a variância do modelo e diminui o risco de sobreajuste (Breiman, 2001).

O XGBoost foi selecionado por sua eficiência computacional e por apresentar resultados competitivos em tarefas de classificação com bases complexas. O algoritmo incorpora regularização explícita para controlar a complexidade das árvores e trata valores ausentes de forma automática (Chen; Guestrin, 2016).

Técnicas como Redes Neurais Profundas (Deep Learning), Support Vector Machines (SVM) e algoritmos de agrupamento não foram adotadas. As Redes Neurais Profundas demandam grandes volumes de dados para uma generalização adequada e apresentam baixa interpretabilidade para uso na gestão escolar. O SVM é sensível à escala e ao ajuste dos parâmetros do kernel e requer procedimentos adicionais diante de classes desbalanceadas. Os algoritmos de agrupamento, por serem não supervisionados, não atendem ao objetivo de classificação binária proposto (Hastie; Tibshirani; Friedman, 2009).

1.5 **ORGANIZAÇÃO DO TRABALHO**

Este trabalho está estruturado em cinco capítulos. O Capítulo 1 apresenta a introdução, a questão de pesquisa, os objetivos, a justificativa, a delimitação e a organização do trabalho. O Capítulo 2 reúne a fundamentação teórica. O Capítulo 3 apresenta a revisão da literatura e os materiais e métodos. O Capítulo 4 expõe e discute os resultados. O Capítulo 5 apresenta a conclusão.

2 FUNDAMENTAÇÃO TEÓRICA

Neste capítulo, apresenta-se a fundamentação teórica dos principais temas deste trabalho: Escolas Técnicas Públicas, Perfil do Aluno da ETEC, Evasão Escolar na ETEC, Gestão Escolar na ETEC e Mineração de Dados Educacionais.

2.1 ESCOLAS TÉCNICAS PÚBLICAS DO ESTADO DE SÃO PAULO

As Escolas Técnicas Públicas Estaduais (ETECs) representam, no Estado de São Paulo, uma alternativa à formação tradicional voltada exclusivamente ao ensino superior, oferecendo cursos técnicos que articulam teoria, prática e desenvolvimento de competências profissionais (Centro Paula Souza, 2022).

A atuação dessas instituições está alinhada às políticas do Centro Paula Souza, que orientam a oferta de cursos voltados aos setores industrial, comercial, de serviços e da tecnologia, atendendo às demandas regionais e contribuindo para a preparação técnica dos alunos (Centro Paula Souza, 2024).

A inserção na comunidade, associada à integração com o setor produtivo, amplia as possibilidades formativas e estabelece vínculos entre escola e contexto social. No entanto, as instituições enfrentam desafios relacionados à permanência e ao rendimento dos alunos, que apresentam perfis diversos e estão inseridos em condições socioeconômicas heterogêneas (Tinto, 2017).

Nesse contexto, o acompanhamento da trajetória acadêmica do aluno é um elemento relevante para identificar necessidades, orientar intervenções e promover condições mais favoráveis de aprendizagem (Romero; Ventura, 2020).

Os sistemas de informação acadêmica desempenham papel estratégico nesse processo ao centralizar, organizar e disponibilizar dados referentes ao desempenho, frequência e progressão dos alunos (Nguyen et al., 2021).

Ao reduzir tarefas manuais e permitir o acesso rápido a registros atualizados, esses sistemas facilitam o trabalho da gestão escolar e oferecem subsídios para análises mais consistentes sobre o percurso formativo; ao reunir esses registros em um mesmo ambiente, favorecem o acesso às informações e o apoio à tomada de decisões por parte da gestão. Sua utilização, entretanto, requer não apenas domínio técnico, mas também a capacidade de interpretar os dados e integrá-los às práticas pedagógicas e administrativas (Carleto, 2011).

Nas ETECs, o Novo Sistema Acadêmico (NSA) concentra dados que compõem a trajetória escolar, como notas, faltas e resultados de recuperação. As informações são atualizadas ao longo do semestre, o que possibilita acompanhar o avanço dos alunos nas diferentes etapas do curso. Além de armazenar os registros,

o NSA auxilia na organização curricular e no fechamento de módulos. Embora o NSA não realize análises automatizadas, ele fornece a base necessária para estudos mais aprofundados, incluindo técnicas de análise de dados (Centro Paula Souza, 2025).

Nesse contexto, o NSA se apresenta como uma ferramenta que organiza as informações acadêmicas nas ETECs, facilita o acompanhamento dos alunos e apoia as rotinas pedagógicas. No entanto, suas limitações, como a ausência de análises automatizadas e certa rigidez, isto é, a baixa capacidade de adaptação às necessidades da escola ou do professor, abrem espaço para o uso de tecnologias como a Mineração de Dados Educacionais, que possibilita uma análise mais aprofundada dos dados e sustenta decisões mais bem fundamentadas (Centro Paula Souza, 2025).

A informatização dos processos escolares, por sua vez, não se limita à adoção de ferramentas tecnológicas. Ela envolve a revisão de práticas institucionais, a adequação de rotinas de trabalho e a articulação entre tecnologia e currículo. Quando esses elementos não estão presentes, o uso dos sistemas tende a ser fragmentado, restrito ao registro burocrático e com pouco impacto na melhoria da aprendizagem (Kirkwood; Price, 2014).

Scherer e Brito (2020) ressaltam que a tecnologia educacional produz resultados quando incorporada a uma cultura de gestão que utiliza dados para compreender necessidades, orientar decisões e monitorar o avanço dos alunos.

Assim, as ETECs, por focarem em um ambiente de formação técnica de habilidades e competências, podem se beneficiar de processos de informatização integrados a práticas de gestão e ensino.

O uso articulado dos sistemas institucionais possibilita ampliar o acompanhamento acadêmico, reconhecer padrões de risco, apoiar ações preventivas e profissionalizar o desenvolvimento dos alunos, contribuindo para a melhoria da qualidade do ensino técnico público no estado (Kirkwood; Price, 2014).

2.1.1 *Perfil do Aluno da ETEC*

O aluno matriculado no Curso Técnico em Química da ETEC Raposo Tavares é majoritariamente jovem, com idade entre 17 e 20 anos, proveniente em grande parte de escolas públicas e inserido em contextos diversificados. A busca pela formação técnica geralmente está associada ao desejo de ampliar oportunidades profissionais, qualificar-se para o mercado de trabalho e, simultaneamente, criar condições para o acesso ao ensino superior (Centro Paula Souza, 2022).

A distribuição de gênero apresenta equilíbrio ao longo dos anos, ainda que certos períodos indiquem participação levemente superior de alunos do sexo masculino. Esses alunos convivem com desafios diversos e, em alguns casos, às limitações estruturais da escola, situações que podem interferir em sua trajetória acadêmica. Apesar disso, observam-se dedicação e interesse pela conclusão do curso, motivados pelo reconhecimento do potencial da formação técnica para ampliar perspectivas profissionais (Centro Paula Souza, 2024).

Para muitos alunos, o curso representa uma possibilidade concreta de mobilidade social, seja através da continuidade dos estudos, seja pela inserção direta no mercado de trabalho. Parte dos concluintes busca vagas em universidades públicas e privadas, utilizando os conhecimentos adquiridos como base para cursos de áreas correlatas, especialmente nos campos das ciências exatas e da saúde. Outros optam por ingressar rapidamente em atividades profissionais, como laboratórios de análises químicas, indústrias, empresas de saneamento ou setores que demandam formação técnica especializada (Centro Paula Souza, 2022).

A ETEC Raposo Tavares foi criada para atender à demanda por formação técnica em uma região marcada pela presença de importantes empresas instaladas ao longo do eixo da Rodovia Raposo Tavares, como a Eurofarma, a Natura, a Schneider Electric e a Votorantim Cimentos, além de indústrias dos setores alimentício, logístico e tecnológico presentes na região de Cotia e entorno.

A oferta de cursos como Administração, Logística, Química e Desenvolvimento de Sistemas apresenta correspondência com esse perfil produtivo, mostrando o alinhamento entre a formação oferecida e as demandas do mercado local. Essas características indicam que a trajetória dos alunos está inserida em condições diversas, que podem influenciar sua permanência e desempenho ao longo do curso (Centro Paula Souza, 2022).

Embora não haja dados públicos detalhados por unidade sobre a contratação de egressos, indicadores do ensino técnico paulista apontam a inserção profissional após a conclusão dos cursos, o que reforça a relação entre a formação técnica e o setor produtivo (São Paulo, 2023).

2.1.2 *Evasão na ETEC*

A evasão escolar aparece de forma recorrente na educação profissional, especialmente em cursos técnicos de nível médio, nos quais a trajetória acadêmica combina componentes gerais e conteúdos específicos da área. A literatura descreve o abandono como um fenômeno complexo, associado a diferentes dimensões que influenciam a trajetória dos alunos e produzem padrões variados conforme o contexto e o perfil do aluno (Feitosa, 2020).

Em instituições públicas de educação profissional, como as ETECs, a evasão costuma se relacionar a dificuldades no acompanhamento das disciplinas técnicas e a trajetórias escolares anteriores marcadas por interrupções ou lacunas de aprendizagem. Estudos indicam que indicadores acadêmicos, como frequência, domínio dos conteúdos introdutórios e participação nas atividades, influenciam o desempenho ao longo dos módulos. Oscilações nesses indicadores aumentam a possibilidade de descontinuidade, sobretudo em cursos que exigem progressão constante (Musso et al., 2020).

No contexto da ETEC, o acompanhamento da evasão tem recebido atenção em razão de taxas de conclusão que, em alguns cursos, ficam abaixo das metas projetadas. Componentes curriculares que exigem maior domínio de matemática, química ou física apresentam índices mais elevados de abandono, especialmente entre alunos que ingressam com dificuldades nessas áreas. Estudos também mencionam a carga horária extensa e a necessidade de conciliar diferentes responsabilidades como fatores que podem comprometer a permanência (Centro Paula Souza, 2024).

A trajetória dos alunos reúne informações que permitem observar movimentos relacionados à continuidade no curso. Variáveis como rendimento, assiduidade, progressão entre módulos e participação indicam momentos de maior vulnerabilidade ao longo da trajetória. Combinações como baixa frequência, desempenho irregular e

repetências, quando acumuladas, costumam anteceder a evasão (Silva e Pereira, 2021).

Além dos aspectos pedagógicos, elementos institucionais também influenciam a permanência. A organização curricular, as condições de oferta, as práticas de acompanhamento e o suporte ao aluno configuram fatores relevantes nesse processo. A literatura indica que a ausência de acompanhamento contínuo resulta em trajetórias menos estáveis, enquanto o uso de informações acadêmicas contribui para ações mais alinhadas às necessidades dos alunos (Lück, 2009; Silva et al., 2020).

Nesse cenário, análises que consideram a trajetória ao longo do tempo permitem observar a progressão entre módulos e identificar mudanças no envolvimento dos alunos. A evasão não se configura como um evento isolado, mas como um processo que apresenta sinais em indicadores como frequência, desempenho e continuidade entre etapas (Romero e Ventura, 2010).

2.1.3 *Gestão Escolar nas ETECs*

Em um cenário de transformações tecnológicas e crescente disponibilidade de dados educacionais, tem-se ampliado a expectativa de que as decisões escolares sejam apoiadas por evidências, para análises mais precisas e ações mais coerentes com as necessidades da comunidade atendida (Silva, 2017).

Lima (2016) destaca que a gestão escolar contemporânea requer instrumentos capazes de interpretar o perfil do aluno, permitindo uma visão mais ampla e da trajetória acadêmica. Nesse contexto, a relação entre gestão e análise de dados é apresentada como uma aproximação que envolve o tratamento de informações referentes ao desempenho e à permanência, compondo parte das práticas utilizadas no âmbito institucional.

Os registros históricos constituem uma fonte da evolução acadêmica ao longo do curso, pois reúnem informações acumuladas sobre frequência, desempenho e continuidade entre módulos. Os dados reunidos ao longo do percurso acadêmico apresentam variações que podem ser observadas em diferentes períodos. Dessa forma, tais informações compõem parte do material considerado em análises realizadas pela instituição (Silva et al., 2020).

A gestão escolar nas ETECs envolve a coordenação de dimensões administrativas, pedagógicas e organizacionais que sustentam o funcionamento da

instituição e a aprendizagem dos alunos. Nesse contexto, a gestão atua articulando recursos, pessoas, práticas pedagógicas e processos institucionais, de modo a promover condições que favoreçam o desenvolvimento acadêmico e a permanência no curso (Centro Paula Souza, 2022).

É importante considerar que a gestão escolar nas ETECs opera em um contexto marcado por diversidade social, expectativas formativas distintas e desigualdades educacionais. Nesse aspecto, identificar o perfil dos alunos e o que pode comprometer sua trajetória acadêmica exige análises sensíveis às especificidades da instituição. O uso de dados torna-se, portanto, um recurso para orientar intervenções pedagógicas e administrativas ajustadas às necessidades locais (Centro Paula Souza, 2022).

Nas ETECs, esses registros ajudam a reconhecer situações que podem levar ao afastamento. Por isso, a Mineração de Dados Educacionais é pertinente, pois permite analisar esses dados e identificar possíveis sinais de risco.

Por fim, a literatura mencionou o uso de informações acadêmicas no contexto da gestão escolar, com direcionamento a aspectos relacionados ao acompanhamento acadêmico. Nessa perspectiva, tais informações foram consideradas na análise da trajetória dos alunos nas ETECs, com referência à aplicação de métodos analíticos sobre dados educacionais (Pereira, 2018).

2.2 **MINERAÇÃO DE DADOS EDUCACIONAIS**

A Mineração de Dados é uma subárea da Inteligência Artificial voltada à identificação de padrões, relações e estruturas relevantes em bases de dados por meio de algoritmos capazes de analisar informações complexas e extrair conhecimento de forma automática (Fayyad et al., 1996).

Essa área está diretamente articulada ao Aprendizado de Máquina, uma vez que fornece os padrões, atributos e relações que servem de base para a construção de modelos capazes de realizar tarefas como previsão, classificação e agrupamento (Romero e Ventura, 2010).

A Mineração de Dados permite identificar relações que não são facilmente observadas em análises convencionais, favorecendo interpretações mais detalhadas sobre o comportamento dos dados (Baker e Siemens, 2014).

Ela é considerada uma das fases do processo de Descoberta de Conhecimento em Bases de Dados ou em inglês *Knowledge Discovery in Databases* (KDD).

O processo de Descoberta de Conhecimento em Bases de Dados (Knowledge Discovery in Databases - KDD) corresponde a um processo não trivial, iterativo e sistemático de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em bases de dados (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

O processo é organizado em cinco fases que incluem a Seleção dos Dados, o Pré-Processamento dos Dados, a Transformação dos Dados e a Interpretação e Avaliação dos Resultados (Fayyad et al., 1996).

Na fase de Seleção dos Dados, define-se o conjunto de dados a ser analisado. O Pré-Processamento dos Dados envolve o tratamento de inconsistências, valores ausentes e ruídos, visando assegurar a qualidade e a confiabilidade dos dados para as análises subsequentes .

A Transformação dos Dados tem por objetivo modificar, reduzir ou criar atributos que tornem os dados mais adequados às tarefas de mineração. Isso pode envolver normalização, engenharia de atributos ou redução de dimensionalidade, dependendo da complexidade da base e do objetivo analítico.

A fase de Mineração de Dados constitui o núcleo analítico do processo de KDD, pois, nesse momento, os dados preparados são submetidos a algoritmos capazes de identificar padrões, relações e tendências que não se evidenciam por meio de análises descritivas tradicionais (Fayyad et al., 1996).

Essa fase envolve decisões metodológicas importantes, como a escolha dos algoritmos adequados, a definição de parâmetros e a validação dos resultados, fatores que influenciam diretamente a qualidade dos padrões identificados e a confiabilidade das inferências produzidas (Witten et al., 2016).

No contexto do KDD, as fases de mineração correspondem a categorias de análise definidas conforme o objetivo da extração de conhecimento, ou seja, o

tipo de padrão ou informação que se pretende identificar nos dados. Entre as principais classificações, destaca-se a classificação, que atribui rótulos a novos registros com base em padrões previamente aprendidos; a regressão, voltada à previsão de valores contínuos; e a clusterização, que identifica grupos de dados com características semelhantes (Han et al., 2012).

Também se incluem a descoberta de regras de associação, que identifica relações frequentes entre variáveis; a detecção de anomalias, que aponta padrões

incomuns; e a sumarização, que produz descrições compactas dos dados (Tan et al., 2005; Witten et al., 2016).

Dessa forma, a escolha adequada das classificações e técnicas, aliada à coerência com os objetivos da pesquisa e à correta interpretação dos padrões identificados, condiciona a geração de conhecimento aplicável (Siemens, 2013).

No contexto educacional, a Mineração de Dados recebe a denominação de Mineração de Dados Educacionais (MDE), que foca a identificação de aspectos do processo de aprendizagem, padrões acadêmicos ou previsões de risco as quais podem incluir queda de desempenho, retenção ou evasão (Romero e Ventura, 2010). A MDE é descrita na literatura como um conjunto de técnicas de análise aplicadas a dados acadêmicos. Romero e Ventura (2010) definem a MDE como

uma especialização da mineração de dados voltada ao ambiente educacional, com foco na identificação de padrões presentes nos registros institucionais.

Baker e Inventado (2014) caracterizam a MDE como um processo que envolve algoritmos voltados à identificação de relações entre variáveis educacionais, incluindo a possibilidade de realizar inferências e previsões com base em comportamentos observados.

Waheed et al.(2020) utilizaram técnicas de *Deep Learning* em ambientes virtuais de aprendizagem, examinando padrões de participação associados ao abandono. Hernández e Cascallar (2020) analisaram trajetórias acadêmicas com base na evolução do desempenho, explorando combinações de variáveis cognitivas e comportamentais para fins preditivos.

Os resultados apresentados por esses trabalhos diferem quanto às variáveis utilizadas nas técnicas. Waheed et al. (2020) trabalharam com indicadores de engajamento e interação em plataformas digitais, enquanto Musso et al. (2020) utilizaram dados referentes ao histórico acadêmico. As diferenças metodológicas destacam a influência do tipo de dado disponível na construção dos modelos.

Além disso, Musso et al. (2020) utilizaram técnicas de aprendizado de máquina para mapear trajetórias de alunos de ensino médio técnico, identificando cinco perfis derivados de variáveis cognitivas, socioemocionais e comportamentais. Andrelo (2022), em trabalho realizado em uma ETEC, empregou a MDE para identificar o perfil dos egressos com base em questionários e registros institucionais, explorando seu uso no apoio à gestão.

Silva, Souza e Pereira (2025) aplicaram algoritmos de classificação para estimar a probabilidade de evasão em bases escolares. Psyridou et al. (2024) analisaram modelos que combinam notas, frequência e engajamento, mostrando variações no desempenho preditivo conforme as variáveis consideradas. Elbouknify et al. (2025) discutem que a constituição dos modelos depende do tipo e da amplitude dos dados disponíveis, considerando que diferentes conjuntos de variáveis produzem diferentes resultados analíticos.

Entre os algoritmos aplicados à fase de classificação, destacam-se a Regressão Logística, o *Random Forest e o XGBoost*. Essas técnicas são amplamente utilizadas em estudos de predição de evasão devido à capacidade de lidar com múltiplas variáveis, dados heterogêneos e cenários de desbalanceamento. Trabalhos recentes indicam o uso combinado dessas técnicas em modelos preditivos, com destaque para métodos baseados em árvores, como *Random Forest e XGBoost*, que apresentam bom desempenho em bases complexas e de grande volume de dados (Cardoso e Nunes, 2025).

Dessa forma, a classificação constitui uma tarefa adequada para análises que buscam estimar o risco de evasão a partir de informações acadêmicas. Ao utilizar registros coletados ao longo do curso, essa técnica possibilita observar como diferentes modelos interpretam os dados e como esses padrões podem ser utilizados para compreender movimentos associados à permanência ou ao desligamento (Baker e Siemens, 2014).

Alshanqiti e Namoun (2020) desenvolveram um modelo híbrido que combina técnicas de regressão e classificação para prever desempenho acadêmico, com acurácia de 89% na identificação de alunos em risco. Waheed et al. (2020) aplicaram redes LSTM (*Long Short-Term Memory*, ou Memória de Curto e Longo Prazo) a dados de ambientes virtuais de aprendizagem, analisaram padrões temporais de interação e registraram acurácia de 93% na previsão de evasão, em comparação com valores inferiores obtidos por Regressão Logística e Árvores de Decisão.

Hoyos e Caicedo-Castro (2024) propuseram uma arquitetura baseada em redes recorrentes associadas a mecanismos de atenção para analisar dados multimodais, incluindo registros acadêmicos, textos qualitativos e interações digitais. Silva et al. (2025) aplicaram técnicas de aprendizado de máquina em bases institucionais de escolas técnicas federais, com acurácia de 92% na previsão de

evasão e uso de variáveis como frequência inicial, desempenho em componentes curriculares e participação em atividades.

Silva e Pereira (2021) também empregaram Árvores de Decisão e *Random Forest* para estimar risco de evasão, utilizando dados de frequência, desempenho e participação. Kloft et al. (2014) analisaram trajetórias de participação em cursos online por meio de séries temporais e modelos probabilísticos, utilizando padrões de engajamento ao longo do tempo como base para estimativas de abandono.

A literatura indica que a MDE é aplicada em tarefas como previsão de desempenho, identificação de alunos em risco e modelagem de trajetórias educacionais, fundamentando o uso de métodos de classificação e agrupamento (Romero e Ventura, 2010; Siemens e Long, 2011).

A aplicação de técnicas de classificação mostra-se adequada quando o objetivo consiste em estimar desfechos binários, como a condição de evasão ou permanência, a partir de registros acadêmicos. Métodos como Regressão Logística, *Random Forest* e *XGBoost* têm sido amplamente utilizados em estudos educacionais, sobretudo pela capacidade de lidar com múltiplas variáveis e com bases de dados desbalanceadas, característica frequente em análises sobre evasão escolar (Fernández et al., 2018).

Esses algoritmos foram utilizados para a geração de estimativas de risco associadas aos atributos acadêmicos e para a produção de classificações relacionadas ao acompanhamento acadêmico, sendo mencionados na literatura em estudos voltados à identificação precoce de alunos em risco (Baker e Siemens, 2014; Costa, 2022).

2.2.1 *TÉCNICAS DE MINERAÇÃO DE DADOS*

A seguir, são descritas as técnicas de MDE adotadas neste trabalho: Random Forest, XGBoost e Regressão Logística.

a) Random Forest

Random Forest (RF) é um algoritmo que utiliza várias árvores de decisão construídas a partir de subconjuntos aleatórios dos dados. Ao final do processo, as previsões das árvores são reunidas por meio de votação, gerando um resultado coletivo (Vadakara; Kumar, 2019).

Cada árvore é construída a partir de uma amostra aleatória dos dados, e a predição final é obtida por votação, no caso de classificação, conforme Equação (1):

$$\hat{y}(x) = \text{mode}\{h_1(x), h_2(x), \ldots, h_B(x)\} \quad (1)$$

Fonte: A autora (2026).

Ou por média, no caso de regressão, conforme Equação (2):

$$\hat{y}(x) = \frac{1}{B}\sum_{b=1}^{B} h_b(x) \quad (2)$$

Fonte: A autora (2026).

Esse processo reduz a variância do modelo e diminui a chance de sobreajuste, pois erros cometidos por árvores individuais tendem a ser compensados pela média do conjunto.

A RF incorpora ideias do Bagging (Breiman, 1996), no qual diferentes subconjuntos aleatórios da base são usados para treinar cada árvore. Além da amostragem dos dados, a técnica também seleciona aleatoriamente um conjunto menor de variáveis em cada divisão, o que reduz a correlação entre as árvores e contribui para melhorar a capacidade de generalização do ensemble. Essa aleatoriedade adicional é um dos principais fatores que diferencia a *Random Forest* de outros métodos baseados em árvores e aumenta sua robustez diante de variações na base de treinamento.

A literatura também mostra a relação da RF com o *Boosting*, embora os métodos sejam distintos: no *Boosting*, exemplos mal classificados recebem pesos maiores ao longo do treinamento (Freund e Schapire, 1997), enquanto a RF mantém pesos iguais e busca diversidade entre as árvores (Trisanto et al., 2021).

O desempenho da RF depende da configuração de hiperparâmetros, o número de árvores (n_estimators), a profundidade máxima (max_depth), o número mínimo de amostras por folha (min_samples_leaf) e a quantidade de variáveis avaliadas por divisão (max_features).

Ajustes inadequados podem levar ao aumento do tempo de processamento ou à perda de desempenho, motivo pelo qual a validação cruzada é recomendada (Vadakara e Kumar, 2019).

A RF também permite estimar a importância das variáveis por meio da redução média de impureza conforme Equação (3):

$$\mathrm{MDI}(j) = \sum_{t \in T:\, j(t) = j} p(t)\, \Delta i(t) \tag{3}$$

Fonte: A autora (2026).

Em que p(t) corresponde à proporção de amostras que alcançam o nó t, e Δi(t) representa a redução de impureza produzida pela divisão. Embora útil, essa medida não indica relações de causa e efeito e pode ser influenciada pela correlação entre os atributos (Strobl et al., 2007).

Quanto ao pré-processamento, a RF é robusta a *outliers* (valores atípicos), valores ausentes e dados categóricos codificados, mas ainda assim pode se beneficiar de normalização, limpeza dos dados e engenharia de atributos. Em bases desbalanceadas, recomenda-se o uso de pesos de classe ou amostragem estratificada para evitar que a classe majoritária domine o processo de classificação (Trisanto et al., 2021).

A RF também pode ser combinada com técnicas de seleção de atributos, métodos híbridos ou outros ensembles quando se busca comparar desempenho entre algoritmos ou explorar diferentes perspectivas analíticas. Essa flexibilidade ajuda a adaptar o método a diferentes contextos e tipos de dado (Vadakara e Kumar, 2019).

b) *Extreme Gradient Boosting*

Extreme Gradient Boosting (XGBoost) é uma técnica baseada em árvores de decisão treinadas em sequência, na qual cada nova árvore busca reduzir os erros das anteriores. Esse princípio deriva do *gradient boosting*, no qual vários modelos simples são somados para compor um modelo mais robusto (Friedman, 2001). O processo é guiado por uma função objetivo que combina o erro de previsão com a regularização do modelo (Chen e Guestrin, 2016).

A função objetivo do XGBoost busca minimizar a perda e, ao mesmo tempo, controlar a complexidade das árvores por meio do termo de regularização $\Omega(f_k)$. Esse equilíbrio reduz o risco de sobreajuste e aumenta a estabilidade das previsões.

A equação da função objetivo do *XGBoost*, apresentada na Equação (4), mostra que o modelo busca simultaneamente minimizar a função de perda l(yi, ŷi) e

controlar a complexidade das árvores por meio do termo de regularização Ω(fk). Esse equilíbrio reduz o risco de sobreajuste e melhora a estabilidade das previsões.

$$\mathcal{L} = \sum_{i=1}^{n} \ell(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

Fonte: A autora (2026).

XGBoost introduz mecanismos como regularização explícita, processamento eficiente de dados esparsos, tratamento automático de valores ausentes e controle detalhado sobre a profundidade e o crescimento das árvores. Esses elementos tornam o algoritmo robusto e adequado para bases heterogêneas.

Durante o treinamento, cada divisão potencial da árvore é avaliada por uma métrica de ganho, responsável por quantificar o quanto essa divisão contribui para a melhoria do modelo. De acordo com Chen e Guestrin (2016), esse cálculo é formalizado pela fórmula do ganho da divisão, apresentada na Equação (5).

$$\text{Gain} = \frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda} + \frac{G_R^2}{H_R+\lambda} - \frac{(G_L+G_R)^2}{H_L+H_R+\lambda}\right] - \gamma \quad (5)$$

Fonte: A autora (2026).

Na qual:

- G_l e G_r: somas dos gradientes dos lados esquerdo e direito;
- H_l e H_r: somas dos hessianos dos lados esquerdo e direito;
- λ e γ: parâmetros de regularização.

Essa fórmula indica quando vale a pena dividir um nó: quanto maior o ganho, mais relevante a divisão.

A interpretação dos modelos baseou-se nas métricas de desempenho, nas matrizes de confusão e nas curvas de avaliação. Para reduzir o risco de sobreajuste, aplicou-se a validação cruzada durante o ajuste dos modelos e a otimização dos hiperparâmetros. Esses procedimentos permitiram verificar a estabilidade das técnicas e selecionar as configurações mais adequadas à base de dados utilizada.

Em relação aos dados, embora o *XGBoost* trate valores ausentes automaticamente, práticas como engenharia de atributos, codificação adequada de categorias e tratamento de outliers podem influenciar positivamente os resultados.

Em bases desbalanceadas, o ajuste de pesos e o uso de métricas como F1 ou precision–recall oferecem avaliações mais apropriadas do desempenho do modelo.

Por fim, é importante destacar que o *XGBoost* não corrige problemas de qualidade da base ou viés amostral. Portanto, seus resultados devem ser interpretados de acordo com o contexto e acompanhados de análise de erros, validação e documentação metodológica adequada (Friedman, 2001; Chen e Guestrin, 2016; Lundberg e Lee, 2017).

c) Regressão Logística

A Regressão Logística é uma técnica estatística de classificação supervisionada utilizada para estimar a probabilidade de ocorrência de um evento binário, como evasão ou permanência, com base em variáveis explicativas inseridas em uma função logística. Sua aplicação é reconhecida em áreas como ciências sociais, epidemiologia e análise educacional (Hosmer; Lemeshow; Sturdivant, 2013).

Seu desenvolvimento teórico está associado à formulação do modelo logit por Cox (1958) e à sistematização apresentada em obras como a de Hosmer, Lemeshow e Sturdivant (2013). O logit corresponde ao logaritmo da razão de chances do evento e permite relacionar uma combinação linear das variáveis explicativas à probabilidade estimada.

A técnica baseia-se na combinação linear dos atributos, transformada pela função logística, que mapeia valores reais para o intervalo entre 0 e 1. A probabilidade do evento Y = 1 é expressa conforme a Equação (6).

$$P(Y = 1 \mid X) = \frac{1}{1 + \exp\left[-\left(\beta_0 + \sum_{j=1}^{p} \beta_j x_j\right)\right]} \qquad (6)$$

Fonte: A autora (2026).

Essa transformação permite interpretar os coeficientes β em termos das alterações nas chances (*odds*) de ocorrência do evento. Cada coeficiente representa a variação no logaritmo da razão de chances associada à variável explicativa correspondente (Hosmer; Lemeshow; Sturdivant, 2013).

O treinamento da técnica consiste em ajustar os parâmetros β por meio da minimização da função de perda logística, apresentada na Equação (7).

$$L(\boldsymbol{\beta}) = -\sum_{i=1}^{n} [y_i \ln(p_i) + (1 - y_i)\ln(1 - p_i)] \quad (7)$$

Fonte: A autora (2026).

Em que pi representa a probabilidade estimada para cada observação. O ajuste é realizado por métodos de otimização, como gradiente descendente ou algoritmos quasi-Newton, que buscam maximizar a verossimilhança das previsões.

A Regressão Logística também admite mecanismos de regularização, como penalidades L1 e L2, que reduzem sobreajuste e estabilizam os coeficientes em bases pequenas ou com variáveis correlacionadas. Esses recursos tornam a técnica completa para diferentes tamanhos de amostra e condições dos dados (Hastie, Tibshirani e Friedman, 2009).

Apesar de sua estrutura simples, a técnica requer atenção ao desbalanceamento entre classes, situação comum em estudos de evasão, podendo utilizar pesos de classe, reamostragem ou métricas alternativas como F1-score ou AUC-ROC, que melhoram o desempenho em cenários assimétricos (Sokolova; Lapalme, 2009).

Por fim, sua aplicação neste trabalho complementou as demais técnicas ao incorporar uma perspectiva baseada em relações lineares entre variáveis, sem prejuízo à comparação entre diferentes modelos preditivos.

2.3 MÉTRICAS PARA AVALIAÇÃO DE DESEMPENHO DE CLASSIFICADORES

A avaliação de desempenho em tarefas de classificação requer métricas que quantifiquem a distribuição de acertos e erros, permitindo caracterizar diferentes dimensões do comportamento preditivo. A seleção dessas métricas varia conforme as características do problema, a proporção entre as classes e o objetivo analítico (Hand; Mannila; Smyth, 2001).

Autores de referência indicam que a matriz de confusão é o ponto de partida para a avaliação, por organizar resultados previstos e observados e sustentar a derivação das principais métricas. Medidas como acurácia, precisão, recall e F1-score são amplamente empregadas em estudos empíricos para comparação de técnicas, sobretudo em bases heterogêneas ou desbalanceadas (Witten; Frank; Hall, 2011; Hand; Mannila; Smyth, 2001).

A literatura indica que a escolha das métricas deve considerar a natureza do problema e as propriedades estatísticas da base, pois cada métrica representa um aspecto do desempenho. Em aplicações práticas, recomenda-se o uso combinado de métricas derivadas da matriz de confusão para obter uma avaliação mais equilibrada das técnicas de classificação.

A seguir, são apresentadas as métricas adotadas para avaliar o desempenho: Matriz de Confusão, Acurácia, Precisão, Recall e F1-score.

a) Matriz de Confusão

A matriz de confusão organiza os acertos e erros do modelo, comparando os valores observados com as previsões realizadas. A partir dela, é possível identificar os Verdadeiros Positivos (VP) e Verdadeiros Negativos (VN), que correspondem às previsões corretas de ocorrência e não ocorrência do evento, respectivamente. Também são indicados os Falsos Positivos (FP) e Falsos Negativos (FN), que representam os erros de classificação (Tan; Steinbach; Kumar, 2009).

Em problemas de classificação binária, a matriz reúne quatro componentes: verdadeiros positivos (VP), falsos positivos (FP), verdadeiros negativos (VN) e falsos negativos (FN). Sua forma geral pode ser visualizada na Equação (8):

		Classe prevista	
		Positivo	Negativo
Classe real	Positivo	VP	FN
	Negativo	FP	VN

(8)

Fonte: A autora (2026).

Formalmente, essa estrutura pode ser representada pela seguinte função de contagem, conforme a Equação (9):

M(i, j) = número de instâncias cuja classe real é i e a classe prevista é j (9)

Fonte: A autora (2026).

Na qual:

i ∈ {positivo, negativo}

j ∈ {positivo, negativo}

A matriz de confusão completa é representada pela Equação (10):

$$M = [[VP, FN], [FP, VN]] \quad (10)$$

Fonte: A autora (2026).

A aplicação da matriz de confusão neste trabalho organizou as previsões geradas pelas técnicas e permitiu identificar a distribuição entre verdadeiros positivos,falsos positivos, verdadeiros negativos e falsos negativos.

Essa estrutura serviu de base para o cálculo das métricas empregadas no trabalho, como acurácia, precisão, recall, F1-score e AUC, que dependem diretamente das frequências registradas na matriz.

A representação matricial possibilitou examinar separadamente os tipos de erro produzidos pelos classificadores e forneceu um ponto de referência para todas as técnicas avaliadas (Hastie; Tibshirani; Friedman, 2009).

Além disso, a matriz de confusão integrou-se às demais etapas de avaliação realizadas no código, que incluíram a construção das curvas ROC, das curvas *precision–recall* e das curvas de calibração, bem como a comparação das predições reais e previstas.

Nesse conjunto de análises, a matriz serviu como elemento estrutural, pois sustentou a derivação das métricas numéricas e permitiu a interpretação conjunta dos gráficos gerados. Dessa forma, ela compôs o núcleo do processo de avaliação dos classificadores utilizados neste trabalho, em conformidade com as recomendações metodológicas apresentadas na literatura (Provost; Fawcett, 2013).

b) Curva ROC

A curva ROC é construída a partir da relação entre duas quantidades derivadas da matriz de confusão: a taxa de verdadeiros positivos (TPR) e a taxa de falsos positivos (FPR). Essas quantidades são definidas pela Equação (11):

$$\mathrm{TPR} = \frac{VP}{VP + FN} \qquad \mathrm{FPR} = \frac{FP}{FP + VN} \tag{11}$$

Fonte: A autora (2026).

A curva resulta do cálculo desses pares (FPR, TPR) ao longo de diferentes limiares aplicados às probabilidades estimadas por uma técnica de classificação.

A AUC-ROC corresponde à área sob essa curva e expressa, em termos formais, a probabilidade de que a técnica atribua um escore maior a uma instância positiva do que a uma instância negativa, interpretação discutida em diversos trabalhos metodológicos (Fawcett, 2006).

Na literatura de aprendizado de máquina, a *ROC* tem sido utilizada como medida para examinar o comportamento discriminativo das técnicas sem depender de um limiar específico. Hastie, Tibshirani e Friedman (2009) destacaram que a ROC permite comparar técnicas com diferentes características estruturais, pois avalia o desempenho ao longo de todos os limiares possíveis.

Bishop (2006) observou que, em classificadores probabilísticos, a ROC é compatível com a ordenação de escores produzida pelas técnicas, uma vez que considera a relação entre distribuições estimadas para as classes.

Em aplicações empíricas, a ROC aparece em estudos envolvendo domínios como diagnóstico médico, detecção de anomalias, classificação de risco e previsão de eventos raros. Provost e Fawcett (2013) discutiram a adequação da AUC em cenários com assimetria na distribuição das classes, enquanto Witten, Frank e Hall (2011) relataram seu uso em experimentos comparativos que avaliam técnicas supervisionadas sob diferentes características dos dados.

Trabalhos aplicados empregaram a ROC como métrica para avaliar a capacidade discriminativa de técnicas supervisionadas. Kloft et al. (2014) utilizaram a ROC na análise de padrões de engajamento em cursos online, comparando classificadores quanto à identificação de risco de abandono.

Costa e Holanda (2021) também adotaram a ROC para comparar Regressão Logística, *Random Forest* e métodos baseados em gradiente na previsão de evasão no ensino superior, destacando diferenças no desempenho das técnicas sob distintos conjuntos de variáveis. De forma semelhante, Lakkaraju et al. (2015) aplicaram a métrica em modelos de risco educacional, considerando cenários com forte desbalanceamento entre classes.

No presente trabalho, a ROC foi utilizada com finalidade equivalente, permitindo avaliar a separação entre as classes de evasão e não evasão ao longo de limiares de decisão. O uso da métrica seguiu a orientação metodológica dos trabalhos mencionados, possibilitando comparar técnicas com estruturas distintas e examinar seu comportamento discriminativo em um contexto educacional com distribuição assimétrica.

c) Acurácia

A acurácia é uma métrica que expressa a proporção de classificações corretas realizadas por uma técnica em relação ao total de instâncias avaliadas. Seu cálculo baseia-se nos elementos da matriz de confusão e é definido conforme a Equação (12).

$$\text{Acurácia} = \frac{VP + VN}{VP + VN + FP + FN} \quad (12)$$

Fonte: A autora (2026).

Na literatura, a acurácia apareceu em diferentes tipos de trabalhos como medida inicial para descrever o comportamento geral de técnicas supervisionadas. Witten, Frank e Hall (2011) mencionaram seu uso recorrente em avaliaçõesexperimentais por fornecer uma visão agregada do desempenho. Hastie, Tibshirani e Friedman (2009) observaram que a acurácia foi frequentemente apresentada em conjunto com métricas derivadas da matriz de confusão, sobretudo em cenários nos quais os erros do tipo I e II podiam ter impactos distintos.

Trabalhos na área educacional utilizam a acurácia como medida inicial para comparação de modelos em estudos sobre evasão, tanto em cursos presenciais quanto online. Esses estudos apresentam a acurácia como uma visão geral do desempenho das técnicas, sendo posteriormente complementada por métricas como precisão e recall (Kloft et al., 2014; Costa; Holanda, 2021).

Em outras áreas aplicadas, como diagnóstico médico e detecção de anomalias, a acurácia também é empregada para comparação inicial entre modelos, antes da análise de métricas mais específicas (Chandola; Banerjee; Kumar, 2009; Han; Kamber; Pei, 2012). De forma semelhante, em problemas de classificação de risco, essa medida é utilizada para comparar o desempenho geral de classificadores em diferentes distribuições de classes (Provost; Fawcett, 2013).

Neste trabalho, a acurácia foi utilizada para quantificar o percentual de acertos das técnicas ao longo das partições de validação, permitindo uma comparação inicial entre os métodos avaliados. Essa medida contribuiu para uma visão geral do desempenho, sendo complementada por outras métricas em um contexto de previsão de evasão com distribuição assimétrica entre as classes.

d) Precisão

A precisão representou a proporção de instâncias classificadas como positivas que, de fato, pertenciam à classe positiva. Essa métrica foi derivada dos elementos da matriz de confusão e expressa na equação (13):

$$\text{Precisão} = \frac{VP}{VP + FP} \quad (13)$$

Fonte: A autora (2026).

Na literatura, a precisão apareceu em diversos trabalhos como medida voltada a analisar especificamente os erros do tipo I, isto é, os casos em que uma instância negativa foi identificada como positiva. Witten, Frank e Hall (2011) relataram seu uso em experimentos destinados a comparar o comportamento de técnicas sob diferentes taxas de falsos positivos.

Hastie, Tibshirani e Friedman (2009) observaram que a precisão era frequentemente empregada em conjunto com o recall para examinar separadamente diferentes dimensões do desempenho, sobretudo em bases de dados com distribuição assimétrica entre as classes.

Em trabalhos aplicados, a precisão foi utilizada em áreas como diagnóstico médico e detecção de anomalias. Han, Kamber e Pei (2012) mencionaram seu uso em classificadores voltados à identificação de registros incomuns, enquanto Chandola, Banerjee e Kumar (2009) relataram a adoção da métrica para avaliar técnicas que buscavam minimizar falsos positivos em processos de detecção.

No contexto de classificação de risco, Provost e Fawcett (2013) também destacaram a precisão como métrica relevante para examinar cenários nos quais a proporção de instâncias positivas era reduzida. Trabalhos na área educacional, incluindo pesquisas sobre previsão de evasão, utilizaram a precisão para verificar a adequação das técnicas na identificação de alunos potencialmente evadidos (Kloft et al., 2014; Costa; Holanda, 2021).

Neste trabalho, a precisão foi utilizada para avaliar a proporção de previsões positivas corretas ao longo das partições de validação. Essa métrica permitiu observar a ocorrência de falsos positivos e contribuiu para a comparação entre as técnicas analisadas, sendo considerada na avaliação do desempenho em um cenário de previsão de evasão com distribuição assimétrica entre as classes.

e) F1-score

A Medida-F1 representou a média harmônica entre precisão e recall, sendo utilizada para avaliar simultaneamente o desempenho da técnica em relação aos falsos positivos e aos falsos negativos. Sua formulação é representada pela equação (14):

$$F_1 = 2\frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}} = \frac{2VP}{2VP + FP + FN} \quad (14)$$

Fonte: A autora (2026).

Na literatura, a F1-score tem sido discutida como métrica adequada para cenários em que a distribuição das classes é assimétrica ou quando existe interesse em controlar, de forma conjunta, diferentes tipos de erro. Witten, Frank e Hall (2011) indicaram que a média harmônica, ao invés da média aritmética, atribuiu maior impacto aos valores mais baixos de precisão ou recall, o que tornou a F1-score sensível a desequilíbrios entre essas quantidades.

Hastie, Tibshirani e Friedman (2009) mencionaram que a F1-score aparece de forma recorrente em trabalhos comparativos quando precisão ou recall, isoladamente, não fornecem uma descrição completa do comportamento da técnica. Esses autores destacaram que a F1-score é particularmente útil quando a taxa de falsos negativos ou falsos positivos apresenta relevância semelhante.

Em trabalhos aplicados, a F1-score tem sido amplamente utilizada em problemas caracterizados por baixa proporção da classe positiva. Chandola, Banerjee e Kumar (2009) relataram seu uso em detecção de anomalias devido à necessidade de avaliar o equilíbrio entre identificar corretamente casos raros e evitar alarmes indevidos.

Provost e Fawcett (2013) discutiram a adoção da F1-score em contextos de classificação de risco, nos quais diferentes limiares de decisão podem alterar de forma significativa a relação entre precisão e recall. Em pesquisas sobre evasão, a F1-score foi empregada por Kloft et al. (2014) e por Costa e Holanda (2021), que analisaram técnicas supervisionadas em bases com proporção reduzida de alunos evadidos e observaram que valores de acurácia, sozinhos, não capturavam adequadamente o comportamento das técnicas nessas condições.

No presente trabalho, a F1-score foi utilizada para avaliar a relação entre acertos na identificação de alunos evadidos e a capacidade de evitar classificações incorretas de alunos não evadidos na classe positiva. A métrica forneceu uma avaliação complementar às medidas derivadas diretamente da matriz de confusão, permitindo examinar o desempenho das técnicas em um contexto caracterizado por desbalanceamento entre as classes.

Dessa forma, a F1-score integrou o conjunto de métricas empregadas para comparar as técnicas analisadas, contribuindo para a análise preditiva em um cenário no qual tanto falsos positivos quanto falsos negativos apresentam relevância prática para a interpretação dos resultados.

f) Recall

O *recall* (Sensibilidade) representou a proporção de instâncias positivas corretamente identificadas pela técnica, sendo, portanto, uma medida diretamente associada aos falsos negativos. Trata-se da razão entre verdadeiros positivos e o total de instâncias que pertenciam à classe positiva, expressa pela equação (15):

$$\text{Recall} = \frac{VP}{VP + FN} \tag{15}$$

Fonte: A autora (2026).

Na literatura, o *recall* tem sido discutido como métrica adequada para avaliar a capacidade de uma técnica de recuperar instâncias raras ou pouco frequentes. Witten, Frank e Hall (2011) destacaram que o recall é especialmente relevante quando o custo de perder um caso positivo é elevado, como em problemas de detecção de risco, fraude, anomalias ou eventos raros.

Hastie, Tibshirani e Friedman (2009) observaram que, em distribuições assimétricas, o recall frequentemente complementa métricas globais como acurácia, pois fornece uma visão direta do comportamento da técnica em relação aos falsos negativos, que podem ser subestimados em avaliações gerais.

Em trabalhos aplicados, o recall foi utilizado em áreas que exigem alta sensibilidade na identificação de instâncias positivas. Chandola, Banerjee e Kumar (2009) discutiram seu uso em detecção de anomalias, dada a necessidade de recuperar casos raros mesmo quando a taxa de falsos positivos permanece controlada.

Na área de classificação de risco, Provost e Fawcett (2013) relataram a importância do recall para examinar o desempenho de técnicas em situações nas quais a classe minoritária possui maior relevância analítica. No campo educacional, estudos de previsão de evasão como os de Kloft et al. (2014) e Costa e Holanda (2021) empregararam o recall para avaliar a capacidade das técnicas de identificar corretamente alunos em risco, dada a baixa proporção de evadidos em muitas bases institucionais.

No presente trabalho, o recall foi utilizado para mensurar a capacidade das técnicas de identificar alunos da classe positiva (evadidos), indicando o quanto cada técnica recuperou dos casos que efetivamente ocorreram. A métrica complementou a análise fornecida pela precisão ao mostrar a relação entre verdadeiros positivos e falsos negativos, permitindo examinar como as técnicas se comportaram em um cenário caracterizado por distribuição assimétrica entre as classes.

Dessa forma, o *Recall* integrou o conjunto de indicadores empregados para comparar o desempenho das técnicas estudadas no contexto da previsão de evasão escolar.

g) Curva de Calibração

A Curva de Calibração foi utilizada para avaliar a correspondência entre as probabilidades estimadas pelas técnicas e as frequências reais de evasão ao longo das partições de validação. Na literatura, a calibração probabilística é definida como a condição em que as estimativas do modelo refletem a proporção real do evento, de modo que, entre as instâncias com probabilidade p, aproximadamente p x 100% pertençam à classe positiva (Niculescu-Mizil; Caruana, 2005). Matematicamente, essa relação é expressa conforme equação (16):

$$\text{Calibração}(p) = P\left(Y = 1 \mid \widehat{P} = p\right) \tag{16}$$

Fonte: A autora (2026).

Em que $\hat{P}$ representa a probabilidade prevista de evasão.

Diversos domínios científicos recorreram à calibração probabilística para avaliar a consistência das estimativas produzidas por modelos supervisionados. A literatura indicou que modelos mal calibrados tenderam a superestimar ou subestimar eventos raros, o que comprometeu sua utilidade prática em processos de tomada de decisão(Caruana; Niculescu-Mizil, 2004; Musso et al., 2020).

Em pesquisas voltadas à análise de permanência de alunos, a calibração foi utilizada para verificar a confiabilidade das probabilidades geradas, sobretudo quando se buscou identificar níveis distintos de risco entre alunos com características semelhantes (Hoyos; Caicedo-Castro, 2024).

No contexto deste trabalho, a calibração apresentou-se como uma etapa adequada para verificar a consistência das probabilidades geradas pelas técnicas aplicadas à previsão de evasão, considerando que níveis distintos de risco atribuídos a alunos podem subsidiar análises e decisões em ambientes educacionais orientados por dados.

Pesquisas em Mineração de Dados Educacionais exploraram a utilidade de probabilidades calibradas para representar níveis diferenciados de risco acadêmico, especialmente em modelos aplicados à identificação de alunos propensos ao abandono (Baker; Inventado, 2014; Aguiar et al., 2015; Kloft et al., 2014; Xing et al., 2016).

3 MATERIAIS E MÉTODOS

Neste capítulo, apresentam-se a revisão da literatura, a caracterização metodológica, a base de dados, a plataforma de ensaios e as fases dos experimentos computacionais.

3.1 REVISÃO DA LITERATURA

A realização da Revisão da Literatura contemplou publicações relacionadas aos seguintes temas centrais: Perfil do Aluno, Trajetória Acadêmica, ETEC, Mineração de Dados Educacionais, Gestão Escolar e Evasão. Estes termos foram utilizados como descritores também na língua inglesa:

Student Profile, Academic Trajectory, Educational Data Mining, School Dropout, Technical Schools e School Management.

A revisão foi realizada considerando as seguintes bases de dados: SCiELO, ERIC, Web of Science, IEEE Xplore e Scopus, com recorte temporal de 2010 a maio de 2026. O longo período considerado se justifica pela escassez de trabalhos relacionados ao tema deste trabalho.

Os critérios de seleção das obras foram a aderência aos temas centrais, a clareza metodológica, a relação direta com análises de trajetória escolar, perfis acadêmicos e aplicação de MDE em contextos de ensino técnico.

As publicações duplicadas ou que não apresentaram relação com os temas centrais foram descartadas, após leitura dos resumos e avaliação preliminar do conteúdo, considerando o foco exclusivo no ensino superior, a abordagem teórica sem aplicação prática, sem relação direta com o uso de dados educacionais para gestão e trabalhos duplicados entre as bases pesquisadas.

O processo de triagem identificou 85 trabalhos. Após a aplicação dos critérios de exclusão, 18 publicações permaneceram diretamente relacionadas ao tema, conforme o Quadro 1.

Quadro 1 — Trabalhos diretamente relacionados ao tema

Autor(es) / Ano	Tema central do trabalho	Relação com este trabalho
Romero e Ventura (2010)	Mineração de Dados Educacionais	Aplicação da MDE na resolução de problemas educacionais.
Baker e Inventado (2014)	Learning Analytics	Learning Analytics e MDE para apoiar o desenvolvimento de pesquisa e aplicação prática na área da educação.
Andrelo (2022)	Perfil do egresso	Análise do perfil dos egressos da ETEC Paulistano com a aplicação da MDE.
Silva e Pereira (2021)	Previsão de evasão	Discute as variáveis associadas ao risco de evasão e a necessidade da previsão do fenômeno.
Alshanqiti e Namoun (2020)	Classificação preditiva	Apresenta técnicas de classificação.
Waheed et al. (2020)	Predição da performance acadêmica	Importância de aplicar técnicas inteligentes na predição da trajetória acadêmica.
Musso et al. (2020)	Trajetória Acadêmica	Técnicas de Aprendizado de Máquinas aplicadas na Previsão da trajetória acadêmica.
Carleto (2011)	Indicadores educacionais	Apoiar a leitura institucional da trajetória acadêmica.
Chen (2020)	Sistemas educacionais digitais	Contextualizar o uso de dados acadêmicos na gestão escolar.
Costa et al. (2012)	Classificação e clusterização	Fundamentar técnicas de classificação preditivas.
Hoyos e Caicedo-Castro (2024)	Desempenho Acadêmico	Mineração de dados para prever o desempenho acadêmico no ensino médio.
Kesavaraj e Sukumaran (2013)	Revisão de classificadores	Justificar a seleção de técnicas de classificação preditiva.
Melo et al. (2022)	Gestão escolar e dados	Relacionar dados da trajetória acadêmica com tomada de decisão.
Moura (2023)	Perfil de aluno ingressante em curso superior	MDE para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores.
Oyewole e Thopil (2023)	Análise de perfis	Interpretar o perfil de alunos.
Pontes et al. (2022)	Educação profissional	Situar o perfil do aluno da ETEC no contexto amplo da educação profissional.
Psyridou et al. (2024)	Indicadores preditivos	Selecionar múltiplos indicadores em conjunto de dados.
Scherer e Brito (2020)	Sistemas de informação escolar	Fundamentar do NSA.

Fonte: A autora (2026)

A seleção dos trabalhos apresentados no Quadro 1 ocorreu a partir da análise da literatura identificada durante a pesquisa bibliográfica. O critério de escolha foi a aderência aos eixos teóricos que sustentaram o presente trabalho Perfil do Aluno, Trajetória Acadêmica, ETEC, Mineração de Dados Educacionais, Gestão Escolar e Evasão.

Essa diretriz de seleção foi compatível com orientações metodológicas presentes em trabalhos de referência, como os de Romero e Ventura (2010) e Baker e Inventado (2014), que destacaram a importância de alinhar o corpo teórico às variáveis e aos fenômenos analisados.

A literatura revisada foi examinada quanto ao alinhamento conceitual, ao tipo de abordagem metodológica e à proximidade com o problema investigado, em consonância com Witten, Frank e Hall (2011).

Foram priorizados trabalhos que discutiram técnicas de Mineração de Dados Educacionais, conforme apresentados por Costa et al. (2012), Han, Kamber e Pei (2012) e Waheed et al. (2020). Também foram incluídas publicações que trataram de previsão de evasão e fatores associados à permanência, como Alshanqiti e Namoun (2020), Silva e Pereira (2021) e Hoyos e Caicedo-Castro (2024).

Além disso, trabalhos que abordaram a trajetória escolar e perfis acadêmicos, como Musso et al. (2020) e Psyridou et al. (2024), contribuíram para a definição das categorias interpretativas utilizadas neste estudo.

Trabalhos relacionados ao contexto da educação profissional e às ETECs, como Andrelo (2022), Pontes et al. (2022) e Moura (2023), foram incluídos por sua proximidade temática com o cenário investigado. Por fim, trabalhos sobre sistemas de informação educacional e gestão escolar, como Scherer e Brito (2020) e Carleto (2011), auxiliaram na análise do uso institucional de dados acadêmicos. Dessa forma, os trabalhos selecionados constituíram o núcleo da fundamentação que sustentou o trabalho desenvolvido.

3.2 CARACTERIZAÇÃO METODOLÓGICA

A natureza desta pesquisa é aplicada, pois busca compreender um problema real relacionado à evasão em uma ETEC. A abordagem adotada é quantitativa, uma vez que utiliza dados provenientes do Novo Sistema Acadêmico (NSA). Quanto aos objetivos, caracteriza-se como descritiva, ao analisar padrões e perfis a partir das informações registradas. Em relação aos procedimentos, envolve pesquisa bibliográfica, para fundamentação teórica, pesquisa documental, com base nos dados do sistema acadêmico, e estudo de caso, ao concentrar a análise na ETEC Raposo Tavares (Prodanov e Freitas, 2013).

Do ponto de vista técnico, o procedimento é experimental, pois utiliza técnicas computacionais para manipular variáveis e observar seus efeitos em um ambiente controlado. Nesse tipo de procedimento, diferentes condições são aplicadas aos dados com o objetivo de analisar o comportamento das variáveis e identificar relações entre elas (Gil, 2008).

3.3 BASE DE DADOS E PLATAFORMA DE ENSAIOS

O hardware utilizado nos experimentos computacionais foi um dispositivo com processador AMD Ryzen 7 5700U com gráficos Radeon, operando a 1,80 GHz. O sistema possui 16,00 GB de memória RAM e o sistema operacional é o Windows 11 de 64 bits. Os *softwares* e bibliotecas utilizados são apresentados na Tabela 1.

Tabela 1 — Softwares e bibliotecas utilizados

Software / biblioteca	Descrição	Utilização no trabalho	Endereço eletrônico
Python	Linguagem de programação.	Desenvolvimento dos algoritmos e processamento dos dados.	python.org
Jupyter Notebook	Ambiente de desenvolvimento interativo.	Execução e organização dos experimentos.	jupyter.org
Pandas	Biblioteca para manipulação de dados.	Leitura, tratamento e organização da base.	pandas.pydata.org
NumPy	Biblioteca para computação numérica.	Operações matemáticas e manipulação de arrays.	numpy.org
Scikit-learn	Biblioteca de aprendizado de máquina.	Classificação, pré-processamento, validação e métricas.	scikit-learn.org
XGBoost	Biblioteca de gradient boosting.	Aplicação da técnica XGBoost.	xgboost.ai

Software / biblioteca	Descrição	Utilização no trabalho	Endereço eletrônico
Matplotlib	Biblioteca de visualização.	Geração dos gráficos dos resultados.	matplotlib.org
Seaborn	Biblioteca de visualização estatística.	Apoio à apresentação gráfica das matrizes.	seaborn.pydata.org
warnings	Módulo nativo do Python.	Controle de mensagens durante a execução do código.	docs.python.org
hashlib	Módulo nativo do Python.	Anonimização dos identificadores por hash SHA-256.	docs.python.org
re	Módulo de expressões regulares.	Identificação, validação e substituição de padrões textuais.	docs.python.org
Canva	Assistente de design gráfico.	Preparação visual de figuras esquemáticas.	canva.com

Fonte: A autora (2026).

Foi enviado um memorando para a diretoria da ETEC Raposo Tavares solicitando autorização para acesso e disponibilidade dos dados do NSA da escola. O documento pode ser visualizado no Apêndice A deste trabalho. O parecer técnico autorizando o uso dos dados do NSA da ETEC Raposo Tavares pode ser visualizado no Apêndice B deste trabalho. O parecer destaca que por não se tratar de pesquisa com seres humanos não houve necessidade de cadastro na Plataforma Brasil, nem submissão ao Comitê de Ética em Pesquisa.

A base contém informações dos anos de 2023 e 2024, referentes a 185 alunos matriculados no Curso Técnico em Química da ETEC Raposo Tavares, do primeiro ao terceiro módulo. Após as fases de pré-processamento e engenharia de atributos, a base passou a conter 298 registros e 16 colunas sendo 15 atributos preditores e 1 variável-alvo (evadiu) , considerando que um mesmo aluno poderia gerar mais de um registro ao longo dos módulos.

A utilização de técnicas de aprendizado de máquina em uma base com 298 registros é adequada quando acompanhada de procedimentos que garantem a generalização dos modelos. Apesar do volume reduzido, a base apresenta múltiplas variáveis relevantes da trajetória acadêmica, permitindo a identificação de padrões consistentes.

Para evitar sobreajuste (*overfitting*), foram adotadas estratégias como validação cruzada, ajuste controlado de hiperparâmetros e limitação da complexidade dos modelos.Além disso, a comparação entre algoritmos com características distintas e o uso de métricas como F1-score, recall e ROC permitiram avaliar o desempenho de forma mais robusta.

Essas medidas asseguram que os resultados não estejam restritos à amostra analisada, mas reflitam padrões com potencial de generalização (Hastie, Tibshirani e Friedman, 2009).

A unidade de análise adotada neste trabalho correspondeu aos registros acadêmicos dos alunos ao longo dos módulos do curso, com base na trajetória acadêmica. Dessa forma, um mesmo aluno pôde ser representado por mais de um registro, de acordo com sua evolução entre módulos. Essa estratégia permitiu a análise de variações de desempenho, frequência e progressão ao longo do tempo, elementos relevantes para a identificação de padrões associados ao risco de evasão. Devido à presença de múltiplos registros por aluno, adotaram-se procedimentos para evitar dependência indevida entre as observações durante o processo de modelagem e avaliação.

Especificamente, utilizou-se validação cruzada com agrupamento por aluno, na qual todos os registros pertencentes a um mesmo aluno permaneceram no mesmo subconjunto de dados, evitando a introdução de viés decorrente do compartilhamento de informações entre os conjuntos de treinamento e teste, conforme recomenda a literatura. Adicionalmente, o atributo identificador do aluno foi excluído do conjunto de variáveis, de modo a evitar a memorização de padrões individuais pelos modelos (Romero e Ventura, 2010; Kuhn e Johnson, 2013).

O uso dos dados educacionais respeitou princípios éticos relacionados à proteção das informações dos alunos e à aplicação responsável dos resultados. Os dados foram anonimizados em conformidade com a Lei Geral de Proteção de Dados (LGPD), o que preservou a identidade dos alunos. Os resultados constituem apoio à gestão escolar e complementam as análises pedagógicas e o acompanhamento individual.

Como a base de dados estava originalmente em formato PDF, os dados foram convertidos em tabelas por meio de técnicas implementadas em Python, envolvendo a extração de informações em documentos semiestruturados.

Em seguida, foram aplicadas expressões regulares (*regex*) para tratar inconsistências na base, como quebras de cabeçalho, colunas mescladas e informações distribuídas em múltiplas linhas. Esse processo permitiu organizar a base em uma estrutura tabular padronizada, adequada para análise. Para a realização dessas etapas, foram considerados procedimentos descritos na literatura sobre pré-processamento e transformação de dados no contexto do processo de KDD (Fayyad et al., 1996).

Após a extração, os dados foram padronizados e tratados com o uso das bibliotecas Pandas e NumPy. Esse processo envolveu a correção de formatos numéricos, a remoção de registros inconsistentes e a uniformização dos nomes das variáveis, com o objetivo de organizar a base para análise. As escalas percentuais foram harmonizadas para o intervalo 0–100, o que permitiu comparações entre diferentes meses e módulos.

Os registros mensais foram reorganizados por meio da pivotagem *long wide,* realizada com a biblioteca Pandas, convertendo múltiplos registros de um mesmo aluno em uma única linha por módulo. Essa transformação reduziu redundâncias e preparou a base.

Em seguida, foram construídas novas variáveis por meio de engenharia de atributos (f*eature engineering*), utilizando *Python*. A partir das informações disponíveis, foram calculadas medidas como média, valores mínimos e máximos de frequência, bem como criadas variáveis relacionadas à evolução da frequência ao longo do módulo. Essas etapas foram realizadas com base em procedimentos descritos na literatura de mineração de dados educacionais, que envolvem a transformação e a construção de atributos a partir de registros acadêmicos (Romero e Ventura, 2010).

Para proteção dos dados, o identificador do aluno foi substituído por um código anônimo gerado por meio de algoritmo de hash criptográfico SHA-256, utilizando a biblioteca *hashlib*. Esse procedimento foi adotado para evitar a identificação direta dos alunos e preservar a confidencialidade das informações utilizadas no trabalho.

Todo o processamento, organização e transformação dos dados ocorreu em Python, que serviu como plataforma única para preparar a base final utilizada nos ensaios preditivos sobre risco de evasão.

A relação completa dos atributos encontra-se na Tabela 2, inicialmente validada pela coordenação do curso, coordenação pedagógica e direção da unidade durante reunião de planejamento realizada em 22/07/2025. Posteriormente, os

atributos foram revisados e ajustados conforme orientações da banca, sendo realizada nova validação em 10/11/2025.

Tabela 2 — Atributos brutos extraídos do NSA

Atributo	Tipo	Descrição	Justificativa metodológica
Id_aluno	Numérico	Código anonimizado do aluno	Identificador anônimo do aluno utilizado para diferenciar registros ao longo dos módulos.
disciplina_cursada	Texto	Nome da disciplina cursada em cada módulo.	Permite relacionar desempenho por área e identificar padrões de progressão.
nota_final_modulo	Numérico	Nota institucional de 0 a 100.	Representa a nota final do aluno ao término do módulo.
menção	Ordinal	MB, B, R, I e AE, convertidas para a escala de 4 a 0.	Representa a avaliação qualitativa institucional em formato numérico.
frequencia_modulo	Numérico	Percentual de presença do aluno no módulo.	Indica a assiduidade ao longo do período.
faltas_modulo	Numérico	Total de ausências registradas.	Registra a quantidade de faltas durante o módulo.

Fonte: A autora (2026).

A Tabela 2 apresenta os atributos brutos extraídos dos relatórios do NSA. A partir desses registros, as fases de pré-processamento e transformação geraram dez atributos derivados utilizados efetivamente na modelagem preditiva, descritos no Capítulo 4.

Neste trabalho, a trajetória acadêmica refere-se ao conjunto de registros produzidos ao longo dos módulos do curso, envolvendo informações de desempenho, frequência e progressão escolar. Esses registros compõem o histórico do aluno e permitem observar sua trajetória formativa ao longo do tempo.

A variável-alvo evadiu foi definida operacionalmente a partir dos registros de situação de matrícula disponíveis no NSA. Foi classificado como evadido (evadiu = 1) o aluno que apresentou encerramento de matrícula por abandono, desistência formal ou ausência prolongada sem registro de transferência para outra unidade ou curso da rede Centro Paula Souza, até o encerramento do módulo em que estava matriculado. Foram classificados como não evadidos (não evadiu = 0) os alunos que concluíram o módulo, que obtiveram aprovação ou reprovação com continuidade de matrícula, ou que realizaram transferência interna para outro curso ou unidade da rede.

Essa distinção é importante do ponto de vista metodológico, pois evita a inclusão de situações administrativas distintas como transferências regulares ou afastamentos temporários na mesma categoria do abandono definitivo.

A separação entre evasão e outras formas de saída do curso segue orientações presentes na literatura de MDE aplicada à educação profissional, que recomendam aadoção de critérios institucionais precisos na construção da variável-alvo para evitar ruído nas técnicas aplicadas (Romero e Ventura, 2010; Silva e Pereira, 2021).

A definição foi validada junto à coordenação do curso e à direção da unidade durante a reunião de planejamento realizada em 22/07/2025, conforme descrito na seção de materiais e métodos, assegurando a aderência do critério adotado à realidade institucional da ETEC Raposo Tavares.

O atributo identificador do aluno (Id_aluno) foi utilizado exclusivamente para fins de organização e rastreabilidade dos registros durante o pré-processamento dos dados. Esse atributo não foi incluído no conjunto de variáveis, uma vez que não possui significado explicativo para o fenômeno analisado e poderia introduzir viés no modelo, permitindo a memorização de padrões em vez da generalização, conforme discutido na literatura de modelagem preditiva (Kuhn e Johnson, 2013).

A Tabela 3 apresenta todos os atributos da base finalizada, com indicação de tipo, origem e descrição de cada variável.

Tabela 3 — Atributos da base de dados finalizada

Atributo	Tipo	Origem	Descrição
id_anon	Texto (hash)	Bruto anonimizado	Identificador do aluno anonimizado via SHA-256 (hashlib)
disciplina_cursada	Texto	Bruto	Nome da disciplina cursada em cada módulo
nota_final_modulo	Numérico	Bruto	Nota institucional de 0 a 100
mencao	Ordinal	Bruto convertido	Menção qualitativa convertida para escala numérica: MB=4, B=3, R=2, I=1, AE=0
frequencia_modulo	Numérico	Bruto	Percentual de presença do aluno na disciplina
faltas_modulo	Numérico	Bruto	Total de ausências registradas durante o módulo
media_frequencia	Numérico (derivado)	Derivado	Média do percentual de frequência ao longo do módulo
total_faltas	Numérico (derivado)	Derivado	Soma total de faltas acumuladas no período
variacao_frequencia	Numérico (derivado)	Derivado	Diferença entre o percentual de frequência no início e no fim do módulo
min_frequencia	Numérico (derivado)	Derivado	Menor percentual de frequência registrado no módulo
max_frequencia	Numérico (derivado)	Derivado	Maior percentual de frequência registrado no módulo
media_desempenho	Numérico (derivado)	Derivado	Média das notas dos componentes curriculares do módulo
disciplinas_criticas	Numérico (derivado)	Derivado	Contagem de disciplinas com frequência inferior a 75%
faltas_mensais	Numérico (derivado)	Derivado	Registro de faltas por mês (desdobrado em colunas mensais)
modulo	Numérico (derivado)	Derivado	Número do módulo cursado pelo aluno (1, 2 ou 3)
evadiu	Binário (alvo)	Derivado	Variável-alvo: 1 = aluno evadido; 0 = aluno não evadido

Fonte: A autora (2026).

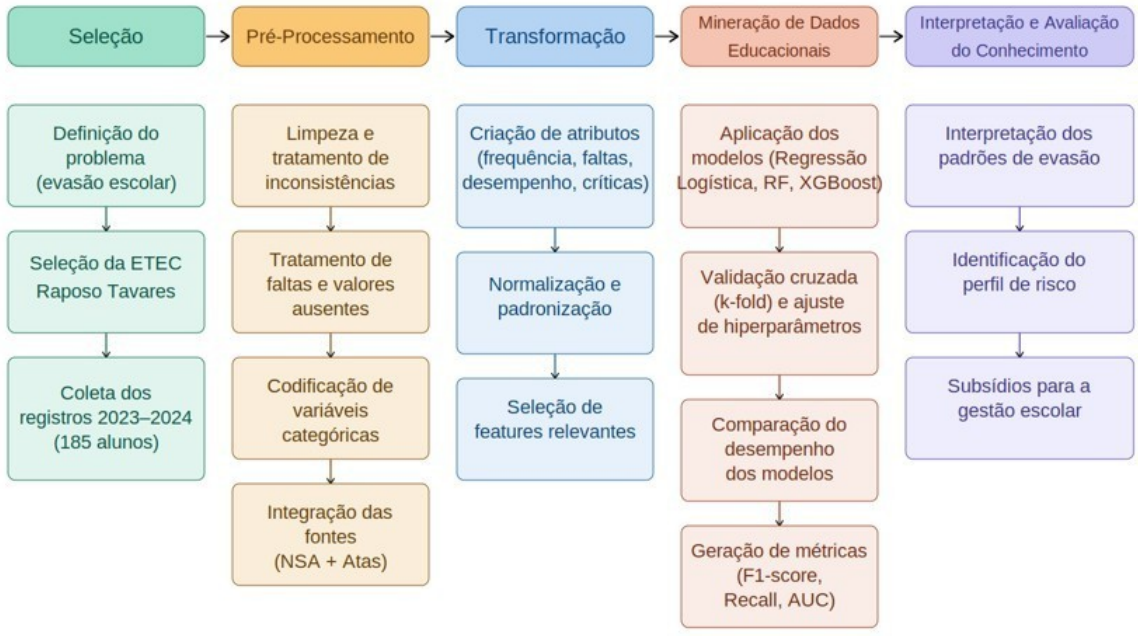
3.4 FASES DE DESENVOLVIMENTO DOS EXPERIMENTOS COMPUTACIONAIS

O desenvolvimento dos experimentos foi organizado em cinco fases, que orientaram este trabalho: (A) seleção dos dados; (B) pré-processamento; (C) transformação dos dados; (D) aplicação das técnicas de Mineração de Dados; e (E) interpretação e avaliação do conhecimento.

Essa divisão seguiu princípios apresentados em modelos de descoberta de conhecimento em bases de dados, nos quais as etapas de preparação, modelagem e análise desempenham papéis complementares no processo (Fayyad; Piatetsky-Shapiro; Smyth, 1996; Han; Kamber; Pei, 2012).

Essas fases definiram o fluxo de trabalho adotado e permitiram delinear as etapas envolvidas no procedimento experimental.A Figura 1 apresenta a sequência das etapas que compuseram o processo desenvolvimento dos experimentos.

Figura 1 – Fases do Processo de KDD



Fonte: A autora (2026).

Fase A: Seleção Dos Dados

Na fase de seleção, identificou-se quais dados disponíveis nos relatórios do NSA poderiam ser convertidas em insumos adequados para os procedimentos computacionais adotados no trabalho.

Essa fase concentrou-se na identificação das variáveis e no processamento com técnicas automatizadas. A literatura de KDD mostra que a seleção deve priorizar dados que apresentem consistência e disponibilidade suficiente para suportar análises posteriores, o que orientou as escolhas realizadas (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

Nos processos foram priorizados valores numéricos, como percentuais, faltas e desempenhos, pois esses campos podem ser manipulados de forma direta na modelagem e constituem variáveis comumente associadas a análises de aproveitamento e participação em estudos educacionais.

Informações textuais extensas ou campos administrativos sem utilidade analítica foram deixados de lado, seguindo recomendações de Han, Kamber e Pei (2012) para redução do volume de dados não informativos (Romero; Ventura, 2020).

A seleção também considerou a regularidade com que cada variável aparecia nos registros institucionais. Variáveis inconsistentes, esparsas ou presentes apenas em parte dos relatórios foram descartadas, conforme a orientação de Han, Kamber e Pei (2012), segundo a qual atributos com baixa completude contribuem pouco para tarefas analíticas.

Da mesma forma, informações de natureza administrativa, como campos internos de controle e identificadores operacionais, foram excluídas por não contribuírem para a modelagem proposta.

Por essa razão, permaneceram apenas os dados que apresentavam regularidade ao longo dos meses, compatibilidade com análises comparativas e potencial para subsidiar a análise da trajetória acadêmica dos alunos. Os relatórios mensais em PDF foram avaliados inicialmente para verificar quais campos possuíam estrutura compatível com extração programada.

Fase B: Pré-processamento dos Dados

Nesta fase, os dados extraídos foram organizados para uso na modelagem. As etapas foram realizadas em *Python*, com uso das bibliotecas pandas, *numpy* e re (Rahm; Do, 2000).

Os registros apresentavam variações de formatação entre páginas e períodos mensais. Foram realizadas operações de limpeza textual, normalização de caracteres e reorganização de campos, incluindo a reconstrução de colunas fragmentadas e a correção de inconsistências.

Em seguida, os valores textuais foram convertidos para formato numérico. Percentuais e faltas acumuladas foram transformados em valores contínuos, com ajuste de separadores decimais, permitindo o cálculo de médias e variações ao longo do tempo (Romero; Ventura, 2020).

Também foram verificados registros duplicados e tratadas observações incompletas por meio de filtragem condicional. Os dados, originalmente organizados em registros mensais, foram reorganizados por módulo, com consolidação das informações de cada aluno ao longo do tempo.

O tratamento de valores ausentes foi feito com a biblioteca *SimpleImputer*, do scikit-learn, que substitui os valores faltantes pela mediana de cada variável numérica. Esse processo ocorre em todos os modelos dentro de cada fold da validação cruzada, com ajuste apenas nos dados de treino e aplicação nos dados de validação. A padronização das variáveis foi desenvolvida com a biblioteca *StandardScaler, também do scikit-learn, e se aplica apenas à* Regressão Logística*, pois Random Forest e XGBoost não dependem da escala dos dados.*

Esses procedimentos integram a etapa de pré-processamento dos dados. Após essas etapas, a base passou de 337 registros mensais para 298 registros, correspondentes a 185 alunos distintos. Foi realizada a verificação de valores fora dos intervalos esperados, com ajuste dos registros inconsistentes. Os nomes das variáveis foram padronizados. Ao final, os dados foram organizados em uma única base para uso nas etapas de transformação e modelagem.

Fase C: Transformação dos Dados

Após o pré-processamento, os dados foram transformados para permitir a organização das informações por módulo, conforme o processo KDD para preparação de bases destinadas à modelagem (Fayyad; Piatetsky-Shapiro; Smyth, 1996).

A transformação envolveu operações em *Python* com a biblioteca pandas, utilizando métodos como *groupby*(), para agrupar os dados por aluno e módulo; agg(), para calcular medidas como média, mínimo e máximo; e *pivot_table*(), para reorganizar os dados em formato tabular. Também foram realizados cálculos estatísticos e manipulações lineares para estruturar a base final.

A estrutura inicial da base, em formato longo, permitiu a realização de cálculos ao longo do tempo para cada aluno. Essas operações utilizaram funções como *mean*(), para cálculo da média; min() e max(), para identificação dos valores mínimo e máximo; e operações de diferença, para comparar variações entre os períodos de abril e dezembro.

A literatura de MDE mostra que indicadores derivados desse tipo capturam oscilações de participação e são úteis para técnicas preditivas (Romero; Ventura, 2020).

Também foram criadas variáveis agregadas relacionadas às faltas, como faltas acumuladas no módulo e faltas mensais, por meio de operações de agregação. Para o desempenho, os percentuais dos componentes curriculares foram reorganizados e sintetizados em medidas como a média de desempenho e a contagem de componentes com frequência inferior a 75%, valor definido com base no critério institucional mínimo de frequência.

Essa contagem foi realizada por meio de condições lógicas, com o objetivo de identificar ocorrências de baixo desempenho ao longo dos módulos.

Fase D: Mineração de Dados Educacionais

Nesta fase foram aplicadas as técnicas de MDE selecionadas: Regressão Logística, *Random Forest e XGBoost.* A utilização de diferentes técnicas permitiu comparar abordagens distintas de modelagem do fenômeno da evasão, em consonância com estudos de MDE (Romero; Ventura, 2020).

Considerando o desbalanceamento entre as classes de evasão e permanência, foi utilizado o parâmetro *class_weight='balanced'* nas técnicas de Regressão Logística e *Random Forest,* de modo a ajustar os pesos das classes durante o treinamento e mitigar o impacto da classe minoritária

A base de dados preparada nas fases anteriores foi avaliada por meio de validação cruzada estratificada por grupo (*StratifiedGroupKFold***)**, com cinco partições. Nesse procedimento, os dados foram organizados em subconjuntos de treinamento e validação de forma rotativa, mantendo a proporção entre as classes da variável-alvo em cada partição.

A validação cruzada (K-fold) é adotada como alternativa que permite bom aproveitamento da amostra e reduz a variabilidade das estimativas, principalmente em bases em que não há uma quantidade suficiente de dados para reservar amostras exclusivas para avaliação, uma vez que a avaliação é realizada em diferentes partições ao longo do processo.

Já a técnica *Holdout* (divisão da base de dados em conjuntos de treino e teste) consiste na separação de uma parte dos dados para avaliação do modelo, sendo utilizada quando há quantidade suficiente de dados para reservar amostras exclusivas para teste.

A separação por grupos impediu que registros de um mesmo aluno aparecessem simultaneamente nos conjuntos de treinamento e validação, evitando vazamento de informação entre as partições, o que é especialmente importante em bases com múltiplos registros por aluno.

Para a avaliação dos modelos de classificação, adotou-se validação cruzada do tipo K-fold, com K = 5. A divisão dos dados ocorreu de forma estratificada em relação à variável alvo, com preservação da proporção entre as classes de evasão e não evasão, conforme práticas recomendadas na literatura de modelagem preditiva (Kuhn e Johnson, 2013).

Considerando que um mesmo aluno pôde apresentar múltiplos registros ao longo dos módulos, a validação cruzada foi conduzida com agrupamento por aluno, de modo que todos os registros pertencentes a um mesmo indivíduo permaneceram no mesmo subconjunto.

Esse procedimento evitou o compartilhamento de informações entre os conjuntos de treinamento e validação, reduzindo o risco de viés e permitindo uma avaliação conforme a capacidade de generalização dos modelos (Kuhn e Johnson, 2013).

A Regressão Logística foi treinada com ajuste de pesos entre as classes (*class_weight="balanced"*), penalização L2 e máximo de 2.000 iterações. *Random Forest* também utilizou ajuste de pesos, com número de estimadores, profundidade e demais parâmetros definidos por meio de busca em grade. O *XGBoost* foi configurado com ajuste de hiperparâmetros, incluindo número de estimadores, profundidade máxima e taxa de aprendizado.

Os modelos foram ajustados por meio da variação controlada de hiperparâmetros, conforme apresentado na Tabela 6 do capítulo 4. Para cada técnica, definiu-se um conjunto de valores candidatos para parâmetros, como número de estimadores, profundidade das árvores e taxa de aprendizado, no caso dos modelos baseados em árvores, e o parâmetro de regularização no caso da regressão logística.

A definição dos intervalos de busca considerou recomendações da literatura e testes experimentais preliminares, com o objetivo de equilibrar desempenho preditivo e capacidade de generalização, evitando o ajuste excessivo aos dados de treinamento, conforme discutido em Trevor Hastie et al. (2009).

O limiar de decisão foi tratado como um parâmetro ajustável para todos os modelos. Inicialmente, foi considerado o valor padrão de 0,50, no qual probabilidades iguais ou superiores a esse valor foram classificadas como evasão.

Para cada modelo, o limiar foi otimizado com base nos dados de validação de cada partição da validação cruzada, por meio da maximização do F1-score. O valor final adotado corresponde à média dos limiares obtidos nos cinco *folds*, resultando em 0,578 para a Regressão Logística, 0,392 para o *Random Forest* e 0,246 para o *XGBoost*.

A avaliação dos modelos foi realizada por meio das métricas de Acurácia, Precision, Recall, F1-score e ROC, complementadas por análise visual a partir da Matriz De Confusão, Curva ROC, Curva Precision-Recall e Curva de Calibração.

Recall e F1-*score* foram utilizados para avaliar o desempenho na classificação binária, enquanto a ROC foi empregada como medida global de desempenho. Além das métricas, o comportamento dos modelos foi analisado em relação ao impacto do limiar de decisão, considerando sua influência na identificação de alunos em risco de evasão.

A seleção do modelo mais adequado considerou o conjunto das métricas obtidas, com ênfase no equilíbrio entre Precisão e Recall, nos valores de F1-score e ROC, bem como na consistência dos resultados ao longo das partições da validação cruzada.

A análise foi conduzida com o objetivo de interpretar o comportamento das classificações em relação à variável-alvo, considerando o equilíbrio entre Precisão e Recall e a identificação dos casos positivos. Também foi considerado o efeito do ajuste do limiar de decisão nos resultados obtidos.

Foi realizada a análise da importância das variáveis, com foco na identificação dos atributos mais relevantes no contexto analisado. Os resultados foram organizados em tabelas e gráficos, possibilitando a comparação entre as técnicas conforme os critérios definidos na etapa anterior.

Fase E: Interpretação e Avaliação do Conhecimento

A fase de Interpretação e Avaliação do Conhecimento é também conhecida como Pós-processamento. Após a fase de aplicação da MDE encontrar padrões nos dados, o pós-processamento tem como objetivo tornar esses padrões úteis, compreensíveis e aplicáveis ao mundo real. Uma maneira de obter a compreensão e interpretação dos resultados é utilizar técnicas de visualização.

Esta fase foi conduzida, considerando o contexto educacional do trabalho, sendo prevista a realização de reunião com a equipe de gestão da ETEC Raposo Tavares, composta pela diretora, pela coordenação pedagógica e pela coordenação de curso.

Com base nos resultados apresentados, busca-se que a equipe de gestão interprete e avalie o conhecimento descoberto, juntamente com a autora deste trabalho, e o aplique no apoio a gestão da ETEC Raposo Tavares no desenvolvimento de estratégias educacionais, à luz da prática institucional e na expertise dos profissionais selecionados.

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Este capítulo apresenta a realização dos experimentos e a discussão dos resultados obtidos, descrevendo o desempenho dos algoritmos por meio dos resultados das métricas de avaliação aplicadas.

4.1 REALIZAÇÃO DOS EXPERIMENTOS

Realizada a Coleta dos Dados descrita na Fase 1, partiu-se para a realização das Fases 2 e 3, respectivamente Pré-Processamento e Transformação da Base de Dados. Após a organização dos registros por módulo, a base utilizada nos experimentos foi estruturada a partir das etapas de pré-processamento e transformação descritas no Capítulo 3.

As variáveis consideradas são numéricas e representam aspectos como engajamento, desempenho, faltas e participação ao longo da trajetória acadêmica, segundo Romero e Ventura (2020).

Os registros mensais apresentavam variações de formatação entre páginas e períodos. Foram realizadas operações de limpeza textual, normalização de caracteres e reorganização de campos, incluindo a reconstrução de colunas fragmentadas e a correção de inconsistências estruturais nos arquivos. Os valores textuais foram convertidos para formato numérico, com ajuste de separadores decimais, e os percentuais e faltas acumuladas foram transformados em valores contínuos, permitindo o cálculo de médias e variações ao longo do tempo.

O tratamento de valores ausentes foi realizado com a classe SimpleImputer da biblioteca scikit-learn, configurada com strategy=median, que substituiu os valores faltantes pela mediana de cada variável numérica. Conforme implementado no código, o ajuste do imputador foi realizado exclusivamente nos dados de treinamento e aplicado nos dados de validação dentro de cada fold, evitando que informações do conjunto de validação influenciassem o pré-processamento.

As variáveis foram convertidas para formatos contínuos ou binários, conforme sua natureza, para manter a compatibilidade com os algoritmos de classificação supervisionada, de acordo com Han, Kamber e Pei (2012). Após a limpeza e a validação, não foram identificados registros incompatíveis com os limites esperados. Ao final das fases de pré-processamento e transformação, a base passou a conter 298 registros e 16 atributos.

Após a extração e a reorganização dos dados por módulo, os percentuais de frequência foram mantidos na escala de 0 a 100, o que assegurou a comparação entre meses e módulos. As menções institucionais foram convertidas para uma escala numérica ordinal: MB = 4, B = 3, R = 2, I = 1 e AE = 0.

A padronização estatística por meio do StandardScaler, da biblioteca scikit-learn, foi aplicada exclusivamente à Regressão Logística, pois essa técnica é sensível à escala dos atributos de entrada. Random Forest e XGBoost, por se basearem em árvores de decisão, não dependem de escalonamento e não exigiram essa etapa.

Conforme o código, o StandardScaler foi ajustado apenas com os dados de treinamento (fit_transform no treinamento e transform na validação) em cada fold da validação cruzada estratificada por grupo. Esse procedimento evitou o vazamento de informações entre as partições.

Na Tabela 3, apresenta-se o resumo dos principais atributos da base final, com suas características gerais, a composição da variável-alvo e os tipos de dados empregados.

Quadro 2 — Síntese da base final

Descrição	Informação
Total de registros	298
Variável-alvo	evadiu (0 = não evadiu; 1 = evadiu)
Classe positiva	24,49% (73 registros)
Classe negativa	75,51% (225 registros)
Variáveis preditoras	15 atributos numéricos contínuos, ordinais ou binários
Unidade de análise	Registro do aluno por módulo
Período dos registros	abril de 2023 a dezembro de 2024

Fonte: A autora (2026).

A partir da base de dados finalizada segue a realização das fases 4 e 5.

4.1.1 ***Fase 4: Mineração de Dados Educacionais***

Nesta fase, foram aplicadas as técnicas selecionadas: Regressão Logística, *Random Forest e XGBoost*. Para a avaliação do desempenho, foi adotado um procedimento de validação cruzada estratificada (*Stratified k-fold cross-validation*). Nesse procedimento, os dados foram particionados em k = 5 subconjuntos (*folds*), mantendo, em cada partição, a proporção entre as classes de evasão e não evasão (Hastie; Tibshirani; Friedman, 2009; Witten; Frank; Hall, 2011).

A utilização da estratificação foi motivada pela distribuição assimétrica da variável-alvo, característica observada na base analisada. Em cada iteração, quatrofolds foram utilizados para treinamento e um para validação, de forma rotativa, permitindo que todas as instâncias fossem utilizadas em ambos os papéis ao longo do processo (Géron, 2019).

Durante a validação cruzada, foi utilizada a métrica F1-score como critério principal para avaliação e seleção dos modelos, por ser mais adequada a contextos de dados desbalanceados, como no caso da evasão escolar.

Esse procedimento foi aplicado tanto na avaliação preliminar dos modelos quanto no ajuste de hiperparâmetros dos algoritmos utilizados, mantendo o mesmo critério de avaliação ao longo das etapas de modelagem. (Fawcett, 2006; Provost; Fawcett, 2013).

Em razão do volume reduzido da base e do desbalanceamento entre as classes, adotou-se a validação cruzada estratificada por grupo (StratifiedGroupKFold, K = 5), da biblioteca scikit-learn. O procedimento substituiu a divisão holdout tradicional, que reservaria um conjunto fixo de teste e reduziria a quantidade de dados disponível para o treinamento. Com 298 registros, a validação cruzada alternou o papel das partições ao longo de cinco iterações e produziu estimativas menos dependentes de uma única divisão.

O agrupamento por aluno utilizou a variável id_anon como parâmetro groups no método split. Todos os registros de um mesmo aluno permaneceram no mesmo fold, o que evitou o compartilhamento de informações entre treinamento e validação.

O atributo id_anon foi excluído das variáveis preditoras antes do treinamento, por meio da instrução X = df.drop(columns=["id_anon", "evadiu"]), para impedir a memorização de padrões individuais.

Na divisão externa, os dados não foram embaralhados, pois o StratifiedGroupKFold foi configurado com cinco partições e sem o parâmetro shuffle. Essa opção preservou o agrupamento por aluno e evitou que registros de um mesmo indivíduo aparecessem simultaneamente nos conjuntos de treinamento e validação. O embaralhamento ocorreu apenas na validação interna do GridSearchCV, por meio do StratifiedKFold com shuffle=True e random_state=42.

Em cada iteração, aproximadamente 80% dos registros compuseram o treinamento e 20% formaram a validação externa. Não houve um conjunto fixo de teste. A proporção de validação variou entre 19,8% e 20,5%, pois o agrupamento por aluno impede uma divisão perfeitamente uniforme entre as partições.

O número de exemplos de treinamento variou entre 237 e 239 registros por fold. O Fold 1 contou com 237 registros de treinamento e 61 de validação; os Folds 2, 3 e 5 contaram com 239 registros de treinamento e 59 de validação; e o Fold 4 contou com 238 registros de treinamento e 60 de validação.

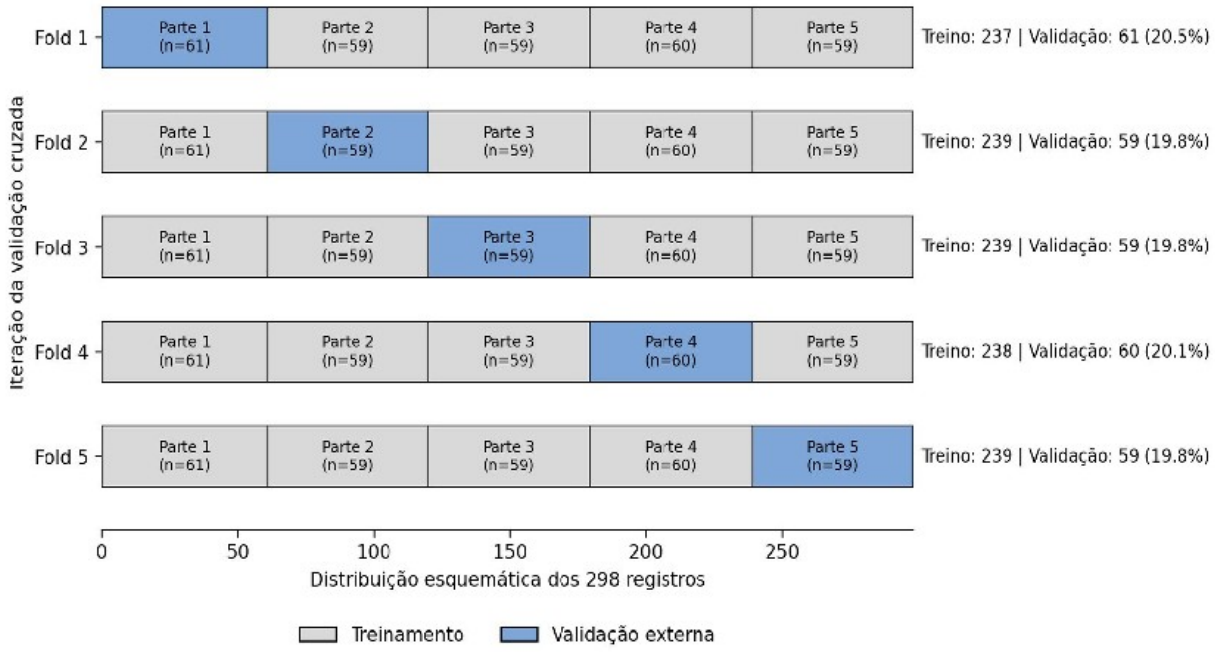
Tabela 4 — Distribuição dos registros nos cinco folds

Fold	Treinamento	Validação	Validação (%)
1	237	61	20,5
2	239	59	19,8
3	239	59	19,8
4	238	60	20,1
5	239	59	19,8

Fonte: A autora (2026)

A Figura 2 representa a rotação das partições ao longo das cinco iterações e evidencia a proporção destinada ao treinamento e à validação externa.

Figura 2 — Validação cruzada estratificada por grupo com cinco folds



Nota: representação esquemática. O agrupamento por aluno impede que registros do mesmo indivíduo apareçam simultaneamente no treinamento e na validação.

Fonte: A autora (2026).

O desempenho de cada fold compôs a média consolidada das técnicas. A Tabela 5 apresenta a acurácia e o recall do Random Forest sob limiar otimizado, técnica que obteve o melhor resultado geral.

Tabela 5 — Desempenho do Random Forest em cada fold sob limiar otimizado

Fold	Registros de validação	Acurácia	Recall
1	61	0,951	1,000
2	59	0,966	1,000
3	59	0,949	1,000
4	60	0,967	1,000
5	59	0,983	1,000
Média	-	0,963	1,000

Fonte: A autora (2026).

O desempenho geral foi considerado satisfatório nas três técnicas. O Random Forest apresentou os melhores resultados consolidados sob limiar otimizado, com acurácia média de 0,963, recall de 1,000, F1-score de 0,931 e AUC de 0,990. A pequena variação da acurácia entre os folds, com desvio padrão de aproximadamente 0,012, indica estabilidade entre as partições.

As matrizes de confusão apresentadas não correspondem ao fold com melhor desempenho. Elas reúnem as previsões acumuladas dos cinco folds, de modo que cada um dos 298 registros participou uma única vez da validação externa. Portanto, as matrizes representam o comportamento consolidado de cada técnica e configuração de limiar.

A acurácia geral corresponde à média aritmética das acurácias dos cinco folds, e não ao resultado isolado do melhor fold. No Random Forest, a média de 0,963 resultou dos valores 0,951, 0,966, 0,949, 0,967 e 0,983. O Fold 5 apresentou a maior acurácia individual, mas esse valor não foi adotado como desempenho geral.

4.1.1.1 **Limiar de Decisão Padrão**

A primeira etapa de avaliação consistiu em analisar o desempenho das técnicas aplicadas à base de dados, utilizando o limiar de decisão padrão de 0,50, conforme adotado pelos classificadores probabilísticos do scikit-learn.

Essa configuração inicial permitiu comparar os modelos sob uma mesma condição, antes de eventuais ajustes de limiar ou definição de custos diferenciados de erro.

O limiar de decisão foi definido para cada técnica e para cada fold por meio da função melhor_limiar_validacao. A função avaliou 100 valores entre 0,01 e 0,99 e selecionou aquele que maximizou o F1-score no conjunto de validação da respectiva iteração. Como o procedimento não adotou um conjunto de teste fixo, os resultados apresentados correspondem às validações externas acumuladas dos cinco folds.As técnicas selecionadas foram aplicadas à base de dados utilizando configurações específicas de parâmetros.

A definição desses parâmetros foi realizada por meio de validação cruzada estratificada, com o uso do método *GridSearchCV*, considerando diferentes combinações de hiperparâmetros para cada algoritmo.

Esse procedimento teve como objetivo otimizar o desempenho dos modelos, tomando como referência a métrica F1-score, além de possibilitar a reprodutibilidade dos experimentos. Os principais parâmetros adotados para cada técnica são apresentados na Tabela 6.

Tabela 6 — Parâmetros utilizados nas técnicas de classificação

Técnica	Configuração adotada
Regressão Logística	Modelo logit com penalização L2, class_weight="balanced", max_iter=2000 e ajuste do parâmetro C por validação cruzada.
Random Forest	Critério de Gini, class_weight="balanced" e busca dos hiperparâmetros n_estimators, max_depth e min_samples_leaf por validação cruzada.
XGBoost	Árvores sequenciais de gradient boosting, eval_metric="logloss" e busca dos hiperparâmetros n_estimators, max_depth e learning_rate por validação cruzada.

Fonte: A autora (2026).

O ajuste dos hiperparâmetros adotou o GridSearchCV da biblioteca scikit-learn, com validação cruzada interna de três partições (StratifiedKFold, n_splits=3, shuffle=True, random_state=42) e F1-score como critério de seleção (scoring="f1"). Em cada fold externo, o GridSearchCV teve acesso apenas aos dados de treinamento e não consultou a partição de validação externa.

A Tabela 7 apresenta os valores candidatos testados e a configuração identificada com maior frequência nos cinco folds.

Tabela 7 — Hiperparâmetros testados e melhores configurações das técnicas

Técnica	Hiperparâmetros testados	Melhor configuração identificada
Regressão Logística	C: [0.1, 1, 10] | max_iter: 2000 | class_weight: balanced	C = 1 (mais frequente nos 5 folds)
Random Forest	n_estimators: [100, 200, 300] | max_depth: [5, 6, 10] | min_samples_leaf: [1, 3] | class_weight: balanced	n_estimators=200, max_depth=6, min_samples_leaf=3
XGBoost	n_estimators: [100, 200] | max_depth: [3, 4, 5] | learning_rate: [0.05, 0.1] | eval_metric: logloss	n_estimators=200, max_depth=5, learning_rate=0.1

Fonte: A autora (2026)

O desbalanceamento entre as classes foi tratado pelo parâmetro class_weight="balanced" na Regressão Logística e no Random Forest. Esse parâmetro atribuiu pesos inversamente proporcionais à frequência de cada classe.

No XGBoost, a análise do desbalanceamento considerou as métricas de avaliação e o ajuste do limiar de decisão.

No Quadro 3 são apresentados os resultados das métricas para os modelos avaliados, considerando tanto o limiar de decisão padrão (0,50) quanto o limiar otimizado.

Quadro 3 — Métricas de desempenho dos modelos de classificação nos cenários de limiar padrão (0,50) e limiar otimizado

Modelo	**Configuração**	**Acurácia**	**Precisão**	**Recall**	**F1**	**ROC**	**Limiar**
Regressão Logística	Padrão (0,50)	0,913	0,771	0,919	0,838	0,968	0,500
Regressão Logística	Otimizado	0,933	0,826	0,932	0,873	0,968	0,578
Random Forest	Padrão (0,50)	0,950	0,890	0,918	0,899	0,990	0,500
Random Forest	Otimizado	0,963	0,871	1,000	0,931	0,990	0,392
XGBoost	Padrão (0,50)	0,929	0,884	0,823	0,849	0,987	0,500
XGBoost	Otimizado	0,956	0,884	0,959	0,917	0,987	0,246

Fonte: A autora (2026).

A Regressão Logística apresentou, no cenário com limiar de decisão padrão (0,50), precisão de 0,771, recall de 0,919, F1-score de 0,838 e ROC de 0,968. Após a otimização do limiar (0,578), observou-se melhora no desempenho, com precisão de 0,826, recall de 0,932 e F1-score de 0,873, mantendo o mesmo valor de ROC.

Esse comportamento é coerente com o que Hastie, Tibshirani e Friedman (2009) descrevem para modelos lineares, que tendem a apresentar estabilidade e ganhos moderados quando ajustados em bases com relações estruturadas entre variáveis. O aumento do recall indica maior capacidade de identificação de alunos em risco de evasão após o ajuste do limiar.

O modelo *Random Forest* apresentou o melhor desempenho geral entre as técnicas avaliadas. No limiar padrão (0,50), obteve acurácia de 0,950, precisão de 0,890, *recall* de 0,918, F1-*score* de 0,899 e ROC de 0,990. Após a otimização do limiar (0,392), o recall atingiu 1,000 e o F1-*score* subiu para 0,931, com acurácia de 0,963 e precisão de 0,871, mantendo o mesmo valor de ROC.

Segundo Breiman (2001), esse desempenho está relacionado à capacidade do

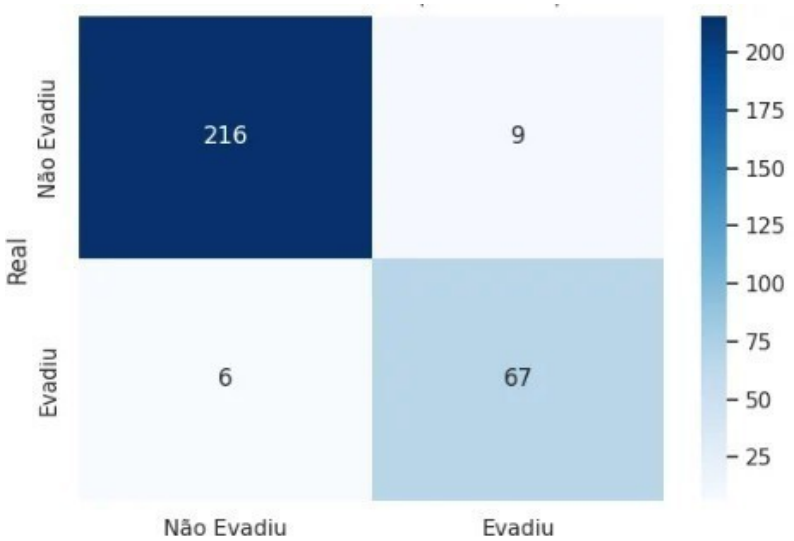
método de explorar interações não lineares e reduzir a variância por meio do uso demúltiplas árvores. Em estudos educacionais, técnicas baseadas em árvores têm sido amplamente utilizadas em tarefas de identificação de risco devido à capacidade de modelar relações complexas entre variáveis como engajamento, frequência e desempenho (Baker; Inventado, 2014).

O *XGBoost* apresentou comportamento distinto em relação aos demais modelos. No limiar padrão (0,50), obteve acurácia de 0,929, precisão de 0,884, recall de 0,823 e F1-score de 0,849, com ROC de 0,987. Após a otimização do limiar (0,246), houve melhora no desempenho, com precisão de 0,884, *recall* de 0,959 e F1-*score* de 0,917.

Esse resultado mostra a sensibilidade do modelo à escolha do limiar de decisão. Conforme discutido por He e Garcia (2009), classificadores aplicados a bases desbalanceadas tendem a concentrar decisões na classe majoritária quando avaliados com o limiar padrão, sendo necessário o ajuste do ponto de corte para melhorar a detecção da classe minoritária, neste caso, os alunos evadidos.

As Figuras 3, 4 e 5 apresentam as matrizes de confusão dos modelos Regressão Logística, *Random Forest e XGBoost,* respectivamente, considerando o limiar de decisão padrão (0,50).

Figura 3 – Matriz de confusão Random Forest no limiar de decisão padrão (0,50)



Fonte: A autora (2026).

Na Figura 3 observa-se a presença de um número reduzido de falsos negativos (6 casos), indicando que a maioria dos alunos que evadiram foi corretamente identificada, resultando em um recall de 0,918.

Esse padrão é importante em sistemas de alerta educacional, pois, conforme argumentam Baker e Inventado (2014), a não identificação de alunos em risco (falsos negativos) é mais prejudicial do que a geração de alertas adicionais para alunos que permaneceriam no curso.

Figura 4 – Matriz de confusão XGBoost no limiar de decisão padrão (0,50)

Fonte: A autora (2026).

Na Figura 4, referente ao modelo XGBoost, observa-se a presença de falsos negativos (13 casos), coerente com o recall de 0,823. Isso indica que parte dos alunos que evadiram não foi identificada pelo modelo no limiar padrão, mostrando menor capacidade de detecção da classe minoritária.

Figura 5 – Matriz de confusão Regressão Logística no limiar de decisão padrão (0,50)

Fonte: A autora (2026).

Na Figura 5, referente à Regressão Logística, observa-se uma distribuição

entre falsos positivos (20 casos) e falsos negativos (6 casos), com recall de 0,918.

Esse comportamento é compatível com a interpretação típica de modelos lineares em bases educacionais, que tendem a apresentar estabilidade e capacidade moderadade separação entre as classes, conforme discutido por Hastie, Tibshirani e Friedman (2009).

De forma geral, os resultados indicam que o impacto do ajuste do limiar de decisão varia conforme o modelo utilizado. Enquanto o *Random Forest* apresentou desempenho estável independentemente do limiar, os modelos Regressão Logística e *XGBoost* apresentaram melhorias após a otimização, especialmente em termos de recall e F1-score.

4.1.1.2 **Análise do Ajuste do Limiar de Decisão**

Em modelos de classificação, o limiar de decisão corresponde ao valor a partir do qual o modelo define se um aluno será classificado como evadido ou não. Em classificadores probabilísticos, esse limiar determina como as probabilidades estimadas são convertidas em classes, influenciando diretamente o equilíbrio entre os diferentes tipos de erro.

O valor padrão adotado é 0,50; contudo, esse ponto de corte nem sempre é o mais adequado, especialmente em contextos com classes desbalanceadas e consequências distintas para os erros de classificação (Fawcett, 2006).

Deixar de identificar um aluno em risco de evasão constitui um erro mais grave do que sinalizá-lo de forma indevida. Um aluno não identificado perde a oportunidade de receber apoio institucional, enquanto uma sinalização incorreta pode ser posteriormente analisada por docentes ou gestores educacionais.

Por essa razão, o limiar foi ajustado para valores inferiores a 0,50, de modo a aumentar a sensibilidade do modelo à identificação de alunos em risco, prática recomendada em cenários com custos assimétricos de erro (Elkan, 2001).

A análise do ajuste do limiar de decisão teve como objetivo avaliar o impacto da variação do ponto de corte no desempenho das técnicas, com ênfase em métricas adequadas a dados desbalanceados, como a medida F1-*score* e a revocação. Inicialmente, considerou-se o limiar padrão de 0,50.

Em seguida, o limiar passou a ser tratado como um parâmetro ajustável, otimizado com base na maximização da medida F1-*score*, calculada a partir das probabilidades estimadas pelos modelos.

A escolha dessa métrica justifica-se por integrar precisão e revocação em uma única medida, permitindo um equilíbrio entre a identificação correta dos alunos evadidos e a redução de classificações incorretas, sendo particularmente adequada para cenários com desbalanceamento de classes (Saito; Rehmsmeier, 2015).

O limiar otimizado foi definido, em cada partição da validação cruzada, como o valor que maximiza a medida F1-score nos dados de validação. Para isso, foram avaliados valores de limiar entre 0,01 e 0,99, com incrementos de 0,01, e selecionado aquele que produziu o maior valor da métrica. Esse procedimento permitiu analisar diferentes configurações de decisão e seus impactos nos erros de classificação, sobretudo nos falsos negativos, que são críticos neste contexto (Sokolova; Lapalme, 2009).

Ao final das cinco partições, calculou-se a média dos limiares obtidos, que passou a constituir o limiar final adotado na avaliação da técnica. Os resultados evidenciaram que a modificação do ponto de corte influenciou diretamente o desempenho das técnicas, especialmente na identificação da classe minoritária.

Após a definição do limiar otimizado, analisou-se o impacto desse ajuste no desempenho das técnicas.

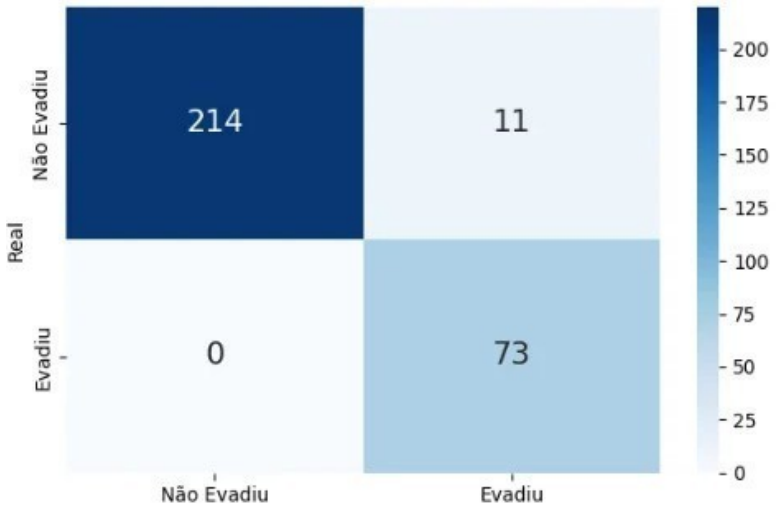
Figura 6 – Matriz de confusão da regressão Logística com limiar otimizado (0,578)

Fonte: A autora (2026).

No caso da Regressão Logística, com o ajuste do limiar de 0,50 para 0,578, a técnica passou a identificar corretamente 68 dos 73 alunos evadidos, reduzindo os falsos negativos para apenas 5 casos, conforme apresentado na Figura 6.

O limiar padrão subestimava situações de risco, classificando alguns alunos como nãoevadidos mesmo apresentando indícios de evasão. Com a otimização, o F1-score aumentou de 0,838 para 0,873, indicando melhora na detecção sem perda relevante de precisão.

Figura 7 – Matriz de confusão da Random Forest com limiar otimizado (0,392)



Fonte: A autora (2026).

A redução do limiar de 0,50 para 0,392 no modelo Random Forest resultou no melhor desempenho entre as três técnicas, com recall de 1,000 isto é, todos os 73 alunos evadidos foram corretamente identificados, sem ocorrência de falsos negativos, conforme apresentado na Figura 7. Esse resultado indica que a técnica identificou integralmente a classe minoritária, em concordância com abordagens que priorizam a redução de falsos negativos em cenários com custos de erro assimétricos (Mazurowski, 2018).

A obtenção de recall máximo, nesse contexto, é justificado do ponto de vista metodológico, pois decorre do ajuste do limiar com base na maximização da medida F1-score em validação cruzada estratificada por grupo, procedimento que contribui para reduzir viés de avaliação e preservar a capacidade de generalização do modelo (Kuhn; Johnson, 2019).

Além disso, a adoção de um limiar inferior a 0,50 aumenta a sensibilidade do modelo, estratégia frequentemente empregada em problemas com classes desbalanceadas e foco na detecção da classe positiva (Fernández et al., 2018).

Em comparação ao limiar padrão, que ainda deixava alguns alunos em risco sem identificação, o ajuste eliminou essa limitação, elevando o F1-score de 0,899 para 0,931. Esses resultados indicam que o modelo *Random Forest* apresenta desempenho adequado para aplicações em que a identificação de alunos em risco constitui prioridade.

XGBoost apresentou a maior redução do limiar, de 0,50 para 0,246, indicando tendência mais conservadora na classificação da classe minoritária quando utilizado o ponto de corte padrão. Após o ajuste, o recall aumentou de 0,823 para 0,959, com redução dos falsos negativos de 13 para 3 casos, conforme apresentado na Figura 8.

Figura 8— Matriz de confusão do modelo XGBoost com limiar otimizado (0,246)

Fonte: A autora (2026).

O F1-score passou de 0,849 para 0,917, enquanto a precisão permaneceu estável em 0,884, indicando que o aumento na identificação de alunos evadidos não implicou crescimento relevante de falsos positivos. Esses resultados demonstram que a otimização do limiar contribuiu para aproximar o desempenho do XGBoost ao observado no Random Forest.

Os resultados consolidados das cinco partições permitiram comparar os modelos nos cenários com limiar padrão e otimizado. A otimização do limiar melhorou principalmente o recall e o F1-score, sem perda relevante de precisão.

Na Regressão Logística, a acurácia aumentou de 0,913 para 0,933 e o F1-score de 0,838 para 0,873, com recall de 0,919 para 0,932 e precisão de 0,771 para 0,826. O limiar médio foi de 0,578. No *Random Forest*, a acurácia passou de 0,950 para 0,963 e o F1-score de 0,899 para 0,931. O *recall* atingiu 1,000, enquanto a precisão reduziu de 0,890 para 0,871. O limiar médio foi de 0,392.

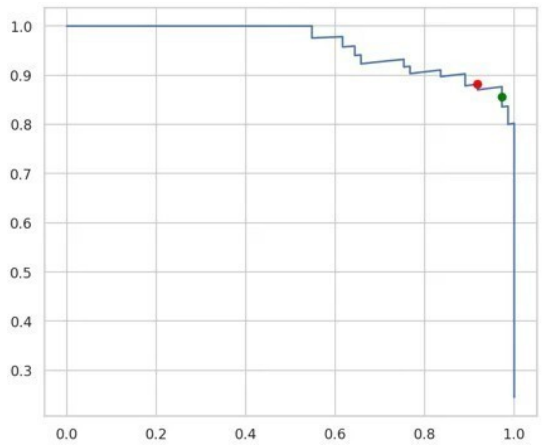
No *XGBoost*, a acurácia aumentou de 0,929 para 0,956 e o F1-score de 0,849 para 0,917. O *recall* passou de 0,823 para 0,959, com precisão mantida em 0,884. O limiar médio foi de 0,246.

A ROC permaneceu estável em todos os modelos, confirmando que a otimização do limiar não alterou a capacidade discriminativa global dos classificadores, conforme esperado, uma vez que a ROC avalia o desempenho ao longo de todos os limiares possíveis (Fawcett, 2006).

De modo geral, a redução do limiar resultou em aumento da revocação e melhoria da F1-score, indicando maior capacidade de detecção de alunos em risco de evasão. Todo o processo de ajuste foi conduzido exclusivamente com base nos dados de validação, sem utilização dos dados de teste, o que preserva a integridade da avaliação e assegura a validade metodológica dos resultados (Kohavi, 1995; Hand, 2009).

A Figura 9 apresenta a curva Precision-Recall do modelo Random Forest, no qual se observa a relação entre precisão e recall ao longo dos diferentes valores de limiar. O ponto correspondente ao limiar padrão (0,50) apresenta menor recall em comparação ao limiar otimizado (0,392), que se posiciona em uma região da curva associada a melhor equilíbrio entre precisão e recall, conforme indicado pelo maior valor de F1-score.

Figura 9 — Curva Precision-Recall Random Forest



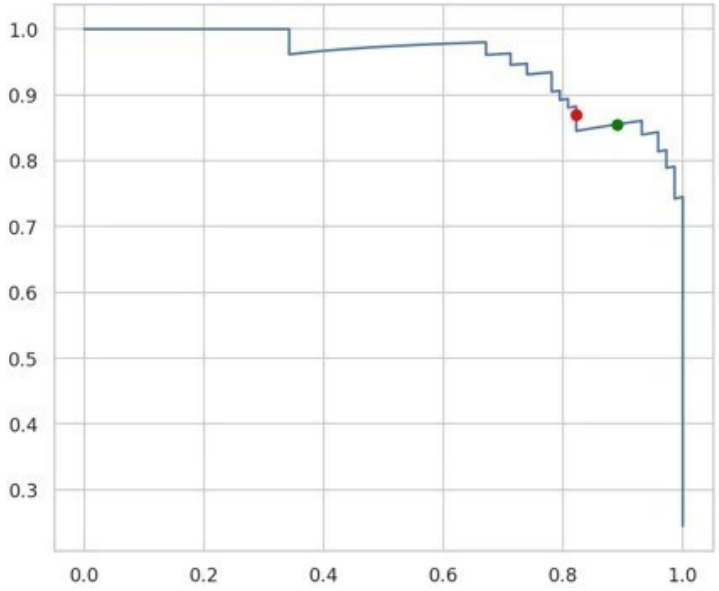
Fonte: A autora (2026).

A curva *Precision-Recall* apresentou a relação entre precisão e *recall* para diferentes valores de limiar de decisão, permitindo analisar como a variação do ponto de corte influenciou o desempenho do modelo em um contexto de dados desbalanceados. Observou-se que o modelo manteve valores elevados de precisão ao longo de uma ampla faixa de *recall*, com redução gradual da precisão à medida que o recall se aproximou de níveis mais altos.

Na figura, foram indicados dois pontos de operação: o limiar padrão de 0,50 (representado em vermelho) e o limiar otimizado (representado em verde), definido a partir da maximização do F1-score. A comparação entre esses pontos mostrou que o ajuste do limiar deslocou o modelo para uma região com maior recall e leve redução na precisão. Esse comportamento foi adequado ao problema analisado, pois aumentou a capacidade de identificação de alunos em risco de evasão, mesmo com o possível aumento de falsos positivos.

A Figura 10 mostra a curva Precision-Recall da técnica XGBoost. Verificou-se que o modelo manteve níveis elevados de precisão nas regiões iniciais da curva, com redução gradual à medida que o recall aumentou. Esse comportamento indicou um equilíbrio entre as métricas, especialmente em faixas mais altas de recall.

Figura 10 – Curva Precision-Recall XGBoost



Fonte: A autora (2026).

No ponto correspondente ao limiar padrão (0,50), observou-se um valor de recall de 0,823, com precisão de 0,884.

Já no limiar otimizado, o modelo apresentou aumento no *recall*, acompanhado de leve redução na precisão, deslocando o ponto de operação para uma região mais

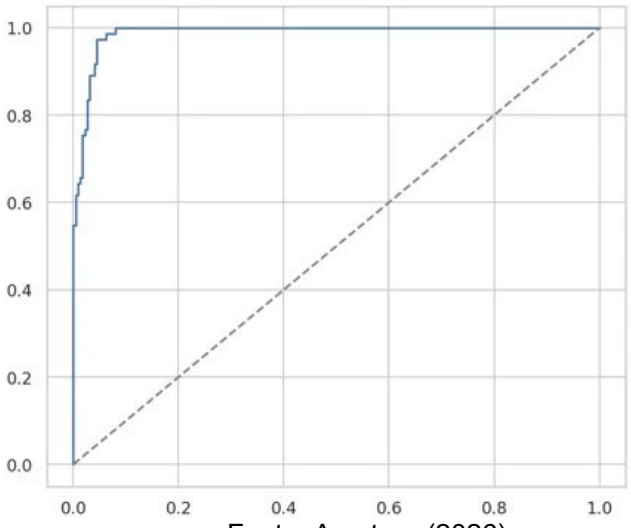
adequada à identificação da classe minoritária, correspondente aos alunos evadidos.

Esse resultado indicou melhora na sensibilidade do modelo, tornando-o mais eficaz na detecção de alunos em risco de evasão, ainda que com um pequeno aumento na taxa de falsos positivos.

À medida que o limiar de decisão foi reduzido, observou-se aumento do *recall*, acompanhado de diminuição gradual da precisão, caracterizando o *trade-off* entre essas métricas. Esse comportamento indicou a necessidade de ajuste do limiar como estratégia para ampliar a capacidade de identificação dos alunos evadidos, conforme verificado nos resultados obtidos com o limiar otimizado.

A Figura 11 mostra a curva Precision-Recall do modelo de Regressão Logística. Observou-se um comportamento estável, com redução gradual da precisão à medida que o recall aumentou, indicando uma transição mais suave entre essas métricas ao longo dos diferentes limiares de decisão.

Figura 11 — Curva Precision-Recall Regressão Logística



Fonte: A autora (2026).

As curvas *Precision-Recall* indicaram que os modelos responderam de forma distinta à variação do limiar de decisão. De modo geral, verificou-se que a redução do limiar resultou em aumento do *recall*, acompanhado de diminuição gradual da precisão, caracterizando o comportamento esperado em cenários com dados desbalanceados.

A técnica *Random Forest* apresentou o melhor desempenho global, mantendo níveis elevados de precisão mesmo em faixas mais altas de *recall*. A curva mostrou-se mais próxima da região superior do gráfico, indicando alta capacidade de identificação da classe minoritária, correspondente aos alunos evadidos, com baixo impacto na taxa de falsos positivos. Além disso, o ajuste do limiar teve impacto

reduzido sobre seu desempenho, evidenciando maior estabilidade do modelo.

O modelo *XGBoost* também apresentou bom desempenho, com curva concentrada em níveis elevados de precisão e *recall*. A redução do limiar promoveu aumento perceptível no recall, com leve redução na precisão, deslocando o ponto de operação para uma região mais adequada à identificação dos alunos evadidos. Esse comportamento indicou ganho relevante em sensibilidade.

Já o modelo de Regressão Logística apresentou desempenho mais moderado, com redução mais acentuada da precisão à medida que o *recall* aumentou. Embora o ajuste do limiar tenha promovido melhora na capacidade de identificação da classe minoritária, essa melhoria ocorreu de forma mais limitada quando comparada aos demais modelos.

De forma geral, os resultados indicaram que a utilização do limiar padrão (0,50) não foi a configuração mais adequada para todos os modelos em um contexto de dados desbalanceados, como o da evasão escolar. O ajuste do limiar contribuiu para aumentar a identificação de alunos evadidos, sendo particularmente relevante para as técnicas XGBoost e Regressão Logística, enquanto a Random Forest apresentou desempenho elevado mesmo sem alterações no ponto de corte.

4.1.1.3 **Resultados Comparativos**

Na comparação entre as técnicas, verificou-se que o *Random Forest* apresentou o melhor desempenho geral, com curva mais próxima da região ideal e maior capacidade de discriminação entre as classes.

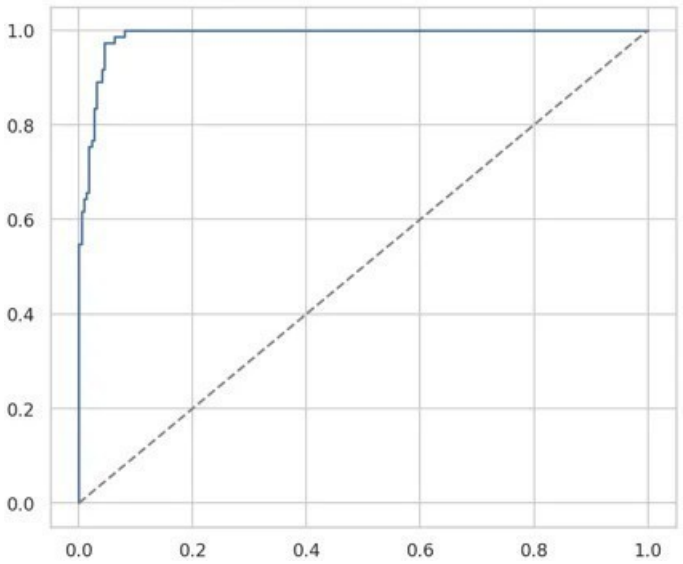
A técnica *XGBoost* também apresentou desempenho elevado, com comportamento semelhante, porém ligeiramente inferior ao *Random Forest* em determinadas regiões da curva. Já a Regressão Logística apresentou desempenho adequado, embora com menor separação entre as classes quando comparada aos demais modelos.

De forma geral, os resultados indicaram que os modelos baseados em árvores apresentaram maior capacidade de capturar padrões complexos nos dados, refletindo em melhor desempenho discriminativo. A seguir, são apresentadas as análises individuais das curvas ROC para cada modelo.

A Figura 12 mostra a curva ROC do modelo Random Forest. Observou-se que a curva permaneceu próxima ao canto superior esquerdo do gráfico, indicando altas taxas de verdadeiros positivos associadas a baixas taxas de falsos positivos. Esse

comportamento indicou elevada capacidade de discriminação do modelo na distinção entre alunos evadidos e não evadidos.

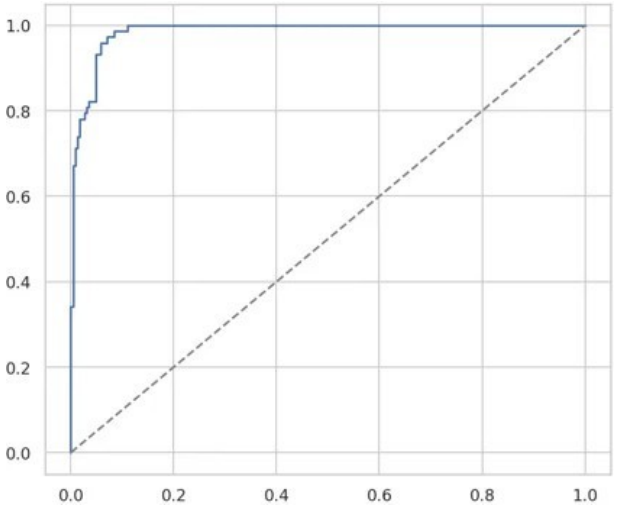
Figura 12 – Curva ROC Random Forest



Fonte: A autora (2026).

A Figura 13 mostra a curva de calibração da técnica Random Forest. Observou-se que as probabilidades previstas apresentaram variações em relação à linha de calibração perfeita, com pontos acima e abaixo da diagonal ao longo dos intervalos analisados. Esse comportamento indicou que, em determinados níveis de probabilidade, o modelo superestimou ou subestimou o risco de evasão.

Figura 13 — Curva de calibração do modelo Random Forest



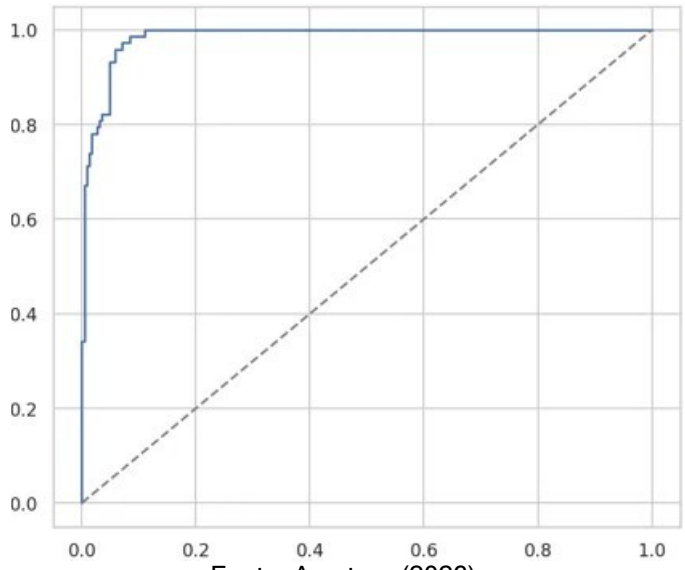
Fonte: A autora (2026).

Apesar dessas oscilações, verificou-se uma tendência de aproximação à linha ideal nos níveis mais elevados de probabilidade, sugerindo estimativas mais consistentes para os casos de maior risco. Esse resultado indicou que, embora o modelo apresente boa capacidade discriminativa, as probabilidades previstas devem ser interpretadas com cautela, especialmente em faixas intermediárias.

De acordo com Alexandru Niculescu-Mizil e Rich Caruana (2005), técnicas bem calibradas são aquelas cujas probabilidades previstas correspondem às frequências observadas, sendo fundamentais em aplicações onde as previsões são utilizadas como estimativas de risco.

A Figura 14 mostra a curva ROC do modelo XGBoost. Observou-se que a curva permaneceu próxima ao canto superior esquerdo do gráfico, indicando altas taxas de verdadeiros positivos associadas a baixas taxas de falsos positivos. Esse comportamento indicou bom desempenho do modelo na diferenciação entre alunos evadidos e não evadidos.

Figura 14 – Curva ROC XGBoost



Fonte: A autora (2026).

Em comparação a *Random Forest,* verificou-se que o desempenho do *XGBoost* foi ligeiramente inferior, com maior afastamento da região ideal em determinados trechos da curva.

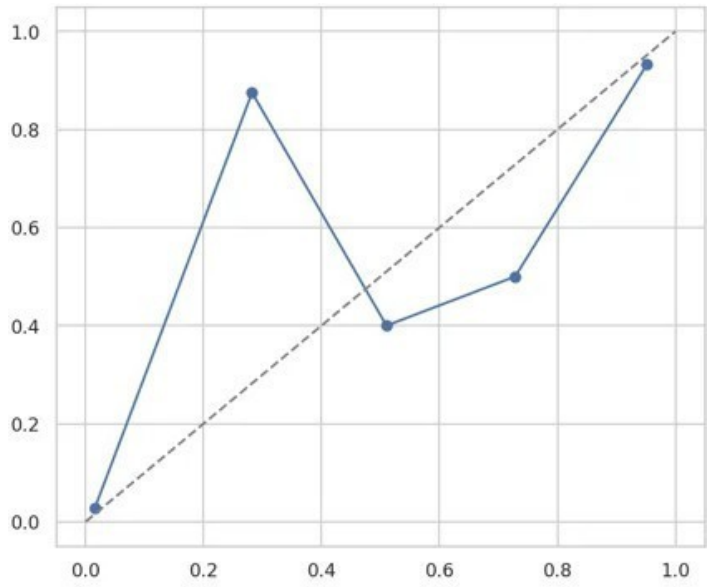
Esse comportamento sugeriu uma menor estabilidade na separação entre as classes, embora o modelo tenha mantido capacidade discriminativa ao longo da maior parte dos limiares avaliados.

Ainda assim, a técnica *XGBoost* mostrou-se adequada para a tarefa, especialmente por sua capacidade de modelar relações não lineares e interações entre variáveis.

De acordo com Tianqi Chen e Carlos Guestrin (2016), a técnica *XGBoost* é um método baseado em *boosting* que combina múltiplos modelos fracos de forma sequencial, otimizando funções de perda e controlando a complexidade da técnica, o que contribui para seu alto desempenho em tarefas de classificação.

Além disso, estudos mais recentes, como Aurélien Géron (2022), mostram que, embora modelos de *boosting* frequentemente apresentem alta capacidade preditiva, sua performance pode variar conforme a configuração de hiperparâmetros e a natureza dos dados, especialmente em cenários com ruído ou desbalanceamento.

Figura 15 – Curva de calibração do modelo XGBoost



Fonte: A autora (2026).

Conforme ilustrado na Figura 15, verificou-se que, em determinadas faixas de probabilidade, a técnica superestimou a probabilidade de evasão, enquanto em outras apresentou subestimação, o que indicou inconsistência na calibração ao longo dosníveis analisados. Esse comportamento indicou desalinhamento entre as probabilidades previstas e as frequências observadas, comprometendo a interpretação direta dessas probabilidades como medidas confiáveis de risco.

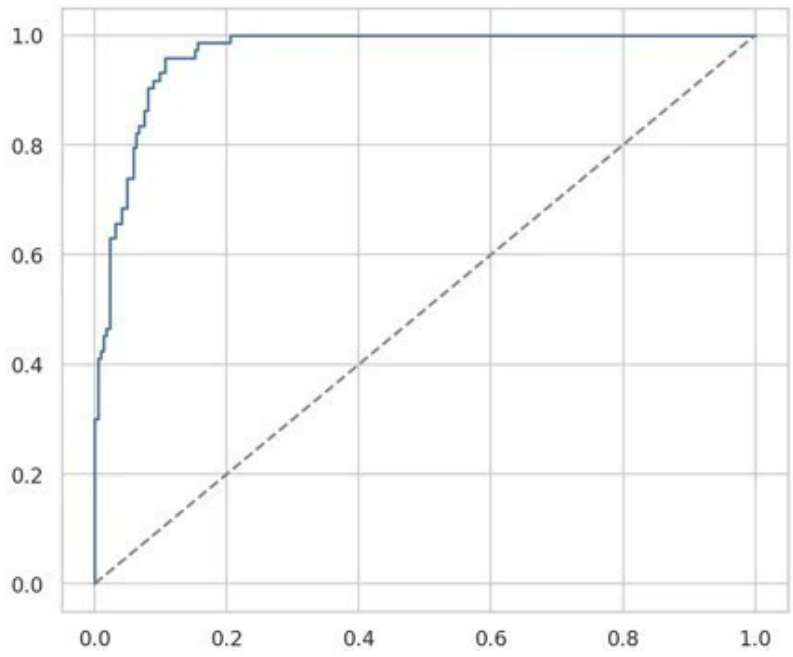
Observou-se que esses desvios ocorreram de forma não uniforme, o que indicou variação na qualidade das estimativas ao longo dos diferentes intervalos de probabilidade. Em níveis intermediários, as discrepâncias foram mais evidentes, o que sugeriu maior instabilidade nas previsões.

Em níveis mais elevados, houve maior aproximação à linha ideal, o que indicou melhor consistência na identificação de casos de maior risco.

Esse resultado indicou que, embora a técnica *XGBoost* tenha apresentado boa capacidade discriminativa, suas probabilidades não refletiram de forma plenamente fiel o risco real de evasão em todas as faixas analisadas. Em aplicações educacionais, nas quais essas probabilidades podem orientar intervenções, essa limitação reforçou a necessidade de cautela na interpretação dos valores previstos, bem como a importância de considerar o uso complementar de outras métricas ou técnicas de ajuste.

A Figura 16 apresenta a curva ROC da técnica de Regressão Logística. Observou-se que a curva se manteve próxima ao canto superior esquerdo do gráfico, indicando boa capacidade de discriminação entre alunos evadidos e não evadidos, com taxas elevadas de verdadeiros positivos associadas a baixas taxas de falsos positivos.

Figura 16 – Curva ROC Regressão Logística



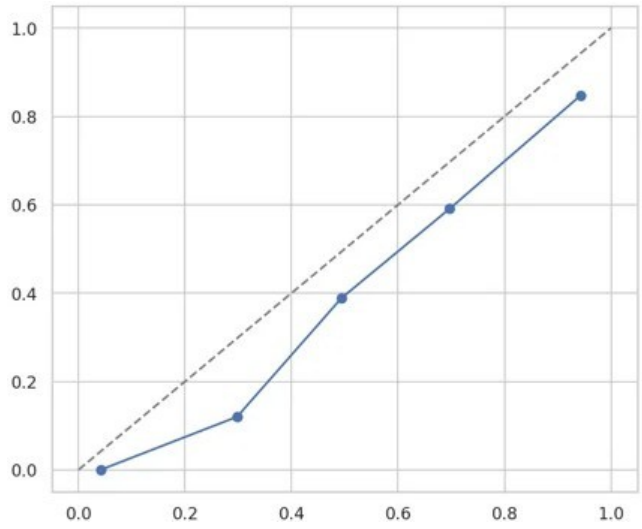
Fonte: A autora (2026).

No entanto, em comparação as técnicas baseadas em árvores, verificou-se

que a curva apresentou maior afastamento da região ideal em determinados trechos, o que indicou menor capacidade de separação entre as classes. Ainda assim, o modelo apresentou desempenho adequado na diferenciação entre alunos evadidos e não evadidos, embora inferior ao observado no *Random Forest e no XGBoost.*

A Figura 17 mostra a curva de calibração da técnica de Regressão Logística. Observou-se que as probabilidades previstas apresentaram desvios em relação à linha de calibração perfeita ao longo dos intervalos analisados, indicando diferenças entre os valores estimados e as frequências observadas.

Figura 17 — Curva de calibração Regressão Logística



Fonte: A autora (2026).

Nas faixas intermediárias de probabilidade, observaram-se diferenças mais evidentes entre os valores previstos e os observados, enquanto nos níveis mais elevados houve maior aproximação à linha ideal. Esse resultado indicou que as probabilidades estimadas variaram ao longo dos intervalos, não representando de forma uniforme o risco real de evasão.

Em comparação ao *Random Forest,* verificou-se maior afastamento da linha ideal, o que indicou menor adequação das probabilidades previstas como medida direta de risco. Conforme discutido na literatura, modelos bem calibrados devem apresentar alinhamento entre probabilidades previstas e frequências observadas, sendo esse um aspecto importante em aplicações baseadas em estimativas de risco (Niculescu-Mizil e Caruana, 2005).

De forma geral, a análise conjunta das curvas ROC e de calibração indicou

que as técnicas apresentaram bom desempenho na identificação de alunos que evadiram (classe positiva) e alunos que permaneceram (classe negativa), com diferenças entre eles. A curva ROC permitiu avaliar a capacidade de separação entre essas duas classes, enquanto a curva de calibração mostrou o alinhamento entre as probabilidades previstas e as frequências observadas.

A técnica *Random Forest* apresentou melhor desempenho, com maior capacidade de discriminação entre alunos evadidos e não evadidos e maior proximidade da linha ideal na calibração, principalmente em níveis mais altos de probabilidade. Esse resultado indicou melhor adequação tanto na classificação quanto na utilização das probabilidades como medida de risco (Breiman, 2001).

A técnica *XGBoost* apresentou capacidade de discriminação, com curva ROC próxima à do *Random Forest*. No entanto, a calibração apresentou desvios ao longo das faixas de probabilidade, o que indicou diferenças entre os valores previstos e observados para a evasão. Esse comportamento limitou o uso direto das probabilidades como estimativa de risco (Niculescu-Mizil e Caruana, 2005).

A Regressão Logística apresentou menor capacidade de discriminação em comparação aos modelos baseados em árvores na separação entre alunos evadidos e não evadidos. A calibração apresentou afastamento da linha ideal em diferentes intervalos, o que indicou menor correspondência entre probabilidades previstas e valores observados (James et al., 2021).

Esses resultados indicaram que a avaliação dos modelos deve considerar tanto a capacidade de classificar corretamente alunos evadidos e não evadidos quanto a qualidade das probabilidades previstas, especialmente em aplicações que envolvem análise de risco, como a evasão escolar.

A importância das variáveis foi analisada como complemento à interpretação dos resultados do modelo *Random Forest,* com o objetivo de identificar os atributos com maior contribuição para a predição da evasão, conforme descrito na literatura sobre métodos baseados em árvores (Breiman, 2001).

A análise utilizou a métrica de redução média de impureza (*Mean Decrease in Impurity MDI*), que quantifica a contribuição de cada variável na redução da impureza de Gini ao longo das árvores da técnica.

Os resultados indicaram que as variáveis mais relevantes foram: (1) percentual médio de frequência; (2) total de faltas; (3) média de desempenho; e (4) número de disciplinas em situação crítica. Essas variáveis concentraram a maior parte da capacidade discriminativa do modelo, em consonância com a literatura sobre evasão em cursos técnicos (Musso et al., 2020; Silva e Pereira, 2021).

Variáveis associadas à variação da frequência ao longo do módulo também apresentaram relevância, indicando que mudanças na trajetória de presença possuem valor preditivo para evasão.

Do ponto de vista da gestão escolar, esses resultados indicam que os principais fatores preditivos correspondem a indicadores já disponíveis nos registros institucionais, como frequência e desempenho, permitindo seu uso para monitoramento e intervenção.

4.1.1.4 **Análise dos Resultados**

A análise das matrizes de confusão indicou diferenças entre as técnicas.

A *Random Forest* apresentou menor número de falsos negativos, com maior capacidade de identificar corretamente os alunos que evadiram. A *XGBoost* apresentou desempenho próximo, porém com maior ocorrência de erros na classificação da evasão. A Regressão Logística apresentou maior número de classificações incorretas, principalmente na identificação de alunos evadidos.

As curvas ROC confirmaram esses resultados. A técnica *Random Forest* apresentou curva mais próxima do canto superior esquerdo, indicando melhor capacidade de separação entre alunos evadidos e não evadidos.

A técnica *XGBoost* apresentou desempenho semelhante, com pequena diferença em relação a técnica anterior. A Regressão Logística apresentou maior afastamento da região ideal, indicando menor capacidade discriminativa.

A utilização da curva ROC para avaliar modelos de classificação é amplamente consolidada na literatura, pois permite analisar o desempenho considerando diferentes limiares de decisão (Fawcett, 2006).

A análise das curvas *Precision-Recall* mostrou diferenças mais evidentes entre os modelos, especialmente na identificação da classe minoritária, representada pelos alunos que evadiram. *Random Forest* manteve valores elevados de precisão mesmo com aumento do recall.

A técnica *XGBoost* apresentou redução mais acentuada da precisão à medida que o recall aumentou. A Regressão Logística apresentou queda mais evidente da precisão, indicando maior dificuldade na identificação correta dos alunos em risco. Esse tipo de comportamento é esperado em bases desbalanceadas, nas quais há menor proporção de casos positivos (He; Garcia, 2009).

As curvas de calibração mostraram diferenças no comportamento das probabilidades previstas. A *Random Forest* apresentou maior aproximação à linha ideal nos níveis mais elevados de probabilidade.

A técnica *XGBoost* apresentou variações ao longo das faixas analisadas, com diferenças entre valores previstos e observados. A Regressão Logística apresentou maior afastamento da linha ideal, principalmente em níveis intermediários, indicando menor correspondência entre as probabilidades estimadas e os valores observados.

A análise de calibração é relevante quando as probabilidades são utilizadas como estimativas de risco, pois avalia o grau de alinhamento entre previsões e ocorrências reais (Niculescu-Mizil; Caruana, 2005).

A análise do ajuste do limiar de decisão mostrou que a utilização do valor padrão de 0,50 não foi a configuração mais adequada para todos os modelos. A redução do limiar resultou em aumento do *recall*, com maior identificação de alunos evadidos.

Na *Random Forest*, o ajuste permitiu alcançar *recall* máximo, com identificação completa dos casos de evasão.

Na técnica *XGBoost* e na Regressão Logística, o ajuste também aumentou o *recall* e o F1-score, embora com impacto na redução da precisão. A definição do limiar deve considerar o objetivo da análise, especialmente em problemas nos quais a identificação da classe de interesse é prioritária (Provost; Fawcett, 2013).

Esses resultados indicaram que os modelos apresentaram comportamentos distintos na identificação de alunos em risco de evasão. *Random Forest* apresentou melhor desempenho geral, considerando a capacidade de classificação e a qualidade das probabilidades previstas.

A técnica *XGBoost* apresentou bom desempenho na discriminação, porém com maior variação nas estimativas de risco. A Regressão Logística apresentou desempenho inferior, especialmente na identificação da classe de evasão.

Dessa forma, a avaliação conjunta das métricas e das curvas permitiu uma análise mais completa do desempenho dos modelos, evidenciando a importância de

considerar diferentes critérios na escolha da técnica mais adequada para o contexto educacional.

Com base nos resultados obtidos, a *Random Forest* foi selecionada como a técnica mais adequada para o contexto. A importância das variáveis, obtida a partir dessa técnica, indicou maior contribuição de atributos relacionados ao desempenho acadêmico e à frequência na predição do risco de evasão, mostrando que oscilações no desempenho e na frequência estão associadas ao aumento do risco de evasão.

Embora a *XGBoost* tenha apresentado desempenho competitivo após a otimização do limiar, com F1-score de 0,917 e ROC de 0,987, a escolha por *Random Forest* fundamenta-se em critérios que vão além do desempenho agregado, considerando especificamente o custo assimétrico dos erros de classificação no contexto educacional.

Em aplicações voltadas à identificação de risco de evasão, os erros do tipo falso negativo casos em que um aluno evadido não é identificado pelo modelo apresentam consequências pedagógicas e institucionais mais graves do que os erros do tipo falso positivo, que resultam em atenção desnecessária a um aluno que permaneceria no curso.

Enquanto o falso positivo representa um custo de intervenção redundante, o falso negativo representa a perda de uma oportunidade de atuação preventiva, potencialmente irreversível no contexto da trajetória acadêmica do aluno (Baker e Inventado, 2014).

Nesse sentido, a *Random Forest* apresentou, no limiar otimizado de 0,392, recall igual a 1,000, o que significa a identificação de todos os casos de evasão presentes nos dados de validação, com zeros falsos negativos.

Esse resultado, combinado com uma precisão de 0,871 e F1-*score* de 0,931, indica que o modelo atingiu o melhor equilíbrio entre sensibilidade à classe minoritária e controle sobre os falsos positivos entre os modelos avaliados.

Além disso, o *Random Forest* apresentou maior estabilidade ao longo das partições da validação cruzada e melhor calibração probabilística em comparação ao *XGBoost,* conforme evidenciado pelas curvas de calibração analisadas na seção anterior. Em contextos nos quais as probabilidades estimadas são utilizadas para orientar decisões de acompanhamento pedagógico, a qualidade da calibração é umcritério relevante, pois determina o grau de confiabilidade das estimativas de risco atribuídas a cada aluno (Niculescu-Mizil e Caruana, 2005).

A escolha pelo *Random Forest* é ainda coerente com a literatura de MDE, que aponta esse algoritmo como uma das técnicas com melhor desempenho em tarefas de classificação com bases desbalanceadas e múltiplas variáveis acadêmicas. Diante do conjunto de evidências apresentado, *Random Forest* constitui a técnica mais indicada para, por exemplo, monitorar risco de evasão no contexto da ETEC Raposo Tavares.

4.1.2 ***Interpretação e Avaliação do Conhecimento***

Os resultados obtidos mostram que a evasão apresenta forte associação com variáveis acadêmicas clássicas, especialmente frequência, número de faltas, desempenho médio e componentes curriculares críticos. Por exemplo, alunos com baixa frequência e maior número de faltas, associados ao acúmulo de disciplinas com desempenho insuficiente, apresentaram maior probabilidade de evasão.

Observou-se diferença entre alunos evadidos e não evadidos, uma vez que os evadidos apresentaram menor frequência, maior número de faltas e maior quantidade de disciplinas em situação crítica, indicando diferentes níveis de integração ao longo da trajetória acadêmica dos alunos.

A presença regular nas aulas não representa apenas cumprimento de carga horária, mas traduz participação ativa na formação, exposição contínua aos conteúdos e manutenção do vínculo institucional. A redução da frequência sugere enfraquecimento progressivo desse vínculo.

Além disso, o acúmulo médio de faltas entre evadidos foi aproximadamente três vezes superior ao observado entre os não evadidos. Esse dado aponta para um processo cumulativo de afastamento, no qual a ausência reiterada compromete a aprendizagem, amplia lacunas e aumenta a probabilidade de dificuldades acadêmicas subsequentes.

Os alunos evadidos também apresentaram maior número médio de disciplinas em situação crítica de 2,6 contra 0,85 entre os não evadidos. Esse indicador mostra que a evasão não se configura como evento abrupto, mas como resultado de um processo progressivo de fragilização do desempenho acadêmico.A existência de disciplinas em situação crítica funciona como marcador objetivo de risco. Quando associadas à baixa frequência, essas ocorrências compõem um padrão de desengajamento gradual.

À medida que o aluno acumula dificuldades em diferentes disciplinas, amplia-se a probabilidade de comprometimento da trajetória acadêmica, reduzindo a percepção de sucesso e aumentando a vulnerabilidade ao abandono. Tal padrão sugere que a evasão pode ser analisada como etapa final de uma trajetória marcada por sinais prévios mensuráveis, conforme apontou Romero e Ventura (2020).

Do ponto de vista institucional, isso indica que a evasão não ocorre de forma repentina, mas é precedida por evidências observáveis, como aumento do número de disciplinas em risco e redução da frequência, o que permite a implementação de mecanismos de monitoramento preventivo e intervenção antecipada, segundo o estudo de He e Garcia (2009).

Entretanto, um dos achados mais relevantes refere-se à identificação de um grupo de alunos que, mesmo apresentando desempenho satisfatório e frequência adequada, evadiram, conforme Baker e Inventado (2014) abordaram em seu estudo. A presença de 40 alunos com média de desempenho superior a 70 e frequência superior a 75% demonstra que a evasão não pode ser explicada

exclusivamente por variáveis acadêmicas.

Esse resultado desloca a análise para além da causalidade exclusivamente pedagógica. Se o desempenho e a frequência são adequados, a evasão não pode ser atribuída somente ao baixo desempenho acadêmico, o que indica a necessidade de interpretar o fenômeno sob uma perspectiva mais ampla, corroborando com o estudo de Tinto (2017).

A presença desse grupo revela que o vínculo com a instituição não se sustenta apenas por indicadores de desempenho. A permanência exige alinhamento entre fatores socioeconômicos, inserção no mercado de trabalho, desalinhamento vocacional ou fragilidade no sentimento de identificação e pertencimento institucional.

Alunos da educação profissional pública podem ter que conciliar a formação técnica com a inserção precoce no mercado de trabalho. Questões relacionadas à renda familiar, necessidade de contribuição financeira, dificuldades de transporte e redefinição de prioridades profissionais podem se sobrepor à trajetória acadêmica, mesmo quando o desempenho é satisfatório (Ramos, 2014; Inep, 2022).

A evasão, nesse sentido, não pode ser analisada apenas como consequência de baixo rendimento ou insuficiência acadêmica. Ela envolve processos relacionados ao sentido atribuído à experiência escolar, à construção de identidade profissional, à percepção de pertencimento e ao alinhamento entre expectativas

individuais e o projeto formativo, conforme apresentou Tinto (2017) e Baker e Inventado (2014) em seus estudos.

Isto implica reconhecer que os dados educacionais capturam manifestações observáveis do comportamento do aluno, mas não apreendem integralmente os significados que os alunos atribuem à sua trajetória. Assim, a MDE deve ser aplicada não para explicação total da evasão, mas como instrumento estratégico dentro de uma metodologia. Seu papel é identificar padrões, sinalizar riscos e fornecer base para recomendações.

Em síntese, os resultados indicam que a análise quantitativa permite identificar estruturas de risco, porém a compreensão da evasão exige considerar também fatores não diretamente observáveis nos dados acadêmicos. Diante dos resultados apresentados, verificou-se que a evasão não ocorre de maneira uniforme, mas assume diferentes configurações conforme a combinação de frequência, desempenho acadêmico e condições externas.

4.1.2.1 **Identificação de Perfis pela Equipe de Gestão**

No dia 24 de fevereiro de 2026, às 10h00 foi realizada uma reunião com a equipe de gestão escolar da ETEC Raposo Tavares, composta pela diretora, pela coordenação pedagógica e pela coordenação do curso, para apresentar os resultados obtidos com a aplicação da MDE.

Durante a discussão, a equipe diretiva comentou que alguns dos pontos apresentados já eram percebidos no cotidiano da escola. Professores e a equipe de gestão frequentemente observavam, por exemplo, que alunos com muitas faltas ou dificuldades em determinadas disciplinas tinham maior chance de abandonar o curso. No entanto, essas percepções surgiam principalmente da experiência diária da escola, ou seja, a partir da observação, sem um levantamento sistemático e comprovação com dados.

Nesse sentido, os resultados apresentados neste trabalho ajudaram a confirmar, com base em dados, hipóteses que já vinham sendo levantadas internamente pela escola.

A partir da reunião com a equipe, foram levantadas medidas de ação voltadas ao acompanhamento dos alunos. Entre elas, destacou-se o acompanhamento mais próximo por parte da orientação educacional, especialmente nos casos de aumento de faltas e queda no desempenho. Também foi mencionada a importância de monitorar a frequência de forma mais contínua, com ações de contato e orientação aos alunos.

Outra medida levantada foi o incentivo à participação dos alunos em atividades diversificadas, como projetos, práticas e ações pedagógicas que favoreçam o engajamento. Além disso, considerou-se a necessidade de acompanhamento mais atento de aspectos sociais que possam impactar a permanência dos alunos na escola.

As medidas discutidas resultaram na definição de ações a serem implementadas, indicando o início de encaminhamentos práticos com base nos resultados obtidos. Essas ações evidenciam o potencial dos dados para apoiar decisões no contexto da gestão escolar.

Por fim, a reunião confirmou a importância do uso de dados educacionais como apoio à tomada de decisão e da aplicação da MDE, contribuindo para o planejamento de estratégias mais alinhadas à realidade dos alunos.

A equipe de gestão com base nos resultados apresentados pela MDE e no conhecimento adquirido ao longo da sua atuação escolar identificou três tipos de perfis de alunos: Alto Desempenho Escolar; Médio Desempenho Escolar; Baixo Desempenho Escolar.

Descreve-se a seguir os três tipos de perfis identificados:

a) **Alto Desempenho Escolar**

Caracterizado como perfil formado por alunos focados que têm ótimo rendimento e alta frequência às aulas. O nível de participação também é alto, com excelente gestão do tempo e em geral nenhuma reprovação em disciplinas críticas. É o perfil de quem se dedica ao curso e busca a formação, considerado com baixíssimo risco de evasão.

b) **Médio Desempenho Escolar**

Caracterizado como perfil formado por alunos com médio rendimento que apresentam oscilações entre desinteresse e momentos de foco. Apresentam frequência regular nas aulas, com médio ou baixo nível de participação. Apresentam também problemas com gestão do tempo e reprovações em disciplinas críticas. Sãoalunos que necessitam de atenção para não migrarem para o baixo rendimento. É o perfil de quem, apesar dos problemas busca formação, considerado com médio risco de evasão.

c) **Baixo Desempenho Escolar.**

Caracterizado como perfil formado por alunos com baixo rendimento, desinteressados e com falta de foco. São alunos que faltam regularmente às aulas, com baixo nível de participação, com problemas de gestão do tempo e reprovações em disciplinas críticas. São alunos que necessitam de intervenção e apoio da equipe pedagógica da escola. É o perfil que pode estar condicionado a desigualdades sociais, falta de apoio familiar, além de desalinhamento vocacional ou fragilidade no sentimento de identificação e pertencimento institucional, considerado com altíssimo risco de evasão

Vale destacar que o fenômeno da evasão escolar não está intimamente ligado somente ao perfil de baixo desempenho escolar. Os três perfis não estão imunes ao fenômeno, já que fatores internos e ou externos que acompanham ou podem surgir de forma inesperada, durante a trajetória podem levar à evasão.

A identificação dos perfis mostra que a evasão assume formas diferentes ao longo do curso. Enquanto alguns casos revelam um enfraquecimento progressivo do vínculo com as atividades acadêmicas, outros indicam que a interrupção pode ocorrer mesmo sem sinais prévios de baixo desempenho.

Assim, a evasão não se apresenta como um evento isolado, mas como um processo que se constrói ao longo da trajetória acadêmica, influenciado por diferentes condições que atuam de forma isolada ou combinada, o que torna o assunto extremamente complexo.

Cabe, então, a ETEC considerar a aplicação de estudos como o desenvolvido neste trabalho para identificar e também prever a evasão escolar, ou seja, a MDE se tornar uma importante aliada para apoiar a gestão escolar no desenvolvimento de estratégias educacionais para reduzir o risco de evasão.

Nesse sentido, com base na análise dos resultados, além das medidas citadas acima, a equipe de gestão recomendou a institucionalização de um protocolo interno de monitoramento permanente com a aplicação da MDE para apoiar a gestão escolar, no qual esses indicadores sejam acompanhados ao longo do período letivo. Tal monitoramento deve estar articulado às instâncias pedagógicas, de modo que os dados produzidos se convertam em estratégias educacionais concretas de apoio ao aluno.

A recomendação do protocolo está alinhado ao estudo de Romero et al., (2020), que ao elencar exemplos de aplicação e tarefas relacionadas à MDE, destaca a de identificar e prever o comportamento dos alunos, visualizar a evolução do aprendizado dos alunos, detectar comportamentos e ou atitudes indesejáveis, como a baixa motivação e a tendência à evasão.

Neste sentido, destaca-se também o estudo de Goldschmidt et al. (2015) que apresenta diferentes grupos e exemplos de interesses relacionados para a aplicação da MDE, que está alinhado com a recomendação do protocolo, ao destacar o uso de ferramentas para análise do aprendizado e do

comportamento dos alunos; identificar estratégias de aprendizado, sugerir tarefas de reforço, identificar e tratar lacunas de aprendizado e identificar alunos que requerem apoio diferenciado.

É fundamental, contudo, que essa estratégia seja orientada por uma lógica preventiva e formativa, e não punitiva. O objetivo central não deve ser classificar ou rotular alunos como de risco, mas oferecer suporte oportuno, reforço acadêmico e acompanhamento individualizado sempre que forem identificados sinais iniciais de vulnerabilidade que podem levar à evasão.

Essa distinção é relevante para a gestão, pois indica que o fenômeno não é homogêneo e, portanto, não deve ser enfrentado por meio de estratégias uniformes. Quando a evasão está vinculada a baixo desempenho e ao acúmulo de disciplinas em situação crítica, as ações tendem a demandar reforço pedagógico estruturado, acompanhamento sistemático e intervenções voltadas à recuperação da aprendizagem.

Por outro lado, quando o aluno apresenta bom desempenho e frequência adequada, mas ainda assim interrompe sua trajetória, a interpretação exige deslocamento do foco estritamente acadêmico para dimensões relacionadas ao sentido atribuído ao curso, às condições de permanência e ao projeto de vida. Nesses casos, tornam-se recomenda-se estratégias de escuta qualificada, orientação educacional e acompanhamento socioeducacional.

Reconhecer a existência desses diferentes perfis permite à instituição evitar metodologias simplificadoras e construir respostas mais sensíveis à complexidade do fenômeno, articulando apoio pedagógico e atenção às dimensões subjetivas da experiência formativa.

A presença de alunos que evadem mesmo apresentando desempenho satisfatório mostra uma lacuna na análise institucional sobre os fatores que influenciam a decisão de desligamento.

Esse dado sugere que os registros acadêmicos, embora relevantes, não são suficientes para explicar integralmente o fenômeno. Nesse contexto, recomenda-se a adoção de um procedimento seja por entrevistas de desligamento, seja por meio de conversas preventivas realizadas antes da formalização da evasão, com o objetivo de ampliar a análise sobre as motivações envolvidas.

A organização e análise dessas informações, ainda que realizadas de maneira simples e integrada às práticas já existentes, poderão oferecer subsídios mais consistentes para a formulação de estratégias de permanência alinhadas à realidade concreta da escola e às múltiplas dimensões que atravessam a trajetória dos alunos.

Os resultados indicam ainda que parte dos desligamentos pode não estar associada a dificuldades acadêmicas, mas ao significado que o aluno atribui à formação e ao vínculo com o curso e com a escola.

Diante disso, recomenda-se o fortalecimento de estratégias de pertencimento e Identidade institucional como eixo das ações de permanência, o que pode ser concretizado por meio da ampliação do contato dos alunos com situações reais da área profissional, da realização de encontros periódicos com egressos e profissionais convidados e da valorização de trajetórias de sucesso vinculadas ao curso.

A escola pode ainda promover rodas de conversa sobre expectativas em relação à formação e ao mercado de trabalho, incentivar a participação em eventos técnicos e integrar relatos de experiências profissionais às disciplinas.

Essas ações, incorporadas ao planejamento pedagógico já existente, favorecem a conexão entre teoria e prática, fortalecem o vínculo com o curso e contribuem para a consolidação da identidade profissional e do sentimento de pertencimento institucional.

5 CONCLUSÃO

A evasão escolar aparece de forma recorrente em cursos técnicos de nível médio, sendo descrita como um fenômeno complexo, associado a diferentes fatores que influenciam a trajetória e produzem padrões variados, de acordo o contexto e o perfil do aluno, o que justifica a aplicação da MDE.

Nesse contexto, a evasão representa a interrupção da trajetória acadêmica, com reflexos na continuidade dos estudos, na inserção profissional e na melhoria da qualidade de vida do próprio aluno e da sua família.

Os procedimentos metodológicos adotados permitiram atender ao objetivo geral proposto, bem como aos objetivos específicos definidos ao longo do trabalho. A organização e o pré-processamento da base de dados possibilitaram a construção de uma base, que apesar de pequena foi adequada para a aplicação de três técnicas de MDE avaliadas com métricas, o que permitiu selecionar a técnica com o melhor desempenho, que por sua vez gerou resultados que possibilitaram à equipe de gestão escolar identificar três perfis de alunos, como resultado da extração de conhecimento.

Os resultados obtidos indicaram que variáveis relacionadas ao desempenho acadêmico e à frequência, observadas de maneira longitudinal, apresentam associação com o risco de evasão ao longo da trajetória escolar. Do ponto de vista aplicado, os achados deste trabalho indicam que dados institucionais disponíveis podem ser utilizados como subsídio para análises no âmbito da gestão escolar quando submetidos ao processo de KDD.

Tal utilização pode ser incorporada às rotinas administrativas existentes, respeitando os fluxos institucionais e os critérios de uso de dados adotados pela unidade escolar, sem pressupor a implantação de sistemas adicionais.

Os resultados devem ser compreendidos considerando o recorte adotado neste trabalho. A análise concentrou-se em uma única unidade escolar e em um curso específico, utilizando exclusivamente variáveis acadêmicas registradas no sistema institucional.

Outros fatores que podem se relacionar à evasão, como aspectos socioeconômicos, familiares ou psicossociais, não foram contemplados, apesar de levantados durante a reunião com a equipe de gestão da escola. Esse recorte, contudo, mostrou-se coerente com os objetivos propostos e com a disponibilidade de dados no contexto analisado. Considera-se que o objetivo geral foi alcançado porque os resultados obtidos com a aplicação da MDE possibilitaram a identificação

de três perfis pela equipe de gestão da ETEC. Outro ponto de destaque foi que os principais motivos desencadeadores de evasão também foram confirmados pela equipe, fato que, anteriormente à realização do estudo figurava como percepção ou observação.

A aprovação da equipe de gestão à aplicação da MDE e, consequentemente aos resultados obtidos com extração de conhecimento trouxe importância para o trabalho, devido a utilização de uma base de dados do mundo real e também da aplicação em um problema do mundo real.

Nesse sentido, além de outras medidas, recomendou-se a institucionalização de um protocolo interno de monitoramento permanente como apoio à gestão escolar, no qual esses indicadores sejam acompanhados ao longo do período letivo.

A resposta para a questão de pesquisa sobre “Como aplicar Mineração de Dados Educacionais para identificar perfis de alunos com risco de evasão do Curso Técnico em Química da ETEC Raposo Tavares, com base na análise de sua trajetória acadêmica, para apoiar a gestão escolar no desenvolvimento de estratégias educacionais?” está embasada nas cinco etapas de realização dos experimentos computacionais, que orientaram desde a coleta e o pré-processamento dos dados até a interpretação e análise do conhecimento descoberto, finalizando com a recomendação de um protocolo interno de monitoramento permanente, além de outras medidas.

As contribuições do trabalho têm vários desdobramentos quando se trata de apoio à gestão educacional para o desenvolvimento de estratégias educacionais ao identificar os perfis dos alunos com risco de evasão e tomar medidas que identifiquem e antecipem a evasão.

A contribuição para os alunos em não evadir do curso pode significar a conquista de uma vaga de emprego, de uma vaga na universidade, empreender, independência financeira e melhora da qualidade de vida de si próprio e dos familiares.

A contribuição para as empresas reside na possibilidade de contarem com candidatos preparados para atuarem em seus cargos. Para a sociedade ao se certificar de que os impostos coletados revertem na melhoria do ensino oferecido pela ETEC Raposo Tavares.

Para a pesquisa acadêmica, face a pequena quantidade de trabalhos que abordam o tema, tornar-se roteiro para a aplicação da MDE em outras ETECS, contribuindo para a descoberta de conhecimento em bases de dados que estão disponíveis, porém pouco acessadas para este fim.

Os resultados obtidos refletem as características do contexto analisado, considerando os dados acadêmicos disponíveis e as especificidades da unidade e do curso, podendo apresentar variações quando aplicados a outras ETECs ou contextos educacionais.

A técnica desenvolvida depende da qualidade e da consistência dos registros institucionais utilizados, podendo apresentar variações conforme a organização dos dados. Além disso, a análise concentrou-se em variáveis acadêmicas, não contemplando aspectos socioeconômicos e contextuais que também influenciam a evasão.

Por fim, o trabalho foi conduzido com uma base de dados de dimensão limitada, considerada como limitação, porém adequada ao contexto que foi a aplicação em um problema do mundo real. Dessa forma, os achados devem ser considerados como resultados situados, com possibilidade de adaptação a outras realidades mediante ajustes nas variáveis e nos dados considerados. Também pode ser considerada como limitação o fato de em função da necessidade de finalização do trabalho, informações sobre a aplicação das medidas discutidas e do monitoramento não estarem presentes no texto.

Como encaminhamentos para trabalhos futuros, sugere-se a ampliação do trabalho para outros cursos, unidades escolares ou períodos letivos, bem como a incorporação de novas variáveis que possam contribuir para uma compreensão mais ampla do fenômeno da evasão. Também se aponta a possibilidade de explorar abordagens complementares de análise, de modo a aprofundar a investigação dos fatores associados à trajetória acadêmica dos alunos ao longo do tempo.

Recomenda-se também, a aplicação da metodologia para o desenvolvimento de projetos em Data Mining, denominada Cross Industry Standard Process for Data Mining (CRISP-DM). A aplicação desta metodologia demandaria tempo além do prazo necessário para o término deste trabalho, o que inviabilizou a sua aplicação.

Os estudos apresentados neste trabalho não têm a pretensão de esgotar o assunto, pelo contrário, buscou-se contribuir para um assunto tão importante, como a evasão escolar.

REFERÊNCIAS

ABU-ZOHAIR, Ahmed. **Mineração de dados educacionais para apoio à gestão acadêmica: revisão de literatura**. Educação e Tecnologia. Belo Horizonte, 2025. Disponível em: https://doi.org/10.7769/gesec.v15i2.3555.. Acesso em: 21 mai. 2025.

ALSHANQITI, a.; NAMOUN, a.. Predicting student performance and its influential factors using hybrid regression and multi-label classification. **IEEE**, v. 8. 203844 p. Disponível em: https://doi.org/10.1109/ACCESS.2020.3036572. Acesso em: 8 mai. 2026.

ANDRELO, Pamela Ferreira Alves. **Mineração de dados educacionais na identificação do perfil dos egressos para apoio à gestão educacional de escola técnica pública**. São Paulo, 2022 Dissertação (Mestrado em Informática e Gestão do Conhecimento) - Universidade Nove de Julho, São Paulo, São Paulo, 2022.

Disponível em: http://bibliotecatede.uninove.br/handle/tede/3050. Acesso em: 30 mai. 2025.

ARAÚJO, Cleide L.. **Evasão escolar: causas e impactos da evasão escolar no Brasil e no mundo**. Revista de Gestão e Secretariado (GeSec). São Paulo.

Disponível em: https://doi.org/10.1016/j.sciaf.2024.e02157. Acesso em: 21 mai. 2025.

ARENDT, Hannah. **A condição humana**. 13 ed. Rio de Janeiro: Forense Universitária, 2010.

BAKER, r. s. j. d; INVENTADO, p. s.. **Educational data mining and learning analytics.**. https://doi.org/10.1007/978-1-4614-3305-7_4.. Disponível em: . Acesso em: 30 mai. 2025.

BAKER, Ryan S. J. d; INVENTADO, Paulo S. ; LARUSSON, Johann Ari. Educational data mining and learning analytics. **Learning Analytics: From Research to Practice**, New York, p. 61-75, 2014. Disponível em: https://doi.org/10.1007/978-1-4614-3305-7_4. Acesso em: 4 set. 2025.

BREIMAN, Leo. Random forests. Machine Learning, Dordrecht, v. 45, n. 1, p. 5-32, 2001. DOI: https://doi.org/10.1023/A:1010933404324. Acesso em: 30 jul. 2026.

CENTRO PAULA SOUZA. **Novo Sistema Acadêmico (NSA)**. São Paulo, 2025. Disponível em: https://nsa.cps.sp.gov.br/professores/professores_menu.aspx. Acesso em: 3 jul. 2025.

CHEN, Tianqi; CARLOS, Guestrin. XGBoost: a scalable tree boosting system. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.**, New York, p. 785-794, 2016. Disponível em: https://doi.org/10.1145/2939672.2939785. Acesso em: 3 set. 2025.COSTA, Evandro. Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. **Sociedade Brasileira de Computação**, Porto Alegre, v. 1, n. 1. 1–29 p, 2012. Disponível em: https://sol.sbc.org.br/livros/index.php/sbc/catalog/book/36. Acesso em: 25 jul. 2025.

ELBOUKNIFY, Imane. **AI-based identification and support of at-risk students: a case study of the Moroccan education system**. Scientific African. 2024. Disponível em: https://doi.org/10.1016/j.sciaf.2024.e02157. Acesso em: 20 mai. 2025.

ETEC RAPOSO TAVARES. **PLANO PLURIANUAL DE GESTÃO 2023-2027**. São Paulo, 2023. 120 p. Disponível em: https://etecraposotavares.cps.sp.gov.br/. Acesso em: 1 ago. 2025.

FARRELL, R. farrell. **Interactive interface to Data Envelopment Analysis modeling**. DOI: 10.1007/978-1-4419-7961-2. 2020. Disponível em: https://cran.r-project.org/. Acesso em: 2 jun. 2025.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. AI Magazine, Palo Alto, v. 17, n. 3, p. 37-54, 1996. DOI: https://doi.org/10.1609/aimag.v17i3.1230. Acesso em: 30 jul. 2026.

54. Disponível em: https://doi.org/10.1609/aimag.v17i3.1230. Acesso em: 4 set. 2025.

FEITOSA, Rômulo. Machine learning: uma aplicação para a prevenção da evasão escolar. **EDUCA – Revista Multidisciplinar em Educação,**, v. 7, n. 17. 901-919 p, 2020.

FERNANDES, m; BARROS, p. Fatores contextuais da evasão em escolas técnicas: um estudo de caso. **Revista de Educação e Contexto**, Goiás, v. 8, n. 2, p. 123-138.

GIL, Antônio Carlos. **Como elaborar projetos de pesquisa**. 5 ed. São Paulo: Atlas, 2008.

GOLDSCHIMIDT, Ronaldo; PASSOS, Emmanuel. Data Mining: Conceitos, técnicas, algoritmos, orientações e aplicações. **Elsevier**, Rio de Janeiro, v. 2, p. 233-255, 2015.GOODFELLOW, Ian; BENGIO, Yoshua; COURVILLE, Aaron. **Deep learning**. Cambridge: MIT Press, 2016. Disponível em: https://doi.org/10.5555/3086952. Acesso em: 6 set. 2025.

GÉRON, Aurélien. **Hands-on machine learning with Scikit-Learn, Keras and TensorFlow**. Sebastopol: O'Reilly Media, 2019. Disponível em: https://doi.org/10.1002/widm.1355. Acesso em: 21 mai. 2025.

HAN, Jiawei; KAMBER, Micheline; JIAN, Pei. Data Mining: Concepts and Techniques. **Morgan Kaufmann**, v. 3, 2012. Disponível em: https://doi.org/10.1016/C2009-0-61819-5. Acesso em: 7 set. 2025.

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. **Data mining**: Concepts and techniques. 3 ed. Burlington: Morgan Kaufmann, 2011.HAND, David J; MANNILA, Heikki; SMYTHV, Padhraic. **Principles of Data Mining**. Cambridge: MIT Press, 2001.

HASTIE, Trevor; TIBSHIRANI, Robert; FRIEDMAN, Jerome. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. **Springer**, New York, v. 2, 2009. Disponível em: DOI: https://doi.org/10.1007/978-0-387-84858-7. Acesso em: 22 jan. 2026.

HE, Haibo; GARCIA, Edwardo. Learning from imbalanced data. **IEEE Transactions on Knowledge and Data Engineering**, New York, v. 21, n. 9. 1263–1284 p, 2009. Disponível em: https://doi.org/10.1109/TKDE.2008.239. Acesso em: 4 set. 2025.

HOYOS, Wilson; CAICEDO-CASTRO, Ignacio. Ajustando modelos de mineração de dados para prever o desempenho acadêmico do ensino médio. **Data**, v. 9, n. 7. https://doi.org/10.3390/data9070086.

ET AL., Elbouknify. **AI-based identification and support of at-risk students: a case study of the Moroccan education system.**. Arxiv. 2025. 233 p. Disponível em: https://arxiv.org/abs/2504.07160. Acesso em: 30 mai. 2025.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA (INEP). **Censo Escolar da Educação Básica 2022: resumo técnico**. INEP. Brasília. Disponível em: https://www.gov.br/inep. Acesso em: 7 jul. 2025.

JACOMINI, Marcia Aparecida. Política educacional na rede estadual paulista e qualidade do ensino sob a Nova Gestão Pública, 1998 a 2018. **Education Policy Analysis Archives**, Arizona State University, v. 30, 2022. Disponível em: https://doi.org/10.14507/epaa.30.6465. Acesso em: 29 mai. 2025.

KHAIRY, Doaa. Previsão do desempenho do aluno em exames usando algoritmos de classificação de mineração de dados. **Education and Information Technologies**, New York, v. 29. 21621-21645 p. Disponível em: https://doi.org/10.1007/s10639-024-12619-w. Acesso em: 13 mai. 2025.

KHALIL, Hiba; EBNER, Martin. **Using learning analytics to improve the educational design of MOOCs**. International Journal of Education and Learning. 2022. 100-108 p. Disponível em: https://doi.org/10.31763/ijele.v4i2.641. Acesso em: 21 mai. 2025.

KINGMA, Diederik P.; BA, Jimmy. **Adam: a method for stochastic optimization**. arXiv. 2014. Disponível em: https://doi.org/10.48550/arXiv.1412.6980. Acesso em: 6 mai. 2025.

KIRKWOOD, Adrian; PRICE, Linda. Technology-enhanced learning and teaching in higher education: what is 'enhanced' and how do we know?. **Learning, Media and Technology,**, v. 39, n. 1, p. 6-36, 2014. Disponível em: DOI: https://doi.org/10.1080/17439884.2013.770404. Acesso em: 4 mar. 2026.

KLOFT, Marius. Predicting MOOC dropout over weeks using machine learning methods. **Workshop on Analysis of Large Scale Social Interaction in MOOCs**,2014. 60-65.

KLUYVER, Thomas. 20th International Conference on Electronic Publishing*. In:* JUPYTER NOTEBOOKS: A PUBLISHING FORMAT FOR REPRODUCIBLE COMPUTATIONAL WORKFLOWS, n. 20. 2016. Anais eletrônicos [...] Amsterdam: IOS Press. 87-90 p. Disponível em: https://ebooks.iospress.nl/volumearticle/43231. Acesso em: 20 mai. 2025.

KLUYVER, Thomas. **Jupyter Notebooks – a publishing format for reproducible computational workflows**. Positioning and Power in Academic Publishing: Players, Agents and Agendas. 2016. Disponível em: https://10.3233/978-1-61499-649-1-87. Acesso em: 20 abr. 2025.

KUHN, Max; JOHNSON, Kjell. Applied Predictive Modeling. **Springer**, New York, 2013. Disponível em: https://doi.org/10.1007/978-1-4614-6849-3. Acesso em: 6 set. 2025.

LIMA, Paulo A.. Gestão e inovação no ensino: desafios e perspectivas para as escolas técnicas públicas. **Educação e Tecnologia**, Belo Horizonte, v. 18, n. 2. 45–59 p, 2016. Disponível em: https://doi.org/10.1234/edutec.2016.01802. Acesso em: 21 mai. 2025.

MARCIA, Carletto. **Educação e tecnologia: um percurso interinstitucional**. Campos dos Goytacazes: Essentia Editora, 2011.

MARCONI, Marina de Andrade; LAKATOS, Eva Maria. **Fundamentos de metodologia científica**. 7 ed. São Paulo: Atlas, 2017.

MARIA, Psyridou. Machine learning predicts upper secondary education dropout as early as the end of primary school. **Scientific Reports**, Londres, v. 14, 2024. Disponível em: https://doi.org/10.1038/s41598-024-63629-0. Acesso em: 4 jul. 2025.

MELO, Eduardo. On the use of explainable artificial intelligence to evaluate school dropout. **IEEE Latin America Transactions**, v. 20, n. 9. 1584–1591 p. Disponível em: https://doi.org/10.1109/TLA.2022.9888863. Acesso em: 17 abr. 2025.

MELO, Elton *et al*. On the use of explainable artificial intelligence to evaluate school dropout. **Education Sciences**, Suíça, v. 12, n. 12. 845 p, 2022. Disponível em:https://doi.org/10.3390/educsci12120845. Acesso em: 21 mai. 2025.

MOURA, Andréia F. de. **Mineração de dados educacionais para apoio à gestão acadêmica na formulação de prognóstico de perfil de aluno ingressante em cursos superiores**. São Paulo, f. 142, 2023 Tese - Universidade Nove de Julho (uninove), São

Paulo. Disponível em: https://bibliotecatede.uninove.br/handle/tede/3236. Acesso em: 7 abr. 2025.

MUSSO, Mariel F; HERNÁNDEZ, Casimiro F. R.; CASCALLAR, Eduardo C.. Predicting key educational outcomes in academic trajectories: a machine-learning approach. **Higher Education**, Dordrecht, v. 80, n. 5. 875–894 p, 2020. Disponível em: https://doi.org/10.1007/s10734-020-00520-7. Acesso em: 21 mai. 2025.NGUYEN, a.; GARDNER, I.; SHERIDAN, d. Design principles for learning analytics information systems in higher education. **Information Systems Journal**, v. 31, n. 4. 1–31 p, 2021. Disponível em: DOI: https://doi.org/10.1111/isj.12315. Acesso em: 4 mar. 2026.

NGUYEN, Vinh; GARDNER, Michael; SHERIDAN, Kim. Proceedings of the ACM Conference on Learning@Scale. *In:* ACM CONFERENCE ON LEARNING@SCALE. 2018, Londres: ACM, 2018. 1,4 p. Disponível em: https://doi.org/10.1145/3277569. Acesso em: 25 mai. 2025.

NICULESCU-MIZIL, Alexandru; CARUANA, Rich. Predicting good probabilities with supervised learning.. **Proceedings of the 22nd International Conference on Machine Learning (ICML)**, New York. 625–632 p, 2005. Disponível em: https://doi.org/10.1145/1102351.1102430. Acesso em: 6 jun. 2025.

Organização para a Cooperação e Desenvolvimento Econômico. Education at a glance 2019: OECD indicators. **OECD Publishing**, Paris, 2019.

PAPAMITSIOU, Zacharoula; ECONOMIDES, Anastasios A.. **Learning analytics and educational data mining in practice: a systematic literature review of empirical evidence**. Journal of Educational Technology & Society. 2014. Disponível em: https://www.jstor.org/stable/jeductechsoci.17.4.49. Acesso em: 21 abr. 2025.

PAULA SOUZA, Centro. **ETEC Raposo Tavares: Relatório de Conclusão de Curso de Química. São Paulo: Centro Paula Souza**. Disponível em: . Acesso em: 1 jun. 2025.

PAULA SOUZA, Centro. Etec tem melhor curso técnico profissionalizante, diz Datafolha. **Centro Paula Souza**, São Paulo, 2025. Disponível em: https://www.cps.sp.gov.br/etec-tem-melhor-curso-tecnico-profissionalizante-diz-datafolha/.. Acesso em: 6 abr. 2025.

PAULA SOUZA, Centro. Plano Plurianual de Gestão 2024-2028. **PPG** , São Paulo, 2024. Disponível em: https://etecraposotavares.cps.sp.gov.br/escola/ppg/. Acesso em: 1 jun. 2025.

PEREIRA, David Eduardo. A comprehensive survey of argument mining in the educational domain: techniques, applications, and future directions. **WIREs Data Mining and Knowledge Discovery**, 2025. Disponível em: https://doi.org/10.1002/widm.70041. Acesso em: 20 jun. 2025.

PONTES, Gustavo. Predictive models for imbalanced data: a school dropout perspective. **Education Sciences, Basel**, v. 9, n. 4. 275 p, 2019. Disponível em: https://doi.org/10.3390/educsci9040275. Acesso em: 20 jun. 2025.

RAMOS, Marília Patta.. Métodos quantitativos e pesquisa em ciências sociais: lógica e utilidade do uso da quantificação nas explicações dos fenômenos sociais. **Mediações**, Londrina, v. 18. 55-65 p, 2014.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining and learning analytics: an updated survey. Wiley Interdisciplinary Reviews: Data Mining and**Knowledge Discovery**, Hoboken, v. 10, n. 3. e1355 p, 2020. Disponível em: DOI: 10.1002/widm.1355. Acesso em: 21 mai. 2025.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining and learning analytics: an updated survey. **WIREs Data Mining and Knowledge Discovery**, v. 10, n. 3, 2020. Disponível em: https://doi.org/10.1002/widm.1355. Acesso em: 13 ago. 2025.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics – Part C:**

Applications and Reviews, New York, v. 40, n. 6, p. 601-618, 2010. Disponível em: https://doi.org/10.1109/TSMCC.2010.2053532. Acesso em: 6 jun. 2025.

ROMERO, Cristóbal; VENTURA, Sebastián. Educational data mining: a review of the state of the art. **IEEE Transactions on Systems, Man, and Cybernetics, Part C**, New York, v. 40, n. 6. 601-618 p, 2010. 10.1109/TSMCC.2010.2053532.

RUSSELL, Stuart J.; NORVIG, Peter. Artificial intelligence: a modern approach. **Pearson**, Harlow, 2020.

SCHERER, Suely; BRITO, Gláucia. Integração de tecnologias digitais ao currículo: diálogos sobre desafios e dificuldades. **Educar em Revista**, Curitiba, v. 36, 2020. Disponível em: https://doi.org/10.1590/0104-4060.76252. Acesso em: 20 jul. 2025.

SILVA, Leandro S. Mineração de dados educacionais para predição de desempenho acadêmico: uma revisão sistemática da literatura. **Revista Brasileira de Informática na Educação**, Porto Alegre, v. 28. . 1-28 p, 2020. Disponível em: https://doi.org/10.5753/RBIE.2020.28.0.1. Acesso em: 20 abr. 2025.

SILVA, Thiago; ZHAO, Liang. Machine Learning in Complex Networks. **Springer**, 2016. Disponível em: https://doi.org/10.1007/978-3-319-17290-3. Acesso em: 20 jul. 2025.

SOKOLOVA, Marina; LAPALME, Guy. A systematic analysis of performance measures for classification tasks.. **Information Processing & Management**, Amsterdam, v. 45, n. 4, p. 427-437. Disponível em: https://doi.org/10.1016/j.ipm.2009.03.002 . Acesso em: 4 jul. 2025.

TINTO, Vicent. Through the eyes of students. **Journal of College Student Retention: Research, Theory & Practice**, v. 19. 254-269 p, 2017. Disponível em: DOI: https://doi.org/10.1177/1521025115621917. Acesso em: 4 mar. 2026.

VADAKARA, Jisha; KUMAR, Dhanasekaran Vimal. Aggrandized random forest to detect the credit card frauds. **Advances in Science, Technology and Engineering Systems Journal**, v. 4, n. 4, 2019. 121-127. Disponível em: https://doi.org/10.25046/aj040414. Acesso em: 28 jun. 2025.

VASWANI, Ashish. Attention is all you need*. In:* CONFERENCE ON NEURAL INFORMATION PROCESSING SYSTEMS (NEURIPS). 2017. 30 ed, Long Beach: Curran Associates. 5998,6008 p. Disponível em: https://arxiv.org/abs/1706.03762. Acesso em: 21 mai. 2025.WAHEED, Hajra; GUITIERREZ, Junaid. Predicting academic performance of students from VLE big data using deep learning models. **Computers in Human Behavior**, v. 104. 106189 p, 2020. Disponível em: https://doi.org/10.1016/j.chb.2019.106189. Acesso em: 21 mai. 2025.

WANLI, Xing. Dropout prediction in MOOCs: using deep learning for personalized intervention. **Journal of Educational Computing Research**, v. 57, n. 3. 547–570 p, 2019. Disponível em: https://doi.org/10.1177/0735633118757015. Acesso em: 7 jul. 2025.

WITTEN, Ian H; FRANK, Eibe; HALL, Mark. A. Data Mining: Practical Machine Learning Tools and Techniques. **Morgan Kaufmann**, Burlington, 2011. Disponível em: https://doi.org/10.1016/C2009-0-19715-5. Acesso em: 7 set. 2025.

GLOSSÁRIO

Acurácia - Métrica de avaliação de classificadores que expressa a proporção de classificações corretas (verdadeiros positivos e verdadeiros negativos) em relação ao total de instâncias avaliadas.

Aprendizado de Máquina (Machine Learning) - Subcampo da Inteligência Artificial que desenvolve algoritmos capazes de aprender padrões a partir de dados sem serem explicitamente programados para cada tarefa.

Árvore de Decisão - Técnica preditiva que representa um conjunto de regras de decisão em estrutura hierárquica ramificada. Cada nó interno representa um teste sobre um atributo; cada ramo representa o resultado do teste; e cada folha representa uma classe ou valor de saída. É a estrutura base dos algoritmos Random Forest e XGBoost.

AUC (Area Under the Curve) - Área sob a curva ROC, que varia de 0 a 1. Representa a probabilidade de que um modelo atribua um escore maior a uma instância positiva do que a uma negativa. Valores próximos de 1 indicam excelente capacidade discriminativa; valores próximos de 0,5 indicam desempenho equivalente ao acaso.

Bagging (Bootstrap Aggregating) - Técnica de ensemble learning em que múltiplos modelos são treinados independentemente sobre subconjuntos aleatórios dos dados originais, gerados por amostragem com reposição. As previsões são combinadas por votação, na classificação, ou por média, na regressão. É o princípio base do Random Forest.

Base de Dados Desbalanceada - Base de dados em que as classes da variável-alvo estão presentes em proporções muito diferentes. Em estudos de evasão escolar, a classe de evadidos tende a ser menor do que a de não evadidos, o que dificulta o aprendizado da classe minoritária.

Boosting - Técnica de ensemble learning em que modelos simples são treinados sequencialmente e cada novo modelo concentra-se nos erros cometidos pelos anteriores. É o princípio base do XGBoost.

Bootstrap - Técnica estatística de reamostragem com reposição, utilizada para gerar subconjuntos de treinamento no Random Forest. Cada árvore é treinada sobre uma amostra do conjunto original, o que amplia a diversidade entre as árvores.

Calibração de Probabilidades - Propriedade que descreve o grau de correspondência entre as probabilidades estimadas pelo modelo e as frequências reais dos eventos observados.

class_weight='balanced' - Parâmetro dos algoritmos scikit-learn que ajusta automaticamente os pesos das classes durante o treinamento, de forma inversamente proporcional à frequência de cada classe.

Classe Minoritária - Em problemas de classificação binária desbalanceada, corresponde à classe com menor número de exemplos. Neste trabalho, refere-se ao grupo de alunos que evadiram.

Classificação - Tarefa de Mineração de Dados que atribui rótulos a novas instâncias com base em padrões aprendidos a partir de dados rotulados.

CPS (Centro Paula Souza) - Autarquia do Governo do Estado de São Paulo responsável pela gestão das Escolas Técnicas Estaduais e Faculdades de Tecnologia.

Curva de Calibração - Representação gráfica que compara as probabilidades previstas por um modelo com as frequências reais observadas nos dados.

Curva Precision-Recall (AUPRC) - Gráfico que representa a relação entre precisão e recall ao longo de diferentes limiares de decisão. É especialmente informativo em bases desbalanceadas.

Curva ROC (Receiver Operating Characteristic) - Representação gráfica da relação entre a taxa de verdadeiros positivos e a taxa de falsos positivos para diferentes valores de limiar de decisão.

Desbalanceamento de Classes - Situação em que as classes da variável-alvo apresentam frequências muito distintas na base de dados.

ETEC (Escola Técnica Estadual) - Instituição pública de educação profissional de nível médio vinculada ao Centro Paula Souza.

Evasão Escolar - Interrupção definitiva da trajetória acadêmica do aluno, sem conclusão do curso. Nesta pesquisa, foi operacionalmente definida como o encerramento da matrícula por abandono, desistência formal ou ausência prolongada, excluídas as transferências regulares.

F1-score (Medida F1) - Métrica que representa a média harmônica entre precisão e recall. Fornece um único valor que equilibra o desempenho do modelo em relação a falsos positivos e falsos negativos.

Falso Negativo (FN) - Erro de classificação em que o modelo prevê a classe negativa para uma instância pertencente à classe positiva.

Falso Positivo (FP) - Erro de classificação em que o modelo prevê a classe positiva para uma instância pertencente à classe negativa.

Feature Engineering (Engenharia de Atributos) - Processo de criação, transformação ou seleção de variáveis preditoras a partir dos dados brutos, com o objetivo de melhorar o desempenho dos modelos.

Fold - Subconjunto gerado pela divisão da base de dados na validação cruzada K-Fold. Em cada iteração, um fold é usado para validação e os demais para treinamento.

Gini (Impureza de Gini) - Medida estatística utilizada como critério de divisão nas árvores de decisão do Random Forest.

GridSearchCV - Método da biblioteca scikit-learn utilizado para busca exaustiva de combinações de hiperparâmetros.

Hash Criptográfico SHA-256 - Função matemática que transforma dados de entrada em uma sequência de 256 bits de tamanho fixo, de forma irreversível e determinística.

Hiperparâmetro - Parâmetro de configuração de um algoritmo definido antes do processo de treinamento, como número de árvores, profundidade máxima e taxa de aprendizado.

Knowledge Discovery in Databases (KDD) - Processo sistemático e iterativo de identificação de padrões válidos, novos, potencialmente úteis e compreensíveis em bases de dados.

LGPD (Lei Geral de Proteção de Dados) - Lei nº 13.709/2018, que regula o tratamento de dados pessoais no Brasil.

Limiar de Decisão (Threshold) - Valor de probabilidade a partir do qual o modelo classifica uma instância como pertencente à classe positiva.

Matriz de Confusão - Tabela que organiza os resultados de um classificador ao comparar as classes reais com as previstas.

MDE (Mineração de Dados Educacionais) - Subárea da Mineração de Dados voltada ao ambiente educacional, com foco na identificação de padrões presentes em registros acadêmicos.

MDI (Mean Decrease in Impurity) - Métrica de importância de variáveis utilizada no Random Forest, baseada na redução média da impureza de Gini.

NSA (Novo Sistema Acadêmico) - Sistema de gestão acadêmica do Centro Paula Souza que centraliza registros das ETECs, como notas, frequência, histórico escolar e situação de matrícula.

Overfitting (Sobreajuste) - Fenômeno em que um modelo se ajusta excessivamente aos dados de treinamento e perde capacidade de generalização.

Perfil Acadêmico - Conjunto de características mensuráveis da trajetória escolar de um aluno, como desempenho, frequência e progressão entre módulos.

Pré-processamento de Dados - Fase do KDD que compreende o tratamento e a organização dos dados brutos para uso na modelagem.

Precisão (Precision) - Métrica que representa a proporção de instâncias classificadas como positivas que efetivamente pertencem à classe positiva.

Random Forest (Floresta Aleatória) - Algoritmo de aprendizado de máquina baseado em ensemble que combina múltiplas árvores de decisão treinadas sobre subconjuntos aleatórios dos dados e das variáveis.

Recall (Sensibilidade) - Métrica que representa a proporção de instâncias positivas corretamente identificadas em relação ao total de instâncias positivas.

Regressão Logística - Técnica estatística de classificação supervisionada que estima a probabilidade de ocorrência de um evento binário.

SimpleImputer - Classe da biblioteca scikit-learn utilizada para o tratamento de valores ausentes.

StandardScaler - Classe da biblioteca scikit-learn utilizada para padronizar variáveis numéricas.

StratifiedGroupKFold - Variante da validação cruzada K-Fold que combina estratificação das classes e agrupamento, mantendo os registros de um mesmo aluno no mesmo fold.

Trajetória Acadêmica - Conjunto de registros produzidos ao longo dos módulos do curso, com informações de desempenho, frequência e progressão escolar.

Validação Cruzada (Cross-Validation) - Técnica de avaliação que divide a base de dados em K subconjuntos e treina e avalia o modelo K vezes, de forma rotativa.

Variável-Alvo (Target Variable) - Variável que o modelo de classificação busca prever. Neste estudo, corresponde à condição de evasão ou não evasão.

Verdadeiro Negativo (VN) - Resultado correto em que o modelo prevê a classe negativa para uma instância que pertence à classe negativa.

Verdadeiro Positivo (VP) - Resultado correto em que o modelo prevê a classe positiva para uma instância que pertence à classe positiva.

XGBoost (Extreme Gradient Boosting) - Algoritmo baseado em gradient boosting que treina árvores de decisão de forma sequencial e incorpora um termo de regularização.

APÊNDICE A — MEMORANDO DE AUTORIZAÇÃO PARA ACESSO AOS DADOS DO NSA DA ETEC RAPOSO TAVARES

SÃO PAULO
GOVERNO DO ESTADO

Governo do Estado de São Paulo
Centro Paula Souza
225 - Etec de Raposo Tavares - Raposo Tavares/SP - Diretoria

MEMORANDO

Nº do Processo: 136.00066491/2025-09

Interessado: Gestão Pedagógica

Assunto: Autorização/Ciência para acesso e disponibilidade de dados do NSA da Etec Raposo Tavares

Prezado Professor Almério Melquíades de Araújo,

Venho por meio desse solicitar o acesso e disponibilidade dos dados do Sistema Acadêmico da Unidade Etec Raposo Tavares para fins de pesquisa do trabalho de mestrado da docente Fernanda Pereira Gomes intitulado: "Mineração de Dados Educacionais na Trajetória Acadêmica para Identificar o Perfil de Alunos de Escola Técnica Pública".

Conforme relatado pela docente e comprovado no projeto de pesquisa anexo o trabalho trata-se de uma pesquisa aplicada, com enfoque na utilização de técnicas de Mineração de Dados Educacionais (Educational Data Mining – EDM)sobre bases de dados acadêmicas já existentes, especificamente dados relacionados ao desempenho e à frequência dos alunos. Esse trabalho pode contribuir muito com coleta de indicadores e posteriormente em desenvolvimento de planejamento estratégico não só para a Unidade em questão mas para todos o Centro Paula Souza.

Me coloco a disposição para maiores esclarecimentos

São Paulo, na data da assinatura digital.

Taís Aparecida de Assis Garcia Moreira
Diretora de Escola Técnica

sei! assinatura eletrônica

Documento assinado eletronicamente por **Tais Aparecida de Assis Garcia Moreira**, **Diretor de Escola Técnica - Etec**, em 06/05/2025, às 11:06, conforme horário oficial de Brasília, com fundamento no Decreto Estadual nº 67.641, de 10 de abril de 2023.

APÊNDICE B — PARECER TÉCNICO AUTORIZANDO O USO DOS DADOS

Governo do Estado de São Paulo
Centro Paula Souza
Gestão Pedagógica

PARECER TÉCNICO

Nº do Processo: 136.00066491/2025-09

Interessado: Fernanda Pereira Gomes

Assunto: Solicitação para autorização de pesquisa de Mestrado

PARECER TÉCNICO 60/2025-GSE/GEPED

A área de Gestão Pedagógica (Geped), do Grupo de Supervisão Educacional, realizou análise da documentação referente ao pedido de autorização de pesquisa feito por **Fernanda Pereira Gomes**, aluna do Programa de Pós-Graduação em Informática e Gestão do Conhecimento (PPGICG), da Universidade Nove de Julho – Uninove, campus São Paulo.

De acordo com o projeto, a pesquisa intitulada "Mineração de dados educacionais na trajetória acadêmica para identificar o perfil de alunos de escola técnica pública" tem como objetivo geral: "aplicar técnicas de Mineração de Dados Educacionais (MDE) para analisar o perfil acadêmico dos alunos da ETEC Raposo Tavares, durante a sua trajetória acadêmica, a fim de apoiar a gestão escolar".

A metodologia utilizada será pesquisa documental consistindo na análise de dados dos alunos do curso técnico de Química, do 1º ao 4º módulo, matriculados no ano de 2024, disponíveis no Sistema Acadêmico NSA: Boletim escolar: para acesso às menções bimestrais por componente curricular; Histórico escolar: para verificar a trajetória de aprovação ou reprovação ao longo dos módulos; Registros de frequência: incluídos no boletim ou obtidos separadamente pelo NSA.

Por não se tratar de pesquisa com seres humanos, não houve necessidade de cadastro para aprovação na Plataforma Brasil, nem submissão ao Comitê de Ética em Pesquisa, uma vez que será realizada pesquisa quantitativa documental.