

UNIVERSIDADE FEDERAL DA FRONTEIRA SUL
CAMPUS CHAPECÓ
PROGRAMA DE MESTRADO PROFISSIONAL EM ADMINISTRAÇÃO PÚBLICA
PROFIAP

JEAN CARLO RODIO

**APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E MACHINE
LEARNING PARA PREDIÇÃO DA EVASÃO DE ESTUDANTES NOS CURSOS DE
GRADUAÇÃO DE UMA IFES**

CHAPECÓ

2026

JEAN CARLO RODIO

**APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E MACHINE
LEARNING PARA PREDIÇÃO DA EVASÃO DE ESTUDANTES NOS CURSOS DE
GRADUAÇÃO DE UMA IFES**

Dissertação apresentada ao Programa de Pós-Graduação em Administração Pública em Rede Nacional (PROFIAP), Mestrado Profissional em Administração Pública, da Universidade Federal da Fronteira Sul (UFFS), como requisito parcial para obtenção do título de Mestre em Administração Pública.

Orientador: Prof. Dr. Moacir Francisco Deimling

Coorientador: Prof. Dr. Giancarlo Dondoni Salton

CHAPECÓ

2026

Bibliotecas da Universidade Federal da Fronteira Sul - UFFS

Rodio, Jean Carlo

APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E MACHINE LEARNING PARA PREDIÇÃO DA EVASÃO DE ESTUDANTES NOS CURSOS DE GRADUAÇÃO DE UMA IFES / Jean Carlo Rodio. -- 2026.

126 f.

Orientador: Doutor Moacir Francisco Deimling

Co-orientador: Doutor Giancarlo Dondoni Salton

Dissertação (Mestrado) - Universidade Federal da Fronteira Sul, Programa de Pós-Graduação Profissional em Administração Pública em Rede Nacional, Chapecó, SC, 2026.

1. evasão acadêmica. 2. mineração de dados educacionais. 3. aprendizado de máquina. 4. modelos preditivos. 5. gestão acadêmica. I. Deimling, Moacir Francisco, orient. II. Salton, Giancarlo Dondoni, co-orient. III. Universidade Federal da Fronteira Sul. IV. Título.

Elaborada pelo sistema de Geração Automática de Ficha de Identificação da Obra pela UFFS com os dados fornecidos pelo(a) autor(a).

JEAN CARLO RODIO

TÍTULO: APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E MACHINE LEARNING PARA PREDIÇÃO DA EVASÃO DE ESTUDANTES NOS CURSOS DE GRADUAÇÃO DE UMA IFES

Dissertação apresentada ao Programa de Pós-Graduação em Administração Pública em Rede Nacional (PROFIAP), Mestrado Profissional em Administração Pública, da Universidade Federal da Fronteira Sul (UFFS), como requisito parcial para obtenção do título de Mestre em Administração Pública.

Este trabalho foi defendido e aprovado pela banca em 17/08/2026.

BANCA EXAMINADORA

Prof. Dr. Moacir Francisco Deimling – UFFS
Orientador

Prof. Dr. Giancarlo Dondoni Salton – UFFS
Coorientador

Prof. Dr. Abdinardo Moreira Barreto de Oliveira – UTFPR
Avaliador

Prof. Dr. Jacir Favretto – UnC
Avaliador

Dedico este trabalho aos meus pais, que, com amor, dedicação e sacrifício, tornaram possível a minha trajetória acadêmica. Cada etapa vencida carrega um pouco do esforço, da confiança e dos valores que recebi deles. Esta conquista também é deles.

AGRADECIMENTOS

Agradeço, primeiramente, à minha família, por todo o zelo, amor e dedicação que sempre tiveram comigo. Pelo apoio incondicional, pelos ensinamentos, pelos exemplos de honestidade e perseverança, e por nunca medirem esforços para que eu pudesse estudar e seguir em frente. Esta conquista também é fruto do cuidado e dos valores que recebi de vocês.

Aos meus amigos Ana Victória, Ivan, Marcel, Marco Antônio, Geovano e Edimar minha sincera gratidão pelo incentivo, pela escuta, pela parceria e pela presença nos momentos em que precisei de apoio. Cada palavra de encorajamento, cada conversa e cada gesto de amizade tornaram esta caminhada mais leve e possível.

Ao meu companheiro Jean C. Vitor, agradeço pelo apoio, pela paciência e pela presença constante ao longo desta caminhada. Sua compreensão nos momentos de ausência, seu incentivo nos períodos de dificuldade e sua parceria diária foram fundamentais para que eu pudesse seguir em frente com mais confiança e tranquilidade. Sou grato por ter ao meu lado alguém que compartilhou comigo os desafios, as renúncias e também as conquistas deste percurso.

À minha psicóloga Eluise, registro minha sincera gratidão pelo acolhimento, pela escuta atenta e pelo apoio ao longo desta caminhada. Sua presença foi importante para que eu pudesse atravessar os desafios deste período com mais clareza, equilíbrio e confiança. Agradeço por contribuir, com sensibilidade e profissionalismo, para o meu amadurecimento pessoal e para que eu encontrasse forças nos momentos em que o percurso parecia mais difícil.

Ao meu orientador e ao meu coorientador, agradeço pela orientação cuidadosa, pela paciência, pela disponibilidade e pelas contribuições essenciais ao desenvolvimento deste trabalho. Suas observações, sugestões e direcionamentos foram fundamentais para o amadurecimento da pesquisa e para a construção deste resultado.

Agradeço também aos professores do Programa de Pós-Graduação em Administração Pública em Rede Nacional - PROFIAP, pelos conhecimentos compartilhados e pelas reflexões proporcionadas ao longo do curso.

Aos colegas de curso, pelas trocas de experiências, debates e aprendizados construídos coletivamente ao longo dessa trajetória.

Por fim, agradeço a todos que, de alguma forma, contribuíram para a realização deste trabalho, seja com palavras de incentivo, apoio técnico ou simples gestos de confiança.

Assim como na música, a vida também é feita de pausas, ritmos, recomeços e momentos em que tudo finalmente encontra harmonia.

RESUMO

A evasão estudantil no ensino superior brasileiro constitui fenômeno de relevância social e econômica, pois representa desperdício de recursos públicos, interrupção da formação de capital humano e aprofundamento de desigualdades estruturais. No âmbito das instituições federais de ensino superior (IFES), o problema é agravado pela ausência, em grande parte dessas instituições, de mecanismos sistemáticos de identificação precoce de estudantes em risco de abandono. A Universidade Federal da Fronteira Sul (UFFS), cuja Taxa de Sucesso na Graduação esteve em torno de 33,11% entre 2015 e 2023, não dispõe de modelo preditivo institucional consolidado para essa finalidade. Nesse contexto, o presente estudo teve como objetivo desenvolver e avaliar modelos preditivos de evasão estudantil nos cursos de graduação da UFFS, *campus* Chapecó, por meio de técnicas de mineração de dados e aprendizado de máquina. Do ponto de vista metodológico, a pesquisa configura-se como aplicada, de natureza quantitativa e caráter descritivo, desenvolvida sob a modalidade de estudo de caso. O processo analítico seguiu a metodologia CRISP-DM. Os dados foram obtidos de quatro sistemas institucionais: SIGAA, SCI, SAS e *Pergamum*, resultando em um conjunto analítico com 235.632 registros referentes a 8.749 matrículas do período de 2014.1 a 2024.2. Foram treinados e comparados seis algoritmos de aprendizado supervisionado: Regressão Logística, Árvore de Decisão, *Random Forest*, XGBoost, SVM Linear e KNN. A validação adotou o protocolo StratifiedGroupKFold com cinco partições agrupadas por matrícula, assegurando que os registros pertencentes a uma mesma matrícula permanecessem exclusivamente no conjunto de treinamento ou no conjunto de teste durante cada partição, reduzindo o risco de vazamento de informações entre registros vinculados ao mesmo estudante. O modelo XGBoost apresentou o melhor desempenho, com AUC-ROC de 0,8878, acurácia de 0,8039 e F1-macro de 0,8037, seguido pela Regressão Logística (AUC-ROC = 0,8681) e pelo SVM Linear (AUC-ROC = 0,8680). A análise de importância das variáveis revelou que o desempenho acadêmico predomina na estrutura preditiva do modelo: a média final nas disciplinas respondeu por aproximadamente um terço do ganho informacional total, seguida por número de faltas e área de conhecimento. O uso do Restaurante Universitário nos primeiros semestres emergiu como preditor de engajamento institucional, superando variáveis sociodemográficas como gênero e etnia. Conclui-se que os modelos desenvolvidos demonstraram capacidade preditiva satisfatória e viabilidade de aplicação institucional. Como produto técnico-tecnológico (PTT), apresenta-se um relato técnico que sistematiza a experiência de aplicação dos modelos preditivos no contexto da UFFS *campus* Chapecó, e

propõe diretrizes para a implementação de um sistema de alerta precoce integrado aos sistemas institucionais, com potencial de replicação em outras IFES, contribuindo para a formulação de políticas de permanência estudantil baseadas em evidências.

Palavras-chave: evasão acadêmica; mineração de dados educacionais; aprendizado de máquina; modelos preditivos; gestão acadêmica.

ABSTRACT

Student dropout in Brazilian higher education is a phenomenon of considerable social and economic relevance, as it represents a waste of public resources, disrupts human capital development, and exacerbates structural inequalities. Within Brazilian Federal Institutions of Higher Education (Instituições Federais de Ensino Superior – IFES), this problem is intensified by the absence, in many institutions, of systematic mechanisms for the early identification of students at risk of dropping out. The Federal University of Fronteira Sul (Universidade Federal da Fronteira Sul – UFFS), whose Undergraduate Success Rate was approximately 33.11% between 2015 and 2023, does not have a consolidated institutional predictive model for this purpose. In this context, this study aimed to develop and evaluate predictive models of student dropout in undergraduate programs at the UFFS Chapecó Campus using data mining and machine learning techniques. Methodologically, this is an applied, quantitative, and descriptive study conducted as a case study. The analytical process followed the Cross-Industry Standard Process for Data Mining (CRISP-DM) methodology. Data were obtained from four institutional information systems—SIGAA, SCI, SAS, and Pergamum—resulting in an analytical dataset comprising 235,632 records related to 8,749 student enrollments from the first semester of 2014 to the second semester of 2024. Six supervised machine learning algorithms were trained and compared: Logistic Regression, Decision Tree, Random Forest, XGBoost, Linear Support Vector Machine (SVM), and K-Nearest Neighbors (KNN). Validation was performed using a five-fold StratifiedGroupKFold protocol, grouped by student enrollment, ensuring that records associated with the same enrollment were included exclusively in either the training set or the test set within each fold. This procedure reduced the risk of information leakage between records associated with the same student. The XGBoost model achieved the best performance, with an AUC-ROC of 0.8878, accuracy of 0.8039, and macro-averaged F1 score of 0.8037, followed by Logistic Regression (AUC-ROC = 0.8681) and Linear SVM (AUC-ROC = 0.8680). The feature importance analysis revealed that academic performance predominated in the model's predictive structure: students' final course grades accounted for approximately one-third of the total information gain, followed by the number of absences and field of study. The use of the University Restaurant during the initial semesters emerged as a predictor of institutional engagement, surpassing sociodemographic variables such as gender and ethnicity. It is concluded that the developed models demonstrated satisfactory predictive capacity and institutional applicability. As a technical and technological product, this study presents a

technical report that systematizes the experience of applying predictive models within the context of the UFFS Chapecó Campus and proposes guidelines for implementing an early warning system integrated with the institution's information systems. The proposed approach has the potential to be replicated in other Brazilian Federal Institutions of Higher Education, contributing to the development of evidence-based student retention policies.

Keywords: academic dropout; educational data mining; machine learning; predictive models; academic management.

LISTA DE FIGURAS

Figura 1 – Fases do modelo de referência CRISP-DM	54
Figura 2 - Comparativo de desempenho dos 6 modelos	81
Figura 3 - <i>Ranking</i> por AUC-ROC	82
Figura 4 - Matriz de confusão do modelo XGBoost.....	84
Figura 5 - Importância das variáveis pelo XGBoost	90
Figura 6 - Coeficientes de correlação de Pearson entre as variáveis	91

LISTA DE QUADROS

Quadro 1 - Estrutura da matriz de confusão para classificação binária	44
Quadro 2 - Rastreabilidade das fases do CRISP-DM na estrutura da dissertação.....	62
Quadro 3 - Relatório de qualidade de dados: variáveis categóricas	65
Quadro 4 - Plano de qualidade de dados	69
Quadro 5 - Variável alvo	73
Quadro 6 - Área de conhecimento	76
Quadro 7 - Região de nascimento	77
Quadro 8 - Resumo das transformações de codificação e engenharia de variáveis	78
Quadro 9 - Configuração dos modelos	80

LISTA DE TABELAS

Tabela 1 - Relatório de qualidade de dados: variáveis contínuas	64
Tabela 2 - Relatório de qualidade de dados: variáveis do Pergamum	67
Tabela 3 - Desempenho dos modelos preditivos	81

LISTA DE ABREVIATURAS E SIGLAS

ABT	<i>Analytics Base Table</i> (Tabela Base Analítica)
AD	Árvore de Decisão
AUC	<i>Area Under the Curve</i> (Área sob a Curva)
AUC-ROC	<i>Area Under the Receiver Operating Characteristic Curve</i> (Área sob a Curva Característica de Operação do Receptor)
COVID-19	<i>Coronavirus Disease 2019</i> (doença pelo coronavírus 2019)
CPF	Cadastro de Pessoas Físicas
CRISP-DM	<i>Cross-Industry Standard Process for Data Mining</i> (Processo Padrão Intersetorial para Mineração de Dados)
DQP	<i>Data Quality Plan</i> (Plano de Qualidade de Dados)
DQR	<i>Data Quality Report</i> (Relatório de Qualidade de Dados)
DRA	Diretoria de Registro Acadêmico
EDA	<i>Exploratory Data Analysis</i> (Análise Exploratória de Dados)
EDM	<i>Educational Data Mining</i> (Mineração de Dados Educacionais)
ENEM	Exame Nacional do Ensino Médio
F1-score	Medida F1, calculada pela média harmônica entre precisão e <i>recall</i>
FIES	Fundo de Financiamento Estudantil
FN	Falso Negativo
FP	Falso Positivo
GBM	<i>Gradient Boosting Machine</i> (Máquina de Reforço por Gradiente)
IA	Inteligência Artificial
IDH	Índice de Desenvolvimento Humano
IES	Instituição de Ensino Superior
IFES	Instituição Federal de Ensino Superior
INEP	Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira
IPEA	Instituto de Pesquisa Econômica Aplicada
IQR	<i>Interquartile Range</i> (Intervalo Interquartil)
ISL	<i>Integral Solutions Ltd.</i>
KDD Dados)	<i>Knowledge Discovery in Databases</i> (Descoberta de Conhecimento em Bases de
KNN	<i>K-Nearest Neighbors</i> (K-Vizinhos Mais Próximos)
LGPD	Lei Geral de Proteção de Dados Pessoais

LIME	<i>Local Interpretable Model-Agnostic Explanations</i> (Explicações Locais Interpretáveis e Agnósticas ao Modelo)
MEC	Ministério da Educação
OECD	<i>Organisation for Economic Co-operation and Development</i> (Organização para a Cooperação e Desenvolvimento Econômico)
PARFOR	Plano Nacional de Formação de Professores da Educação Básica
PEC-G	Programa de Estudantes-Convênio de Graduação
PIN	Programa de Acesso e Permanência dos Povos Indígenas
PNAES	Política Nacional de Assistência Estudantil
PROFIAP	Mestrado Profissional em Administração Pública em Rede Nacional
Prouni	Programa Universidade para Todos
PTT	Produto Técnico Tecnológico
REUNI Federais	Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais
RF	<i>Random Forest</i> (Floresta Aleatória)
RL	Regressão Logística
ROC	<i>Receiver Operating Characteristic</i> (Característica de Operação do Receptor)
RU	Restaurante Universitário
SAEB	Sistema de Avaliação da Educação Básica
SAS	Sistema de Análise Socioeconômica
SCI	Sistema de Credenciamento Institucional
SEMESP	Instituto Semesp
SEMMA	<i>Sample, Explore, Modify, Model, Assess</i> (Amostragem, Exploração, Modificação, Modelagem e Avaliação)
SETI	Secretaria Especial de Tecnologia e Informação
SHAP	<i>SHapley Additive exPlanations</i> (Explicações Aditivas de Shapley)
SIG	<i>Special Interest Group</i> (Grupo de Interesse Especial)
SIGAA	Sistema Integrado de Gestão de Atividades Acadêmicas
SISU	Sistema de Seleção Unificada
SVM	<i>Support Vector Machine</i> (Máquina de Vetores de Suporte)
TAE	Técnico-Administrativo em Educação
TCU	Tribunal de Contas da União
TFP	Taxa de Falsos Positivos
TSG	Taxa de Sucesso na Graduação

TVP	Taxa de Verdadeiros Positivos
UFFS	Universidade Federal da Fronteira Sul
UF	Unidade da Federação
VN	Verdadeiro Negativo
VP	Verdadeiro Positivo
XGBoost	<i>Extreme Gradient Boosting</i> (Reforço Extremo por Gradiente)

SUMÁRIO

1 INTRODUÇÃO.....	19
1.1 PROBLEMA DE PESQUISA.....	22
1.2 OBJETIVOS.....	23
1.2.1 Objetivo geral.....	23
1.2.2 Objetivos específicos.....	23
1.3 JUSTIFICATIVA E RELEVÂNCIA DA PESQUISA.....	24
2 FUNDAMENTAÇÃO TEÓRICA.....	28
2.1 EVASÃO ESCOLAR NO ENSINO SUPERIOR.....	28
2.1.1 Dimensão social da evasão.....	30
2.1.2 Dimensão institucional da evasão.....	32
2.1.3 Dimensão sistêmica e estrutural da evasão.....	34
2.2 APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E <i>MACHINE LEARNING</i> NA EDUCAÇÃO.....	37
2.3 MODELOS PREDITIVOS E SUA APLICAÇÃO À EVASÃO NO ENSINO SUPERIOR.....	39
2.3.1 Modelos lineares e baseados em regras.....	40
2.3.2 Métodos de conjunto (<i>Ensemble methods</i>).....	40
2.3.3 Modelos de alta complexidade.....	41
2.3.4 Métodos baseados em instâncias.....	42
2.3.5 Métricas de avaliação de modelos preditivos.....	43
2.3.5.1 Matriz de confusão.....	43
2.3.5.2 Acurácia.....	44
2.3.5.3 Precisão (<i>Precision</i>).....	45
2.3.5.4 Sensibilidade (<i>Recall</i>).....	45
2.3.5.5 F1-score.....	46
2.3.5.6 AUC-ROC.....	47
2.4 <i>CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING</i> (CRISP-DM).....	48
2.4.1 Compreensão do negócio (<i>Business understanding</i>).....	49
2.4.2 Compreensão dos dados (<i>Data understanding</i>).....	50
2.4.3 Preparação dos dados (<i>Data preparation</i>).....	51
2.4.4 Modelagem (<i>Modeling</i>).....	52
2.4.5 Avaliação (<i>Evaluation</i>).....	53
2.4.6 Implantação (<i>Deployment</i>).....	53
3 PROCEDIMENTOS METODOLÓGICOS.....	56
3.1 CARACTERIZAÇÃO DA PESQUISA.....	56
3.2 COLETA, ANÁLISE E INTERPRETAÇÃO DE DADOS.....	57
3.2.1 Aspectos éticos e proteção dos dados.....	58
3.2.2 Metodologia CRISP-DM.....	58
4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS.....	63
4.1 RELATÓRIO DE QUALIDADE DE DADOS.....	63

4.1.1 Variáveis contínuas.....	64
4.1.2 Variáveis categóricas.....	64
4.1.3 Colunas de uso do RU e sistema de análise socioeconômica (SAS).....	66
4.1.4 Colunas de empréstimos da biblioteca.....	66
4.2 IDENTIFICAÇÃO DE PROBLEMAS DE QUALIDADE.....	67
4.2.1 Valores ausentes.....	67
4.2.2 Cardinalidade irregular.....	68
4.2.3 <i>Outliers</i>	68
4.3 PLANO DE QUALIDADE DE DADOS.....	69
4.3.1 Tratamento dos dados.....	70
4.3.1.1 Remoção de variáveis sem variação.....	70
4.3.1.2 Consolidação das variáveis do sistema de análise socioeconômica.....	70
4.3.1.3 Tratamento de <i>outliers</i>	70
4.3.1.4 Imputação de valores ausentes.....	71
4.3.1.5 Consolidação das variáveis do RU.....	71
4.3.1.6 Consolidação das variáveis da biblioteca.....	72
4.4 CODIFICAÇÃO E ENGENHARIA DE VARIÁVEIS.....	72
4.4.1 Definição da variável alvo.....	72
4.5 ESTRATÉGIAS DE CODIFICAÇÃO ADOTADAS.....	73
4.5.1 Remoção de colunas de alta cardinalidade.....	74
4.5.2 <i>Binary encoding</i>	74
4.5.3 <i>Label encoding</i>	74
4.6 NOVAS VARIÁVEIS CRIADAS POR ENGENHARIA.....	75
4.6.1 Área de conhecimento: variável <i>area_disciplina</i>	76
4.6.2 Região geográfica de nascimento.....	76
4.6.3 Engenharia de variável - idade no ingresso.....	77
4.6.4 Visão geral das transformações.....	78
4.7 RESULTADOS DA MODELAGEM PREDITIVA.....	79
4.7.1 Resultados dos modelos.....	80
4.7.2 Matriz de confusão do modelo XGBoost.....	82
4.8 ANÁLISE POR MODELO.....	84
4.9 ANÁLISE INTEGRADA DOS ACHADOS DA MODELAGEM PREDITIVA.....	86
4.9.1 Interpretação das variáveis preditivas do XGBoost.....	87
4.9.2 Contribuição original do estudo.....	91
4.9.3 Implicações para a gestão acadêmica da UFFS.....	92
5 SEÇÃO PROPOSITIVA (PTT).....	95
6 CONSIDERAÇÕES FINAIS.....	99
REFERÊNCIAS.....	103
APÊNDICE A - PROTOCOLO DE IMPLEMENTAÇÃO DE UM SISTEMA DE ALERTA PRECOCE DE EVASÃO ACADÊMICA NA UFFS CAMPUS CHAPECÓ.	111

1 INTRODUÇÃO

O ensino superior ocupa posição estratégica no projeto de desenvolvimento do Brasil. A formação de profissionais qualificados é condição necessária para o aumento da produtividade, a geração de inovação tecnológica, a redução das desigualdades sociais e o fortalecimento da capacidade competitiva do país no cenário global. Ao longo das últimas décadas, estudos de diversas áreas do conhecimento demonstraram que países com maiores taxas de escolarização de nível superior apresentam indicadores mais robustos de crescimento econômico, desenvolvimento humano e coesão social (OECD, 2023). No Brasil, essa relação é ainda mais significativa diante das históricas assimetrias regionais e socioeconômicas que marcam a estrutura social do país.

Reconhecendo esse papel estratégico, o Estado brasileiro promoveu, especialmente a partir dos anos 2000, uma expansão sem precedentes da oferta de educação superior. No âmbito público, o Programa de Apoio a Planos de Reestruturação e Expansão das Universidades Federais (REUNI), instituído em 2007, viabilizou a criação de novas universidades federais, a interiorização dos *campi* existentes e a ampliação expressiva do número de vagas em cursos de graduação. Entre 2003 e 2014, o número de universidades federais no Brasil passou de 45 para 63 instituições, evidenciando o processo de expansão da oferta pública de ensino superior no país, com abertura de novos *campi* em municípios que nunca haviam contado com presença federal de ensino superior (INEP, 2023). Paralelamente, programas como o Programa Universidade para Todos (ProUni) e o Fundo de Financiamento Estudantil (FIES) ampliaram o acesso nas instituições privadas, incorporando ao sistema universitário estudantes das camadas populares que, até então, encontravam barreiras econômicas intransponíveis para ingressar nesse nível de formação.

Nesse contexto de expansão, foi criada em 2009 a Universidade Federal da Fronteira Sul (UFFS), com *campi* nos estados do Rio Grande do Sul, Santa Catarina e Paraná. Concebida como uma universidade popular e voltada ao desenvolvimento da Mesorregião Grande Fronteira do Mercosul, a UFFS nasce com a missão explícita de democratizar o acesso ao ensino superior em uma região historicamente marcada pela predominância da agricultura familiar e pela baixa escolarização de nível superior. Seu perfil de ingressantes é, em grande medida, composto por estudantes oriundos de escolas públicas, filhos de agricultores familiares e trabalhadores rurais, muitos dos quais são a primeira geração de suas famílias a acessar uma universidade federal (UFFS, 2025).

Entretanto, a expansão do acesso ao ensino superior brasileiro não foi acompanhada, na mesma proporção, pela garantia de permanência e conclusão dos cursos. O aumento do número de vagas e de estudantes matriculados revelou, com maior nitidez, um problema estrutural que desafia o sistema educacional há décadas: a evasão. Nem todas as vagas criadas são efetivamente preenchidas e, entre os estudantes que ingressam, uma parcela significativa não chega a concluir o curso em que se matriculou. Esse fenômeno representa, simultaneamente, o desperdício de recursos públicos investidos na formação acadêmica, a frustração de projetos de vida individuais e o não cumprimento da missão social das instituições de ensino superior.

Os dados disponíveis confirmam a relevância do problema. O Mapa do Ensino Superior no Brasil 2024 mostra que, no ciclo 2018-2022, a taxa de desistência acumulada no ensino superior brasileiro alcançou 57,2% no conjunto das instituições, chegando a 52,6% nos cursos presenciais e a 64,0% na modalidade a distância (INSTITUTO SEMESP, 2024). A edição de 2025 reforça a persistência do problema ao indicar, para o ciclo 2019-2023, taxa de desistência acumulada de 58,3% no conjunto do sistema, com 54,2% nos cursos presenciais e 63,7% na EAD. Mesmo na rede pública, os índices permanecem preocupantes, com desistência acumulada de 41,9% no total, sendo 41,6% nos cursos presenciais e 46,6% na EAD (INSTITUTO SEMESP, 2025). Gomes e Hirata (2022) estimam que cerca de 60% dos estudantes que ingressaram no ensino superior em 2010 não o concluíram em até oito anos, dado que evidencia que a evasão não é um evento pontual, mas um processo que se estende ao longo de toda a trajetória acadêmica.

A evasão representa, portanto, mais do que o abandono individual de uma trajetória acadêmica. Ela implica o desperdício de investimento público per capita realizado em cada estudante, a limitação na produção de capital humano qualificado e o distanciamento entre as metas de acesso e as de formação efetiva. O Tribunal de Contas da União reconhece esse impacto ao recomendar, por meio do Acórdão 1868/2024, que as instituições federais de ensino superior elaborem políticas estruturadas de prevenção à evasão, fundamentadas em evidências e sustentadas por sistemas de monitoramento contínuo dos indicadores acadêmicos (TCU, 2024).

A literatura científica evidencia que a evasão é um fenômeno multicausal, resultante da articulação dinâmica de fatores acadêmicos, socioeconômicos, pessoais e institucionais. O modelo de integração estudantil proposto por Tinto (1975) constitui o principal referencial teórico para compreender esse processo: segundo o autor, a decisão de abandonar o curso é influenciada pelo grau de integração do estudante à vida acadêmica e social da instituição,

sendo que quanto menor essa integração, maior a probabilidade de evasão. Estudos recentes corroboram e ampliam esse modelo. Krause Bierhalz e Bruno (2025) argumentam que a decisão de evadir-se não se restringe ao espaço universitário, sendo frequentemente determinada por condições externas à vida acadêmica, como as exigências do mercado de trabalho e as responsabilidades familiares. Baggi e Lopes (2011) reforçam que a análise do abandono precisa considerar circunstâncias anteriores à vida universitária, incluindo o percurso escolar prévio e as expectativas construídas antes do ingresso. Barlem et al. (2012) identificam que fatores como imaturidade na escolha profissional, dificuldades de adaptação ao ambiente universitário, problemas financeiros e conflitos entre vida acadêmica e vida laboral figuram entre as principais razões relatadas pelos estudantes que abandonam seus cursos.

No contexto específico da UFFS, a evasão discente assume características que reforçam a necessidade de acompanhamento institucional. O Relatório de Auditoria Interna da instituição identificou a evasão discente como tema de alto risco institucional, com reflexos diretos sobre a eficácia, a eficiência e a efetividade das políticas de graduação. O documento registra que a média histórica da Taxa de Sucesso na Graduação (TSG) entre 2015 e 2023 foi de apenas 33,11%, valor que indica que menos de um terço dos estudantes concluem o curso no tempo previsto. O mesmo relatório aponta que a UFFS ainda não dispõe de política institucional consolidada nem de sistema formal e contínuo de monitoramento da evasão, o que limita a capacidade da instituição de identificar precocemente estudantes em risco e de estruturar intervenções preventivas tempestivas. Entre os fatores associados à evasão na UFFS, destacam-se a dificuldade de conciliar trabalho e estudo, as dificuldades financeiras, os problemas pessoais ou familiares, as dificuldades de organização do tempo, a não identificação com o curso e as dificuldades de acompanhamento das atividades acadêmicas (UFFS, 2025).

Esses elementos demonstram que a evasão, na UFFS, não pode ser compreendida apenas como um fenômeno individual, mas como um desafio institucional que demanda mecanismos mais sistemáticos de acompanhamento, análise e apoio à tomada de decisão. A identificação precoce dos estudantes com maior probabilidade de abandono é condição necessária para que a instituição possa intervir de forma preventiva, antes que o processo de desengajamento se consolide e o abandono se concretize.

Diante desse quadro, a Mineração de Dados Educacionais (*Educational Data Mining - EDM*) e o Aprendizado de Máquina emergem como estratégias analíticas com potencial comprovado de transformar a forma como as instituições de ensino superior lidam com a

evasão. Ao contrário das abordagens reativas, que identificam o problema apenas após o abandono, os modelos preditivos permitem antecipar o risco com base em dados históricos, identificando padrões que, embora imperceptíveis à análise humana direta, são estatisticamente associados ao abandono futuro (Baker; Yacef, 2009). Algoritmos como florestas aleatórias, árvores de decisão, regressão logística, máquinas de vetores de suporte e métodos de *gradient boosting* possibilitam processar grandes volumes de dados institucionais, e construir modelos capazes de estimar, para cada estudante, a probabilidade de evasão em semestres subsequentes (Kelleher; Mac Namee; D'Arcy, 2015).

Pesquisas recentes desenvolvidas no Brasil e no exterior demonstram que técnicas de EDM permitem identificar estudantes em situação de vulnerabilidade acadêmica com elevada precisão, viabilizando a atuação preventiva das instituições antes que o abandono se concretize. Bitencourt, Silva e Xavier (2021) e Soares e Carvalho (2023) evidenciam resultados expressivos na predição de evasão utilizando dados de sistemas acadêmicos de universidades brasileiras. Romero e Ventura (2020), em revisão abrangente da literatura internacional, confirmam que a predição de evasão constitui uma das aplicações mais consolidadas e impactantes da EDM no ensino superior, com crescente sofisticação metodológica e ampliação do escopo de variáveis analisadas. Kalita et al. (2025) e Quimiz-Moreira et al. (2025) avançam nessa agenda ao incorporar técnicas de explicabilidade algorítmica que permitem não apenas prever, mas também interpretar os fatores preditivos, tornando os modelos mais acionáveis pelos gestores acadêmicos.

A UFFS dispõe de um conjunto diversificado de dados institucionais que ainda não foi explorado de forma integrada para fins preditivos. A análise conjunta dessas informações representa uma oportunidade concreta de identificar precocemente estudantes em risco de evasão e apoiar a gestão acadêmica na adoção de intervenções preventivas.

1.1 PROBLEMA DE PESQUISA

A evasão no ensino superior brasileiro é reconhecida como um dos maiores desafios para as políticas públicas educacionais, pois implica desperdício de recursos, descontinuidade na formação de capital humano e aprofundamento das desigualdades sociais (Zago, 2006; Saccaro; França; Jacinto, 2019). Dados do Instituto de Pesquisa Econômica Aplicada (IPEA) mostram que, entre os estudantes que ingressaram em 2010, aproximadamente 60% não concluíram seus cursos em até oito anos, evidenciando a magnitude do problema no sistema federal de ensino (Gomes; Hirata, 2022).

Diversos fatores contribuem para a evasão, incluindo condições socioeconômicas, dificuldades de adaptação acadêmica, escolhas profissionais mal definidas e limitações institucionais (Baggi; Lopes, 2011; Barlem et al., 2012). Frente a esse cenário multifatorial, cresce o interesse pelo uso de metodologias inovadoras que permitam a identificação precoce de estudantes em risco, favorecendo ações preventivas. Pesquisas brasileiras recentes confirmam o potencial dessas técnicas: Bitencourt, Silva e Xavier (2021) demonstraram a eficácia da IA na previsão de evasão universitária, enquanto Soares e Carvalho (2023), aplicaram algoritmos de aprendizado de máquina em bases institucionais, alcançando alta precisão na classificação de estudantes propensos ao abandono.

Diante desse cenário, surge o seguinte problema de pesquisa: Como a aplicação de técnicas de Mineração de Dados e Aprendizado de Máquina sobre dados institucionais integrados pode subsidiar a identificação precoce de estudantes em risco de evasão nos cursos de graduação da Universidade Federal da Fronteira Sul, *campus* Chapecó?

1.2 OBJETIVOS

1.2.1 Objetivo geral

Aplicar técnicas de Mineração de Dados e Aprendizado de Máquina sobre dados institucionais integrados para a predição do risco de evasão discente nos cursos de graduação da Universidade Federal da Fronteira Sul, *campus* Chapecó.

1.2.2 Objetivos específicos

- a) Caracterizar o fenômeno da evasão no ensino superior público federal, com ênfase nos fatores acadêmicos, socioeconômicos e institucionais associados à permanência discente;
- b) Analisar os dados acadêmicos e demográficos disponíveis, realizando o tratamento necessário à sua utilização em algoritmos de aprendizado de máquina;
- c) Desenvolver um modelo preditivo, com base em técnicas de aprendizado supervisionado, para estimar a probabilidade de evasão discente;
- d) Validar o desempenho do modelo preditivo desenvolvido, verificando sua aplicabilidade como ferramenta de apoio à gestão acadêmica em uma IFES.
- e) Propor o Produto Técnico-Tecnológico decorrente do modelo validado.

1.3 JUSTIFICATIVA E RELEVÂNCIA DA PESQUISA

A justificativa desta pesquisa apoia-se em três dimensões centrais: social, acadêmica e institucional.

No âmbito social, a evasão no ensino superior representa não apenas o abandono de uma trajetória individual, mas também a perpetuação das desigualdades de acesso (Zago, 2006). Dessa forma, garantir a permanência estudantil significa ampliar oportunidades para populações vulneráveis. E, reduzir a evasão, contribui para formar profissionais capazes de atender às demandas regionais e nacionais, o que reforça a relevância direta desta investigação para a inclusão e a mobilidade social (Silva et al., 2022).

Nessa mesma direção, destaca-se o impacto regional da instituição estudada. A Universidade Federal da Fronteira Sul foi criada com a missão de atuar em uma região historicamente marcada por desigualdades socioeconômicas e pela carência de mão de obra qualificada, buscando consolidar-se como agente de integração e desenvolvimento territorial (UFFS, 2019). Pesquisas evidenciam que a presença da instituição tem contribuído para dinamizar o desenvolvimento local, ao ampliar o acesso ao ensino superior e formar profissionais em áreas estratégicas (Krajevski, 2021). Nesse sentido, a experiência da UFFS dialoga com análises nacionais que relacionam a interiorização das universidades públicas à redução de assimetrias regionais e ao fortalecimento do capital humano (Macedo; Monteiro Neto; Vieira, 2022).

Do ponto de vista acadêmico, embora haja uma longa trajetória de estudos conceituais sobre evasão, o desenvolvimento de protótipos operacionais ainda é relativamente recente considerando sua relevância. Revisões sistemáticas recentes destacam que a aplicação de modelos preditivos representa uma evolução substancial no combate ao abandono (Alves et al., 2024). Estudos empíricos demonstram o uso de algoritmos como regressão logística, redes neurais e algoritmos genéticos na construção de modelos com alto grau de acurácia para identificar estudantes em risco (Morais; Souza, 2024). Outros, ainda, propõem *dashboards* interativos e modelos preditivos em tempo real que auxiliam a gestão universitária, permitindo intervenções rápidas (Ladeira; Lima; Krohling, 2022).

A contribuição teórica desta pesquisa situa-se na articulação entre o campo dos estudos sobre evasão no ensino superior e o campo da Mineração de Dados Educacionais e do Aprendizado de Máquina, com foco em um contexto institucional: uma universidade federal multicampi do interior do Brasil, com perfil socioeconômico predominantemente vulnerável. A literatura clássica e contemporânea demonstra que a evasão é um fenômeno multifatorial,

influenciado por variáveis acadêmicas, institucionais, socioeconômicas e individuais, o que exige modelos analíticos capazes de lidar com elevada complexidade relacional (Saccaro; França; Jacinto, 2019; Nierotka; Salata; Klitzke Martins, 2023). Ao propor a utilização de técnicas de Mineração de Dados Educacionais e do Aprendizado de Máquina para a predição da evasão em cursos de graduação de uma IFES, esta pesquisa contribui teoricamente ao aproximar a tradição explicativa dos estudos sobre evasão das abordagens preditivas orientadas por dados, ampliando a compreensão sobre como diferentes dimensões do fenômeno podem ser operacionalizadas analiticamente em contexto institucional concreto (Bitencourt; Silva; Xavier, 2021; Vaarma; Li, 2024).

Em segundo lugar, o estudo dialoga com achados de revisão sistemática sobre predição de evasão no ensino superior, a qual evidencia a recorrência da comparação entre diferentes técnicas de aprendizado de máquina e o uso de métricas como acurácia, sensibilidade/*recall*, precisão, F1-score e AUC para avaliação dos modelos (Silva; Roman, 2021). A presente pesquisa busca avançar em relação a esse enfoque predominantemente técnico ao traduzir os resultados dos modelos em um protocolo institucional aplicável à gestão acadêmica. A revisão de Quimiz-Moreira et al. (2025) reforça esse diagnóstico ao mostrar que a literatura acumulou um grande volume de fatores e modelos preditivos, mas ainda apresenta desenvolvimento mais limitado no campo da explicabilidade e praticamente inexistente no uso de modelos de simulação para apoiar decisões universitárias. Nesse cenário, a presente pesquisa oferece contribuição teórica ao não tratar a predição apenas como exercício computacional, mas como instrumento analítico vinculado à gestão acadêmica, ao discutir como a modelagem preditiva pode ser incorporada à interpretação institucional da evasão em uma universidade pública federal (Quimiz-Moreira et al., 2025; Wegner; Alves Júnior; Fernandes, 2024).

Uma terceira contribuição teórica decorre do recorte institucional adotado. Estudos sobre evasão em instituições públicas brasileiras têm demonstrado que a permanência e a conclusão dependem de interações complexas entre características dos estudantes e condições institucionais, inclusive em universidades federais interiorizadas, como a UFFS, cuja missão social está ligada à ampliação do acesso e à redução de desigualdades regionais (Nierotka; Salata; Klitzke Martins, 2023). Ao desenvolver um modelo voltado a uma IFES, esta pesquisa contribui para a literatura ao deslocar o foco de estudos excessivamente genéricos ou centrados em contextos isolados para uma análise situada, capaz de dialogar com especificidades da educação superior pública brasileira. Teoricamente, isso é relevante porque permite examinar a evasão não apenas como desfecho estatístico, mas como processo

institucionalmente condicionado, fortalecendo o debate sobre a relação entre políticas de permanência, produção de indicadores e capacidade preditiva dos modelos em contextos públicos (Nierotka; Salata; Klitzke Martins, 2023; Saccaro; França; Jacinto, 2019).

Além disso, a pesquisa contribui para o avanço teórico ao incorporar a discussão sobre explicabilidade e uso responsável de dados educacionais. A literatura de *learning analytics* e de IA aplicada à educação já reconhece que modelos preditivos não devem ser avaliados apenas por sua performance, mas também por sua inteligibilidade, transparência e aderência ética, especialmente quando seus resultados podem influenciar intervenções institucionais sobre trajetórias estudantis (Slade; Prinsloo, 2013). Assim, ao discutir a aplicabilidade de modelos preditivos para apoio à gestão acadêmica em uma IFES, esta dissertação amplia a literatura ao sustentar que a utilidade teórica de modelos preditivos depende não apenas de prever corretamente, mas de tornar compreensíveis os fatores mais associados ao risco de evasão, de modo compatível com a realidade institucional e com princípios de uso responsável da informação (Slade; Prinsloo, 2013; Quimiz-Moreira et al., 2025).

Por fim, no campo teórico, esta pesquisa contribui ao reforçar uma agenda de investigação que combina análise longitudinal, dados institucionais e predição orientada à intervenção precoce. Estudos recentes mostram que a importância relativa das variáveis associadas à evasão pode variar ao longo do tempo, o que desafia abordagens estáticas e reforça a necessidade de modelos capazes de captar mudanças no risco de abandono conforme a trajetória acadêmica evolui (Vaarma; Li, 2024). Da mesma forma, pesquisas em *learning analytics* apontam que a identificação antecipada de estudantes em risco é especialmente relevante quando os dados disponíveis são integrados de forma consistente e interpretados à luz do contexto educacional em que são produzidos (Queiroga et al., 2020). Nesse sentido, o presente estudo avança a literatura ao propor uma leitura da evasão que combina capacidade preditiva, contextualização institucional e potencial de uso para gestão, contribuindo para consolidar o entendimento de que técnicas de mineração de dados e *machine learning*, no campo educacional, devem ser compreendidas como recurso analítico para qualificar a interpretação da permanência estudantil em instituições públicas de ensino superior, e não apenas como mecanismo automatizado de classificação (Bitencourt; Silva; Xavier, 2021; Vaarma; Li, 2024; Queiroga et al., 2020).

No âmbito institucional, a relevância se manifesta ao oferecer à Universidade Federal da Fronteira Sul (UFFS) um instrumento de análise preditiva para o monitoramento do risco de evasão discente. A administração universitária carece de mecanismos que permitam agir de forma preventiva, e um modelo baseado em técnicas de Mineração de Dados e Aprendizado

de Máquina possibilita identificar estudantes em risco antes da concretização do abandono. A adoção de medidas antecipadas favorece a redução da evasão e otimiza a alocação de recursos institucionais, alinhando-se às metas do Plano Nacional de Educação (BRASIL, 2014; BRASIL, 2024a) e às recomendações da Auditoria Interna da própria instituição, que identificou a evasão como tema de maior grau de risco institucional (UFFS, 2025).

Nesse mesmo eixo, destaca-se o alinhamento com os desafios contemporâneos da gestão pública. A sociedade demanda eficiência, transparência e inovação na aplicação de recursos, e o uso de técnicas de mineração de dados e *machine learning* em contextos educacionais responde a essa necessidade ao propor soluções fundamentadas em evidências. Nesse sentido, a pesquisa dialoga com a agenda de transformação digital do setor público, conforme estabelecido pela Lei de Governo Digital e pela Estratégia Federal de Governo Digital 2024-2027 (Brasil, 2021; Brasil, 2024b).

2 FUNDAMENTAÇÃO TEÓRICA

Esta fundamentação organiza-se em quatro eixos. No Eixo 1 (2.1), apresenta-se a evasão no ensino superior como fenômeno multifatorial, com ênfase em definições, determinantes e dimensões pessoal, institucional e sistêmica/estrutural. No Eixo 2 (2.2), discutem-se a Mineração de Dados e o *Machine Learning* na educação e a Mineração de Dados Educacionais (EDM) como campo de suporte analítico e de gestão, contemplando potencialidades e limites. O Eixo 3 (2.3) aprofunda os modelos preditivos empregados para identificar risco de evasão, destacando critérios de interpretabilidade e desempenho e também as métricas avaliativas utilizadas nos modelos. Por fim, o Eixo 4 (2.4) apresenta o CRISP-DM como referência metodológica que estrutura o ciclo do projeto, da compreensão do negócio à implantação, assegurando coerência entre o problema público, de permanência estudantil, e as soluções analíticas.

2.1 EVASÃO ESCOLAR NO ENSINO SUPERIOR

A evasão no ensino superior constitui fenômeno complexo, que envolve aspectos acadêmicos, administrativos e financeiros e que afeta estudantes, instituições, sistema educacional e sociedade como um todo. Lobo (2012) já alertava, há mais de uma década, que se tratava de problema de gestão institucional cujo enfrentamento não poderia se restringir à decisão individual do estudante nem a questões estritamente financeiras. A autora observava, à época, a escassez de estudos sistemáticos sobre o tema no Brasil, quadro que se modificou substancialmente desde então, com a consolidação da Mineração de Dados Educacionais como campo de pesquisa e o crescente interesse acadêmico pela predição da evasão a partir de dados institucionais (Romero; Ventura, 2020; Kalita et al., 2025; Quimiz-Moreira et al., 2025). Apesar desse avanço, compreender as causas da evasão e formular intervenções eficazes ainda exige análise multidimensional, articulando dimensões acadêmicas, sociodemográficas, institucionais e gerenciais.

Do ponto de vista de Fritsch, Rocha e Vitelli (2015), os autores caracterizam a evasão como a desistência do estudante por qualquer motivo, exceto a conclusão ou diplomação, configurando-se como processo de exclusão determinado por fatores e variáveis internas e externas às instituições de ensino. Os autores observam que, mesmo com a expansão do ensino superior brasileiro, a relação entre ingressantes e concluintes mantém-se discrepante, com reflexo tanto no ensino público quanto no privado.

No cenário internacional, tanto os Estados Unidos quanto países da União Europeia enfrentam desafios relacionados às taxas de evasão no ensino superior, ainda que motivados por fatores distintos. Nos Estados Unidos, o fenômeno é fortemente associado ao endividamento estudantil causado pelo elevado custo das mensalidades, o que leva muitos alunos a abandonar os cursos mesmo diante de políticas de crédito educacional. Já na União Europeia, políticas integradas de inclusão, apoio e monitoramento, incluindo sistemas de alerta precoce, têm obtido maior sucesso em reduzir a evasão, embora as taxas permaneçam elevadas em cursos das áreas técnicas e exatas. Esses dados reforçam que a evasão é um fenômeno global, multifatorial e persistente (OECD, 2023).

Outro aspecto relevante refere-se à diferença dos índices de evasão entre áreas de conhecimento, pois cursos das ciências exatas apresentam taxas mais elevadas quando comparados com outras áreas de conhecimento, observando, portanto, uma conexão na problemática ensino/aprendizagem. Assim, percebe-se uma falta de ações de enfrentamento voltadas à qualidade do ensino, com foco nas dificuldades de aprendizagem e no tratamento adequado das defasagens da formação básica, levando em conta o cuidado para evitar a sobrecarga de conteúdos (Garcia; Gomes, 2022).

Conforme Cunha e Morosini (2014), nem sempre seria possível identificar claramente todas as causas e motivações do abandono, mas é possível observar padrões que se repetem. Esses fatores podem ser agrupados em três dimensões: (1) pessoais/sociais, como escolhas equivocadas de carreira, ausência de orientação vocacional, expectativas irreais ou dificuldades de adaptação ao curso; (2) institucionais, relacionadas à localização da instituição, limitações estruturais ou falta de vínculos afetivos com o ambiente acadêmico; e (3) gerais, vinculadas às lacunas na formação básica, dificuldades de acompanhamento das disciplinas, desmotivação, baixa autoestima e limitações financeiras. Em diálogo com essa tipologia, outros estudos também mostram que a evasão pode ser compreendida a partir de dimensões acadêmicas, institucionais e sociodemográficas, ainda que utilizem classificações distintas. Silva Filho et al. (2007), ao analisarem dados do ensino superior brasileiro, indicam que características do curso e da instituição influenciam os índices de evasão, observando, por exemplo, associação entre menor demanda pelos cursos e maiores taxas de abandono. Andriola (2009), com base na percepção de docentes e coordenadores, destaca a importância do apoio pedagógico, da orientação acadêmica e do fortalecimento de ações institucionais de acompanhamento discente no enfrentamento da evasão. Em estudos quantitativos mais recentes, Nierotka, Bonamino e Carrasqueira (2023) verificam, com base em análise longitudinal, que variáveis como desempenho acadêmico, apoio social, tendem a reduzir a

probabilidade de evasão, ao passo que desigualdades sociodemográficas também afetam esse risco. De modo semelhante, Saccaro, França e Jacinto (2019), por meio de análise de sobrevivência, identificam que fatores como sexo, idade e apoio financeiro influenciam o tempo de permanência no ensino superior. Assim, a literatura indica que essas dimensões não apenas descrevem o fenômeno da evasão, mas produzem efeitos mensuráveis sobre a probabilidade de abandono, sendo já examinadas quantitativamente em diferentes contextos institucionais.

2.1.1 Dimensão social da evasão

Cunha e Morosini (2014) destacam que a evasão universitária não tem como exclusiva característica a responsabilidade dos estudantes, isso devido a estar ligada a uma série de fatores institucionais que ampliam o efeito do fenômeno. Muitas vezes, os alunos que abandonam seus cursos vêm de situações sociais mais vulneráveis, cujas trajetórias educacionais são marcadas por deficiências na educação básica e, a falta de preparação adequada para enfrentar o ensino superior, agrava ainda mais a situação, comprometendo a autoestima e a motivação dos estudantes.

Entre os principais fatores associados à evasão, a literatura destaca a inadequação curricular, a baixa identificação com o curso escolhido, o ingresso em curso que não constitui a primeira opção do estudante, as dificuldades financeiras, os problemas de adaptação ao ambiente acadêmico, a distância geográfica da família e a necessidade de conciliar estudo e trabalho (Cunha; Morosini, 2014; Gilioli, 2016). Tais fatores, contudo, apresentam diferentes possibilidades de mensuração. Algumas dimensões podem ser operacionalizadas quantitativamente a partir de dados institucionais, como curso de ingresso, modalidade de acesso, município de origem, apoio financeiro e situação ocupacional, quando disponível (Gilioli, 2016; Saccaro; França; Jacinto, 2019). Outras dimensões, de caráter mais subjetivo, como identificação com o curso e adaptação à vida universitária, tendem a requerer indicadores indiretos ou instrumentos complementares de coleta, uma vez que não costumam estar registradas de forma direta nas bases administrativas (Cunha; Morosini, 2014; Nierotka; Bonamino; Carrasqueira, 2023). Cursos como as licenciaturas e aqueles das áreas de ciências exatas e biológicas apresentam taxas de evasão particularmente elevadas, chegando a superar 50% em alguns casos. A insuficiência de políticas institucionais de apoio, como assistência estudantil e o acompanhamento pedagógico, também contribui para o problema (Gilioli, 2016).

Neste cenário, não é apenas uma falha individual do estudante, mas um reflexo de uma estrutura educacional que não está suficientemente preparada para atender às necessidades dos alunos, especialmente àqueles que enfrentam dificuldades econômicas. Esse é um fator bastante agravante, já que o impacto da carência de recursos financeiros, leva muitos alunos a desistirem de seus cursos por falta de opções viáveis para seguir com sua formação. Desse modo, nota-se que as instituições de ensino superior têm um papel importante na mitigação desse problema e há uma necessidade de assumir a responsabilidade por suas políticas e práticas que impactam diretamente a permanência dos estudantes.

Outro aspecto central no debate sobre evasão no ensino superior é sua estreita relação com as desigualdades sociais. A literatura tem demonstrado que o abandono não pode ser explicado apenas por dificuldades acadêmicas imediatas, pois está também associado às condições sociais desiguais que antecedem o ingresso na universidade e continuam a produzir efeitos ao longo da trajetória estudantil (Paula; Picanço, 2024; Nierotka; Bonamino; Carrasqueira, 2023). Nesse sentido, a evasão deve ser compreendida não apenas como um evento individual ou circunstancial, mas como expressão de processos mais amplos de estratificação social que influenciam o acesso, a permanência e a conclusão no ensino superior (Knop; Collares, 2019; Paula; Picanço, 2024).

Estudos recentes indicam, de forma quantitativa, que a origem social permanece estatisticamente associada ao risco de evasão mesmo após o ingresso no ensino superior. Paula e Picanço (2024), ao analisarem dados pareados do Exame Nacional do Ensino Médio e do Censo da Educação Superior em um grupo de ingressantes de 2014 acompanhada até 2019, estimaram, por meio de modelos de regressão logística multinível, razões de chance de evasão do sistema de ensino superior de 1,17 (IC95%: 1,15-1,19) para estudantes com pais que concluíram, no máximo, o ensino médio, e de 1,22 (IC95%: 1,20-1,24) para aqueles com pais até o ensino fundamental, ambas em comparação a estudantes cujos pais possuíam ensino superior ($p < 0,001$). Os autores também identificaram efeito expressivo da renda familiar, com aumento progressivo das chances de evasão à medida que diminui a faixa de renda, chegando a 1,63 (IC95%: 1,60-1,66) para estudantes com renda familiar de até um salário mínimo, em comparação aos com mais de cinco salários mínimos. Esses achados reforçam a compreensão de que a democratização do ingresso, por si só, não elimina as desigualdades de permanência.

A renda familiar também exerce papel relevante nesse processo. Knop e Collares (2019), ao investigarem a influência da origem social na probabilidade de conclusão dos diferentes cursos de ensino superior, demonstram que a renda familiar e a escolaridade

parental afetam as chances de conclusão, ainda que de forma diferenciada conforme o tipo de curso e o setor institucional. Em termos analíticos, isso significa que estudantes de famílias com menos recursos econômicos e culturais tendem a enfrentar condições menos favoráveis para permanecer até a diplomação, o que, em sentido inverso, está relacionado a maiores riscos de evasão. Assim, a evasão não decorre apenas de dificuldades de aprendizagem ou adaptação acadêmica, mas também de restrições materiais e simbólicas que limitam a permanência estudantil ao longo do curso (Paula; Picanço, 2024; Knop; Collares, 2019).

No caso das instituições públicas, essa relação também aparece em estudos longitudinais. Nierotka, Bonamino e Carrasqueira (2023), ao acompanharem um grupo de estudantes de ensino superior público, verificaram que fatores acadêmicos e sociais interagem na definição das trajetórias de permanência e evasão. Os autores identificaram, entre outros aspectos, que estudantes com melhor desempenho acadêmico e maior apoio social tendem a apresentar menores chances de evasão, enquanto desigualdades sociodemográficas permanecem associadas ao abandono. Isso sugere que a evasão deve ser analisada como fenômeno multidimensional, no qual vulnerabilidades econômicas, trajetória escolar prévia e condições institucionais atuam de forma combinada.

Nessa perspectiva, compreender e enfrentar a evasão implica reconhecer que ela reflete desigualdades históricas que atravessam o sistema educacional brasileiro. A literatura mostra que estudantes oriundos de grupos socialmente menos favorecidos não apenas acessam o ensino superior em condições desiguais, mas também permanecem mais expostos a obstáculos financeiros, acadêmicos e institucionais que ampliam o risco de desligamento (Paula; Picanço, 2024; Nierotka; Bonamino; Carrasqueira, 2023). Por essa razão, combater a evasão significa também enfrentar desigualdades estruturais, o que exige articulação entre políticas de assistência estudantil, apoio pedagógico, inclusão social e estratégias institucionais de acompanhamento acadêmico (Gilioli, 2016; Nierotka; Bonamino; Carrasqueira, 2023). Dessa forma, a permanência estudantil não pode ser tratada apenas como questão de desempenho individual, mas como responsabilidade institucional e política vinculada à promoção de maior equidade no ensino superior brasileiro (Gilioli, 2016; Paula; Picanço, 2024).

2.1.2 Dimensão institucional da evasão

O fenômeno da evasão no ensino superior não pode ser tratado como algo inevitável ou natural, pois isso impede que as instituições compreendam plenamente a dimensão do

problema, por isso, é necessário adotar estratégias e medidas de apoio aos alunos, como a implementação de programas de assistência estudantil, o investimento na formação contínua dos professores e a criação de um ambiente acadêmico que favoreça a identidade profissional dos estudantes (Cunha; Morosini, 2014).

Os fatores institucionais também exercem papel determinante no combate à evasão no ensino superior, pois cursos com altas taxas de reprovação em disciplinas iniciais, como cálculo, estatística e física, funcionam como barreiras de permanência, sendo assim observa-se que muitos estudantes desistem nos primeiros semestres. Programas de nivelamento e tutoria se tornam fundamentais, por isso, universidades que investem em reforço pedagógico conseguem reduzir índices de abandono, desse modo a evasão está diretamente relacionada ao suporte oferecido pelas instituições, contudo a ausência de acompanhamento contínuo agrava o risco de exclusão e perpetua ciclos de reprovação e desistência (Colpani, 2018).

De acordo com Pinto (2021), a evasão universitária possui um impacto direto e significativo no orçamento público, pois resulta no desperdício de recursos financeiros destinados à educação superior. O abandono de cursos compromete os investimentos realizados em infraestrutura, contratação de docentes e manutenção dos programas acadêmicos, além de afetar o planejamento orçamentário das instituições de ensino. Esse fenômeno acarreta em uma ineficiência do gasto público, já que os recursos alocados para a formação de estudantes não são devidamente aproveitados, impedindo que o potencial educacional e os benefícios de longo prazo sejam alcançados.

Além disso, as causas que geram a evasão universitária estão interligadas com questões econômicas e sociais, que influenciam diretamente o desempenho das políticas públicas. A falta de adaptação dos estudantes ao ambiente acadêmico, somada às dificuldades financeiras, acarreta uma pressão sobre os cofres públicos, uma vez que o abandono de cursos de graduação pode resultar em maior demanda por serviços de apoio social e em um aumento nas taxas de desemprego e subemprego. As políticas públicas focadas em reduzir a evasão precisam, portanto, considerar essa interdependência entre os custos individuais e coletivos da desistência universitária, a fim de promover não apenas a permanência dos alunos, mas também a maximização do retorno dos investimentos públicos no setor educacional (Pinto, 2021).

2.1.3 Dimensão sistêmica e estrutural da evasão

Conforme discutido anteriormente, a evasão escolar no ensino superior é um fenômeno abrangente, não se tratando apenas de uma escolha individual do estudante, mas de um reflexo de condições estruturais que exigem uma análise aprofundada para ser compreendido integralmente. Nesse contexto, a aplicação de modelos preditivos se torna uma ferramenta eficaz para identificar, de maneira antecipada, os fatores de risco associados à desistência escolar, permitindo às instituições tomar medidas preventivas e personalizadas que favoreçam os resultados educacionais (Morais; Souza, 2024).

Silva Filho et al. (2007), em análise do período 2000-2005, identificaram uma taxa média anual de evasão de 22% no sistema, sendo de 12% nas instituições públicas e de 26% nas privadas. Os autores observam que, embora a falta de recursos financeiros seja apontada como principal causa tanto por instituições públicas quanto privadas, essa resposta constitui uma simplificação, uma vez que questões de ordem acadêmica, as expectativas do aluno em relação à sua formação e sua integração à instituição figuram, na maior parte das vezes, como os fatores determinantes da decisão de evadir.

Nessa mesma direção, Baggi e Lopes (2011) situam a evasão no ensino superior como um fenômeno social complexo, cuja análise não pode ser desvinculada do contexto histórico mais amplo nem dos níveis anteriores de ensino, uma vez que reflete desigualdades estruturais que se acumulam ao longo da trajetória escolar. Em uma perspectiva sistêmica, as autoras chamam atenção para a escassez de instituições brasileiras que mantêm programas institucionais regulares de combate à evasão, com planejamento de ações, acompanhamento de resultados e sistematização de experiências bem-sucedidas. Apontam, ainda, fragilidades na produção e na disponibilização de dados oficiais sobre o fenômeno: os números do Inep não explicitam a saída de alunos de forma direta, exigindo cálculos aproximados a partir de matriculados, ingressantes e concluintes, e não permitem o rastreamento do percurso do estudante entre instituições, de modo que a evasão em uma IES pode, na prática, corresponder à matrícula em outra. Essa opacidade compromete tanto a comparabilidade entre instituições quanto a formulação de políticas públicas mais consistentes, o que reforça a necessidade de articular os estudos sobre evasão aos processos de avaliação institucional como meio de produzir diagnósticos mais qualificados e subsidiar intervenções no âmbito do sistema educacional brasileiro.

Estudos recentes reforçam a necessidade de ampliar essa leitura comparativa entre os diferentes segmentos do ensino superior brasileiro, especialmente no caso das instituições

públicas. Melo Gonçalves (2024), ao analisar quantitativamente a evasão nas Instituições de Ensino Superior públicas do Brasil com base nas Sinopses Estatísticas do Ensino Superior do INEP no período de 2016 a 2020, identificou uma média anual nacional de evasão de 17,19%, com variações regionais expressivas: Norte, Sul e Centro-Oeste apresentaram índices superiores à média nacional, enquanto Nordeste e Sudeste ficaram abaixo dela. Em uma perspectiva sistêmica, o autor chama atenção para a ausência de consenso metodológico na mensuração da evasão, argumentando que essa limitação compromete a formulação de políticas públicas eficazes, uma vez que dificulta tanto a comparação entre instituições e contextos quanto a identificação de padrões consistentes no sistema. Destaca, ainda, a escassez de trabalhos com dados em nível nacional sobre as IES públicas e a heterogeneidade dos índices entre regiões e estados, defendendo a necessidade de análises mais regionalizadas e específicas, capazes de captar as particularidades de cada contexto na formulação de estratégias de redução da evasão. Tais resultados evidenciam que, mesmo no âmbito da rede pública, a evasão se manifesta de forma marcadamente desigual no território nacional, o que reforça a sua dimensão sistêmica e estrutural: a forma como o ensino superior está organizado regionalmente, a fragilidade dos indicadores disponíveis e a ausência de parâmetros metodológicos uniformes constituem elementos que extrapolam a esfera individual ou institucional e condicionam, de modo concreto, a capacidade do próprio sistema de diagnosticar o problema e formular respostas adequadas.

Trazendo para o âmbito do setor público, Gilioli (2016), em estudo técnico sobre as IFES, analisa o fenômeno da evasão nos cursos de graduação das Instituições Federais de Ensino Superior (IFES) e, de certo modo, ainda é multifatorial e complexo. O autor indica que a evasão nesse âmbito pode ser categorizada em três modalidades principais: a evasão do curso (abandono de um curso específico), a evasão da instituição (desligamento da IES) e a evasão do sistema (abandono definitivo ou temporário do ensino superior). Além disso, distingue entre evasão anual (percentual de alunos que não se matriculam no ano seguinte) e evasão total (alunos que não concluem o curso no período regular).

Gilioli (2016) também apresenta que um ponto chave para a análise nas instituições federais foi a implementação do Sistema de Seleção Unificada (SISU), pois trouxe mudanças significativas no perfil discente das IFES, ampliando a mobilidade geográfica e a diversidade socioeconômica dos ingressantes. Embora haja percepções de que o SISU tenha aumentado a evasão devido ao deslocamento de estudantes, nota-se que essa relação não é linear nem determinante. O impacto varia conforme o curso, a região e a existência de políticas locais de permanência. O autor conclui que a redução da evasão depende menos de políticas nacionais

isoladas e mais de ações institucionais articuladas, como revisão curricular, programas de assistência e integração acadêmica, no âmbito da autonomia das IFES.

Para complementar a análise no âmbito das IFES, apresenta-se uma análise do Tribunal de Contas da União (TCU), órgão que exerce o controle externo e realiza a fiscalização contábil, financeira, orçamentária, operacional e patrimonial dos entes e entidades públicas, avaliando a legalidade dos atos, a legitimidade das ações e a boa aplicação dos recursos sob o princípio da economicidade (TCU, 2022). O TCU, por meio do Acórdão 1868/2024, realizou uma auditoria operacional com foco no desempenho de 69 instituições federais de ensino superior brasileiras. Nesse procedimento, identificou a evasão escolar no ensino superior como um dos principais desafios identificados.

O relatório aponta a inexistência de políticas estruturadas e informações suficientes para o gerenciamento efetivo do fenômeno, tanto por parte do Ministério da Educação (MEC) quanto das próprias instituições de ensino federais. Verificou-se que as taxas de desistência acumulada superaram as taxas de conclusão entre 2010 e 2017, indicando uma ineficiência na retenção de estudantes (TCU, 2024).

A análise do TCU (2024) revelou que as ações de combate à evasão concentram-se em grande parte em programas de transferência de renda, como a Política Nacional de Assistência Estudantil (PNAES), sem que haja uma estratégia nacional integrada com diagnóstico, metas definidas ou instâncias de monitoramento. Além disso, 51% das universidades declararam não possuir diagnóstico do perfil dos alunos evadidos ou das causas subjacentes ao abandono, o que dificulta ainda mais a adoção de medidas de prevenção.

Conforme o relatório, um aspecto crítico destacado é a limitação do conceito adotada pelo Instituto Nacional de Estudos e Pesquisas Educacionais (Inep), que define evasão apenas como a saída antecipada do curso, desconsiderando motivos como transferência interna entre cursos ou instituições. Essa abordagem prejudica a compreensão do fenômeno e a formulação de políticas adequadas, além de subestimar o real desempenho das universidades na formação de profissionais de nível superior (TCU, 2024).

Desse modo, em sua auditoria, o TCU (2024) recomenda a elaboração de uma política nacional de prevenção e redução da evasão universitária, com base em parâmetros técnicos e evidências, incluindo a padronização de conceitos, a integração de bases de dados e a promoção de ações específicas conforme o perfil e as causas da evasão. A superação dessas fragilidades é essencial para melhorar a eficiência e a efetividade das universidades federais no cumprimento de sua missão educacional.

2.2 APLICAÇÃO DE TÉCNICAS DE MINERAÇÃO DE DADOS E *MACHINE LEARNING* NA EDUCAÇÃO

A Inteligência Artificial (IA) é um campo interdisciplinar da ciência da computação que busca desenvolver sistemas capazes de realizar tarefas que exigem raciocínio humano, como aprendizado, reconhecimento de padrões e tomada de decisão, simulando processos cognitivos de maneira automatizada (Russell; Norvig, 2016). Dentro desse campo amplo, destacam-se dois subcampos diretamente relacionados ao presente estudo: a Mineração de Dados, que se ocupa da descoberta de padrões úteis em grandes volumes de dados, e o *Machine Learning*, que envolve o desenvolvimento de algoritmos capazes de aprender com os dados e realizar previsões sobre novos casos (Hastie; Tibshirani; Friedman, 2009; Kelleher; Mac Namee; D'Arcy, 2015). A aplicação dessas técnicas tem se expandido em diferentes setores da sociedade, incluindo saúde, indústria, segurança e educação, com transformação significativa na forma como se compreendem fenômenos complexos a partir de dados, tornando-se essencial investigar suas potencialidades para enfrentar problemas estruturais como a evasão estudantil.

No contexto educacional, essas técnicas vêm sendo aplicadas como ferramentas de apoio à aprendizagem e à gestão acadêmica, possibilitando desde a personalização de conteúdos até a identificação precoce de dificuldades acadêmicas. Segundo Heggler, Szmoski e Miquelin (2025), o uso dessas técnicas na educação apresenta uma dualidade: potencializa novas formas de interação e análise de dados e levanta preocupações relacionadas a vieses algorítmicos e desigualdades no acesso. Na mesma linha, Bandeira e Aquino (2025) analisam o papel transformador da Inteligência Artificial no futuro da educação, ressaltando a necessidade de que os docentes atuem como mediadores críticos frente às inovações tecnológicas.

Estudos recentes no Brasil discutem a incorporação dessas técnicas ao ambiente educacional com diferentes finalidades, incluindo a personalização do ensino, a avaliação de desempenho e a melhoria da acessibilidade (Pacheco et al., 2024). Os autores destacam o potencial transformador dessas ferramentas, mas apontam também a necessidade de cuidado com questões éticas, especialmente quanto à privacidade e à segurança dos dados dos estudantes. Albuquerque, Abreu e Lima (2024) complementam essa discussão ao analisarem a aplicação dessas técnicas na personalização do ensino, evidenciando que sistemas de tutoria inteligentes e métodos de análise de dados educacionais têm potencial para melhorar o engajamento e o desempenho acadêmico, oferecendo subsídios valiosos para a tomada de

decisão dos educadores. Em perspectiva complementar, Valente (2018) argumenta que as tecnologias digitais, quando integradas de forma intencional às práticas pedagógicas, podem inovar os processos de ensino e aprendizagem, ressaltando que sua efetividade depende do alinhamento entre os recursos tecnológicos disponíveis e as demandas pedagógicas concretas das instituições de ensino.

O avanço das tecnologias digitais e a crescente disponibilidade de grandes volumes de dados acadêmicos consolidaram a *Educational Data Mining* (EDM), ou Mineração de Dados Educacionais, como um campo interdisciplinar que articula estatística, ciência da computação e ciências da educação. Seu objetivo central é explorar dados oriundos de ambientes de aprendizagem, como histórico escolar, participação em plataformas virtuais, interações em fóruns, tempo de permanência em atividades e dados socioeconômicos, com vistas à identificação de padrões capazes de subsidiar decisões pedagógicas e administrativas (Romero; Ventura, 2020).

Uma revisão sistemática, publicada recentemente por Kalita et al. (2025), mostra que, na última década, houve um crescimento expressivo das pesquisas em EDM, com destaque para a aplicação de técnicas de aprendizado de máquina em tarefas como previsão de desempenho, detecção de evasão e personalização do ensino. Os autores salientam que a EDM tem se tornado um processo cada vez mais iterativo, envolvendo não apenas a construção de modelos preditivos, mas também a avaliação contínua de sua eficácia em contextos reais.

Novas fronteiras têm sido exploradas pela EDM em associação com tecnologias imersivas. Lampropoulos e Evangelidis (2025), identificam que a combinação entre EDM, *Learning Analytics*, que é processo de coletar, analisar e interpretar dados sobre o aprendizado para otimizar a educação e melhorar o desempenho dos alunos, e ambientes de realidade aumentada, realidade virtual e metaverso¹ tem ampliado as possibilidades de acompanhamento da aprendizagem, promovendo maior engajamento e autorregulação por parte dos estudantes. Essa ampliação evidencia o potencial da EDM em contextos híbridos e inovadores de ensino.

Além do aspecto pedagógico, estudos têm ressaltado o impacto da EDM nas dimensões administrativas das instituições. Almaghrabi et al. (2024) destacam que os modelos derivados de mineração de dados, não apenas apoiam o acompanhamento do

¹Segundo Mystakidis (2022), o metaverso pode ser compreendido como um ambiente digital persistente e multiusuário, voltado à interação social em tempo real, no qual os usuários, por meio de avatares, participam de experiências imersivas que combinam realidade virtual (VR) e realidade aumentada (AR).

estudante, mas também fornecem insumos relevantes para a gestão universitária, auxiliando na formulação de políticas públicas e na distribuição eficiente de recursos.

No Brasil, pesquisas têm buscado alinhar a EDM com os desafios da proteção de dados e da Lei Geral de Proteção de Dados Pessoais (LGPD). Tertulino (2025) realizou um estudo comparativo entre abordagens de aprendizado centralizado e federado aplicadas a microdados do Sistema de Avaliação da Educação Básica (SAEB), concluindo que a EDM pode ser desenvolvida em conformidade com princípios de privacidade sem perda significativa de desempenho nos modelos preditivos. Essa discussão é particularmente relevante para instituições públicas de ensino, que lidam com informações sensíveis de milhares de estudantes.

Em síntese, a EDM evoluiu de um campo emergente para uma área consolidada da pesquisa em educação, expandindo-se para além da predição de desempenho individual e alcançando dimensões institucionais, tecnológicas e éticas. Ao articular inovação pedagógica, suporte à gestão e responsabilidade com dados sensíveis, a EDM configura-se como ferramenta estratégica para enfrentar problemas estruturais do ensino superior brasileiro, como a evasão escolar.

Sendo assim, a aplicação de técnicas de Mineração de Dados e *Machine Learning* na educação não deve ser compreendida apenas como recurso tecnológico, mas como instrumento estratégico de apoio à gestão acadêmica voltada à permanência estudantil. Ao viabilizar a construção de sistemas de alerta precoce, essas técnicas tornam possível identificar estudantes com tendência à evasão antes que o abandono se concretize, oferecendo aos gestores informações objetivas e tempestivas, que conferem maior agilidade ao processo decisório e permitem o direcionamento de ações preventivas. Nesse sentido, a disponibilização dos resultados da Mineração de Dados Educacionais aos coordenadores de curso e demais gestores constitui elemento essencial para que o conhecimento gerado por esses modelos efetivamente se traduza em intervenções capazes de mitigar o fenômeno da evasão (Schmitt; Bernardi; Kantorski, 2019).

2.3 MODELOS PREDITIVOS E SUA APLICAÇÃO À EVASÃO NO ENSINO SUPERIOR

A utilização de técnicas de Aprendizado de Máquina (*Machine Learning*) para prever a evasão no ensino superior representa um avanço significativo na gestão acadêmica, inserindo-se no campo da EDM. Este campo se dedica a desenvolver métodos para explorar dados gerados no ambiente educacional, com o objetivo de compreender os processos de

aprendizagem e otimizar os resultados institucionais (Baker; Yacef, 2009). A predição da evasão é, talvez, sua aplicação mais impactante, pois permite que as IES migrem de um modelo reativo, que analisa a evasão após sua ocorrência, para um modelo proativo, que identifica os estudantes em risco e possibilita a implementação de intervenções pedagógicas e de apoio de forma precoce e direcionada (Magalhães et al., 2025).

A eficácia desses modelos preditivos reside na sua capacidade de processar e identificar padrões em conjuntos de dados multidimensionais, que integram variáveis acadêmicas (desempenho, frequência), demográficas (idade, gênero, localidade) e socioeconômicas (renda, auxílios). A seguir, detalham-se as principais classes de algoritmos empregadas para esta finalidade.

2.3.1 Modelos lineares e baseados em regras

Os primeiros modelos aplicados à predição de evasão foram, em grande parte, aqueles que ofereciam maior transparência nos resultados. A interpretabilidade é um fator fundamental no contexto educacional, pois as decisões baseadas nas predições do modelo (como contatar um aluno para oferecer apoio) exigem uma justificativa clara e compreensível para os gestores acadêmicos.

A Regressão Logística é um dos algoritmos mais tradicionais nesta categoria. Trata-se de um modelo estatístico que estima a probabilidade de um evento binário ocorrer (no caso, evadir ou permanecer) com base em um conjunto de variáveis independentes. Sua principal vantagem é a clareza: os coeficientes gerados para cada variável indicam a direção (positiva ou negativa) e a magnitude de sua influência na probabilidade de evasão (Kotsiantis, 2007).

As Árvores de Decisão, por sua vez, oferecem uma interpretabilidade ainda mais intuitiva. Este algoritmo funciona particionando recursivamente o conjunto de dados em subconjuntos cada vez mais homogêneos, com base em critérios de divisão (como o Índice Gini ou a Entropia) aplicados às variáveis (Witten; Frank; Hall, 2011). O resultado é uma estrutura hierárquica de regras "se-então" que pode ser facilmente visualizada e compreendida.

2.3.2 Métodos de conjunto (*Ensemble methods*)

As duas principais abordagens nesta categoria são o *bagging* (*Bootstrap Aggregating*) e o *boosting*.

A primeira abordagem, *bagging*, consiste em treinar múltiplos modelos de forma independente e paralela, cada um em uma subamostra aleatória dos dados. O algoritmo mais proeminente desta classe é a Floresta Aleatória (*Random Forest*), proposta por Breiman (2001), que se baseia na combinação de múltiplas árvores de decisão. Essa técnica reduz o risco de sobreajuste e garante classificadores mais estáveis, além de oferecer maior interpretabilidade ao permitir a análise da importância relativa dos atributos utilizados na classificação.

Em contraste com a abordagem paralela do *bagging*, os métodos de *boosting* constroem os modelos de forma sequencial e adaptativa. A lógica do *boosting*, popularizada por algoritmos como o *Gradient Boosting Machine* (GBM) e suas variantes modernas como o XGBoost, consiste em treinar um modelo inicial e, em seguida, treinar um segundo modelo para corrigir os erros do primeiro. Este processo iterativo continua, com cada novo modelo focando nos casos mais difíceis de classificar. Essa especialização progressiva resulta em modelos de altíssima performance preditiva (Chen; Guestrin, 2016).

2.3.3 Modelos de alta complexidade

Esta categoria inclui algoritmos capazes de modelar relações extremamente complexas e não-lineares, mas que, em contrapartida, operam de maneira menos transparente.

A Máquina de Vetores de Suporte (SVM) é um classificador que busca um "hiperplano" ótimo para separar as classes com a maior margem possível, utilizando o "truque do kernel" (*kernel trick*) para lidar com dados não-lineares (Hastie; Tibshirani; Friedman, 2009).

As Redes Neurais Artificiais e a Aprendizagem Profunda (*Deep Learning*) estão entre os modelos de maior complexidade no campo do aprendizado de máquina. Inspiradas de forma abstrata no funcionamento de neurônios biológicos, essas técnicas são compostas por camadas de unidades computacionais capazes de aprender padrões hierárquicos nos dados. O treinamento ocorre por meio da retropropagação do erro (*backpropagation*), em conjunto com métodos de otimização que ajustam os pesos da rede para reduzir a diferença entre a saída prevista e a saída esperada. Apesar de seu potencial para identificar relações complexas e interações sutis entre variáveis, sua aplicação a dados tabulares acadêmicos pode ser desafiadora, pois esses modelos tendem a exigir maior volume de dados, maior custo computacional e apresentam menor interpretabilidade em comparação a modelos lineares ou baseados em árvores simples (Heaton, 2017).

2.3.4 Métodos baseados em instâncias

Os métodos baseados em instâncias, também denominados aprendizado por memória (*memory-based learning*) ou aprendizado preguiçoso (*lazy learning*), constituem uma categoria de algoritmos de classificação supervisionada distinta das abordagens apresentadas nas subseções anteriores. Diferentemente de algoritmos que constroem explicitamente uma função de decisão durante o treinamento, os algoritmos dessa classe armazenam os exemplos de treinamento e adiam todo o processamento para o momento da predição, classificando cada nova instância com base em sua similaridade com os exemplos já observados (Hastie; Tibshirani; Friedman, 2009).

Entre os algoritmos dessa classe, destaca-se o K-Vizinhos Mais Próximos (*K-Nearest Neighbors* - KNN), introduzido por Fix e Hodges (1951) sobre classificação não paramétrica e posteriormente formalizado por Cover e Hart (1967). O KNN classifica uma nova instância identificando os k exemplos de treinamento mais próximos a ela no espaço de variáveis, onde proximidade é tipicamente medida por uma função de distância, como a distância euclidiana ou a distância de Manhattan, adotada neste estudo, e atribuindo à instância a classe majoritária entre esses vizinhos (Kotsiantis, 2007). A escolha do valor de k é um hiperparâmetro crítico: valores muito baixos tornam o modelo sensível a ruído nos dados (*overfitting*), enquanto valores muito altos suavizam excessivamente as fronteiras de decisão (*underfitting*).

A principal limitação do KNN em contextos de alta dimensionalidade, como o presente estudo, está associada ao fenômeno conhecido como maldição da dimensionalidade (*curse of dimensionality*), termo introduzido por Bellman (1961) e amplamente discutido por Hastie, Tibshirani e Friedman (2009). À medida que o número de variáveis aumenta, as distâncias entre as observações tornam-se progressivamente menos informativas para fins de discriminação, reduzindo a capacidade dos métodos baseados em vizinhança de identificar padrões locais relevantes. Como consequência, a eficácia do KNN tende a diminuir em conjuntos de dados com elevado número de atributos.

Adicionalmente, o KNN é um algoritmo de natureza computacionalmente intensiva na fase de predição: como não constrói um modelo durante o treinamento, exige a comparação da nova instância com os exemplos armazenados no conjunto de treinamento, tornando-o pouco adequado para aplicações em tempo real com grandes volumes de dados (Kelleher, Mac Namee e D'Arcy, 2015). Apesar dessas limitações, o KNN continua sendo amplamente utilizado como referência de comparação em estudos de classificação, por ser um algoritmo não-paramétrico que não faz suposições sobre a distribuição dos dados.

2.3.5 Métricas de avaliação de modelos preditivos

A avaliação do desempenho de modelos preditivos constitui etapa indispensável no desenvolvimento de soluções de aprendizado de máquina, pois permite mensurar objetivamente a capacidade do modelo de generalizar para dados não vistos durante o treinamento (Kelleher, Mac Namee e D'Arcy, 2015). A escolha das métricas deve ser orientada pelas características do problema, em especial a distribuição das classes e os custos assimétricos dos diferentes tipos de erro, e não apenas pelo desempenho agregado do modelo (Hastie; Tibshirani; Friedman, 2009).

No campo da Mineração de Dados Educacionais (EDM), essa escolha assume relevância particular. Baker e Yacef (2009) argumentam que, em aplicações orientadas à intervenção institucional, o erro de não identificar um estudante em risco, denominado falso negativo, é frequentemente mais custoso do que o erro de gerar um alerta desnecessário, o falso positivo. Essa assimetria justifica a adoção de métricas que penalizam explicitamente os erros na classe positiva, como o recall e o F1-score macro, em complemento às métricas globais de desempenho como a acurácia e o AUC-ROC (Fawcett, 2006).

As seções a seguir apresentam os fundamentos teóricos de cada métrica utilizada na avaliação dos modelos preditivos deste estudo.

2.3.5.1 Matriz de confusão

A matriz de confusão é a estrutura fundamental a partir da qual todas as métricas de avaliação de modelos de classificação são derivadas. Conforme Kelleher, Mac Namee e D'Arcy (2015), trata-se de uma tabela que registra, para cada combinação possível entre classe real e classe prevista pelo modelo, a contagem de instâncias correspondente. Para problemas de classificação binária, a matriz assume uma estrutura 2×2 , conforme ilustrado no Quadro 1.

Quadro 1 - Estrutura da matriz de confusão para classificação binária

		Classe Predit	
		Positivo	Negativo
Valor Real	Positivo	VP - Verdadeiro Positivo	FN - Falso Negativo
	Negativo	FP - Falso Positivo	VN - Verdadeiro Negativo

Fonte: Adaptado de Kelleher, Mac Namee e D'Arcy (2015).

Cada célula da matriz representa um tipo específico de resultado da predição:

- Verdadeiro Positivo (VP): instâncias corretamente classificadas como pertencentes à classe positiva. No contexto da predição de evasão, corresponde aos estudantes evadidos identificados corretamente pelo modelo.
- Verdadeiro Negativo (VN): instâncias corretamente classificadas como pertencentes à classe negativa. Corresponde aos estudantes não evadidos classificados corretamente.
- Falso Positivo (FP): instâncias incorretamente classificadas como positivas, sendo na realidade negativas, também denominado erro Tipo I ou alarme falso. Corresponde a estudantes que não evadiram, mas foram sinalizados como em risco.
- Falso Negativo (FN): instâncias incorretamente classificadas como negativas, sendo na realidade positivas, também denominado erro Tipo II. Corresponde a estudantes evadidos não identificados pelo modelo. Em aplicações de alerta precoce, este é o erro de maior custo institucional, pois representa oportunidades perdidas de intervenção.

Hastie, Tibshirani e Friedman (2009) destacam que a análise da matriz de confusão revela não apenas o desempenho geral do modelo, mas também a natureza específica dos erros cometidos, permitindo ajustar o modelo ou o limiar de decisão conforme as prioridades institucionais.

2.3.5.2 Acurácia

A acurácia é a métrica de avaliação mais intuitiva de um modelo de classificação, representando a proporção de predições corretas em relação ao total de instâncias avaliadas, independentemente da classe a que pertencem (Kelleher, Mac Namee e D'Arcy, 2015). Sua formulação é dada por: $Accuracy = (VP + VN) / (VP + VN + FP + FN)$.

Embora seja a métrica de mais fácil comunicação para gestores e tomadores de decisão, a acurácia possui uma limitação relevante: pode ser enganosa em *datasets* com classes fortemente desbalanceadas. Conforme Hastie, Tibshirani e Friedman (2009), um modelo que classifica todas as instâncias como pertencentes à classe majoritária pode atingir acurácia muito elevada sem ter aprendido nenhum padrão discriminativo, fenômeno denominado paradoxo da acurácia. Por essa razão, Kelleher, Mac Namee e D'Arcy (2015) recomendam que a acurácia seja sempre interpretada em conjunto com outras métricas, especialmente recall e F1-score, para garantir uma avaliação completa e não enviesada do modelo.

2.3.5.3 Precisão (*Precision*)

A precisão, também denominada valor preditivo positivo, mensura a proporção de instâncias corretamente classificadas como positivas em relação ao total de instâncias classificadas como positivas pelo modelo, sejam elas verdadeiramente positivas ou não (Fawcett, 2006). Sua formulação é: $Precision = VP / (VP + FP)$.

Uma precisão elevada indica que o modelo gera poucos alarmes falsos, ou seja, quando sinaliza um estudante como em risco de evasão, essa previsão tende a estar correta. No contexto institucional, essa propriedade é relevante para garantir a eficiência no uso dos recursos de apoio estudantil, evitando que equipes pedagógicas e de assistência social sejam sobrecarregadas com alertas desnecessários (Baker; Yacef, 2009).

Na modalidade macro, adotada neste estudo por ser mais robusta em cenários de classes com frequências distintas, a precisão é calculada independentemente para cada classe e, em seguida, computa-se a média aritmética simples entre os valores obtidos:

$$Precision_macro = (Precision_classe0 + Precision_classe1) / 2$$

2.3.5.4 Sensibilidade (*Recall*)

O *recall*, também denominado sensibilidade ou taxa de verdadeiros positivos, mensura a proporção de instâncias realmente positivas que foram corretamente identificadas pelo modelo como positivas (Fawcett, 2006). Sua formulação é: $Recall = VP / (VP + FN)$.

No contexto da predição de evasão acadêmica, o *recall* é a métrica de maior relevância prática para sistemas de alerta precoce. Baker e Yacef (2009) argumentam que, em aplicações de EDM orientadas à intervenção, o erro mais custoso é o falso negativo, deixar de identificar

um estudante em risco significa perder a oportunidade de ofertar apoio institucional que poderia reverter o processo de abandono. Essa lógica é corroborada por Slade e Prinsloo (2013), que enfatizam a responsabilidade ética das instituições de ensino em agir proativamente diante de indicadores de risco.

Hastie, Tibshirani e Friedman (2009) ilustram com clareza por que o *recall* é frequentemente mais relevante que a acurácia em problemas com custos assimétricos de erro: um modelo que nunca identifica nenhuma instância como positiva pode atingir acurácia elevada em *datasets* desbalanceados, mas apresenta *recall* nulo, sendo completamente inútil para fins de alerta precoce. Na modalidade macro: $\text{Recall_macro} = (\text{Recall_classe0} + \text{Recall_classe1}) / 2$.

2.3.5.5 F1-score

O F1-score é uma métrica que sintetiza a precisão e o *recall* em um único indicador, calculado como a média harmônica entre os dois valores (Kelleher, Mac Namee e D'Arcy, 2015). A média harmônica é mais restritiva que a média aritmética: um F1-score elevado exige que tanto a precisão quanto o *recall* sejam simultaneamente elevados, o modelo não pode compensar baixo *recall* com alta precisão, ou vice-versa:

$$F1 = 2 \times (\text{Precision} \times \text{Recall}) / (\text{Precision} + \text{Recall})$$

O F1-score é especialmente útil em contextos onde tanto os falsos positivos quanto os falsos negativos têm custos significativos (Fawcett, 2006). No presente estudo, são reportadas duas modalidades de agregação:

A primeira, F1-score Macro, que calcula o F1 individualmente para cada classe e computa a média aritmética simples, atribuindo peso igual a todas as classes independentemente de sua frequência no dataset: $F1_macro = (F1_classe0 + F1_classe1) / n_classes$

Conforme Kelleher, Mac Namee e D'Arcy (2015), essa modalidade é a mais adequada para *datasets* com algum nível de desbalanceamento, pois não permite que a classe majoritária domine a métrica agregada. Por essa razão, o F1-score macro é adotado como métrica principal de comparação entre modelos neste estudo.

E a segunda, F1-score Weighted, que calcula a média ponderada do F1 de cada classe, utilizando a frequência relativa de cada classe como peso: $F1_weighted = \sum (\text{suporte_i} / \text{total}) \times F1_i$

Essa modalidade tende a produzir valores mais elevados quando a classe majoritária tem melhor desempenho, podendo mascarar dificuldades do modelo na classe minoritária. Hastie, Tibshirani e Friedman (2009) recomendam seu uso como métrica complementar, reportada em conjunto com o F1-score macro para maior transparência na avaliação.

2.3.5.6 AUC-ROC

A AUC-ROC (*Area Under the Receiver Operating Characteristic Curve*) é amplamente considerada a métrica mais robusta e completa para avaliação de modelos de classificação binária. Diferentemente das métricas anteriores, que avaliam o desempenho do modelo para um único limiar de decisão fixo, a AUC-ROC avalia o desempenho ao longo de todos os limiares possíveis, fornecendo uma medida global da capacidade discriminativa do modelo (Fawcett, 2006).

A curva ROC é construída plotando, para cada limiar de decisão possível, a Taxa de Verdadeiros Positivos (TVP, equivalente ao *recall*) no eixo Y contra a Taxa de Falsos Positivos ($TFP = FP / (FP + VN)$) no eixo X. Cada ponto na curva representa o par (TFP, TVP) obtido para um determinado limiar, e a curva completa ilustra o *trade-off* entre sensibilidade e especificidade ao longo de todo o espectro de limiares possíveis (Fawcett, 2006). A diagonal principal, que representa a linha $y = x$, corresponde ao desempenho de um classificador aleatório ($AUC = 0,50$), enquanto o ponto (0, 1) representa o classificador perfeito ($AUC = 1,0$). Quanto mais a curva se aproxima do canto superior esquerdo do gráfico, maior a capacidade discriminativa do modelo.

A AUC é calculada como a integral da curva ROC e possui uma interpretação probabilística direta: representa a probabilidade de que o modelo atribua uma pontuação mais alta a uma instância positiva escolhida aleatoriamente do que a uma instância negativa escolhida aleatoriamente (Fawcett, 2006): $AUC = P(\text{score}(\text{positivo}) > \text{score}(\text{negativo}))$

Conforme Kelleher, Mac Namee e D'Arcy (2015), não há critérios rígidos para definir um valor aceitável de AUC-ROC, sendo a decisão dependente da aplicação. Como regra prática, os autores indicam que valores acima de 0,70 caracterizam um modelo forte, enquanto valores abaixo de 0,60 indicam um modelo fraco. A faixa intermediária (entre 0,60 e 0,70) constitui uma zona em que a aceitabilidade do modelo deve ser avaliada ao contexto específico de aplicação.

Fawcett (2006) aponta que a métrica AUC-ROC é adequada para avaliar modelos de classificação porque permite analisar o desempenho do modelo considerando diferentes

possibilidades de limiar de decisão. Ela não depende da escolha de um único ponto de corte para definir se um estudante será classificado como provável evadido ou não evadido. Essa característica é importante em sistemas de alerta precoce, pois a definição desse ponto de corte pode variar conforme a capacidade da instituição de acompanhar os estudantes sinalizados e conforme o equilíbrio desejado entre identificar mais casos de evasão e reduzir alertas incorretos.

Dessa forma, o AUC-ROC foi adotado como métrica principal para a comparação dos modelos preditivos desenvolvidos neste estudo.

2.4 *CROSS INDUSTRY STANDARD PROCESS FOR DATA MINING (CRISP-DM)*

O CRISP-DM (*Cross Industry Standard Process for Data Mining*) foi desenvolvido na segunda metade da década de 1990, período em que a mineração de dados emergia como uma disciplina promissora, porém carente de um modelo metodológico unificado. Conforme descreve Shearer (2000), até então não havia um padrão amplamente aceito que orientasse o ciclo de vida de projetos de mineração de dados de forma estruturada, documentada e independente de ferramentas ou setores específicos. Para atender a essa lacuna, quatro grandes organizações: Daimler-Benz, Integral Solutions Ltd. (ISL), NCR e OHRA, uniram-se em 1996 para criar um modelo de referência aberto, não proprietário e amplamente documentado, capaz de padronizar as etapas envolvidas no processo analítico e promover as chamadas *best practices* em *data mining*. Esse consórcio, posteriormente ampliado com a criação do CRISP-DM *Special Interest Group* (SIG), envolveu mais de duzentos profissionais e pesquisadores do setor, incluindo desenvolvedores de software, consultores e usuários de diferentes segmentos, o que garantiu à metodologia um caráter verdadeiramente transindustrial e colaborativo.

O resultado desse esforço foi a elaboração de um modelo robusto e amplamente replicável, publicado oficialmente em 2000 como versão 1.0 do CRISP-DM. De acordo com Shearer (2000), o modelo se estrutura em seis fases principais: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação, que formam um ciclo contínuo de aprendizado e melhoria. As fases estão interconectadas por fluxos iterativos, permitindo revisões sucessivas sempre que novas informações ou limitações são descobertas. Essa característica torna o CRISP-DM uma metodologia dinâmica e adaptativa, em oposição aos modelos lineares e prescritivos que predominavam em iniciativas

anteriores. Sua proposta é que o conhecimento gerado em cada etapa retroalimente o processo, refinando as análises e favorecendo tomadas de decisão mais assertivas.

Para Goldschmidt e Passos (2005), o CRISP-DM representa a evolução natural das metodologias de análise de dados, ao combinar aspectos técnicos e estratégicos de maneira integrada. Esses autores destacam que o modelo transcende o simples uso de algoritmos, pois enfatiza o entendimento do problema organizacional e a tradução de objetivos institucionais em metas de mineração de dados. Ao adotar uma lógica de gestão por processos, o CRISP-DM propõe uma sequência de tarefas com entregas bem definidas, que podem ser auditadas e replicadas, o que contribui para a maturidade analítica das instituições. Além disso, a sua estrutura modular favorece a documentação sistemática de todas as decisões tomadas, criando uma trilha de rastreabilidade essencial para ambientes corporativos e acadêmicos que demandam transparência e validação científica.

A consolidação do CRISP-DM como referência metodológica global decorre justamente de sua neutralidade tecnológica e setorial, permitindo sua adoção tanto por empresas privadas quanto por instituições públicas de ensino. O modelo provou ser especialmente útil em contextos em que a complexidade dos dados exige abordagens iterativas e baseadas em evidências, como ocorre na gestão universitária e na análise de evasão estudantil. Em ambientes educacionais, essa metodologia possibilita alinhar os objetivos institucionais, como a redução da evasão, às metas de mineração de dados, transformando grandes volumes de informações acadêmicas em indicadores estratégicos. Essa capacidade de integração entre ciência de dados e políticas de gestão educacional reforça a atualidade do CRISP-DM, mesmo após mais de duas décadas de sua criação.

Por sua natureza aberta e flexível, o CRISP-DM continua sendo, segundo levantamentos recentes de portais especializados como o Data Science PM, o modelo de mineração de dados mais utilizado no mundo, superando outras metodologias clássicas como o SEMMA, desenvolvido pela SAS, e o KDD (Knowledge Discovery in Databases) (DATA SCIENCE PM, 2024).

2.4.1 Compreensão do negócio (*Business understanding*)

A fase de Compreensão do Negócio (*Business Understanding*) é a etapa inaugural da metodologia CRISP-DM. Seu principal objetivo é alinhar o projeto de mineração de dados com as metas e necessidades estratégicas. Esta etapa vai além da simples identificação de um

problema, exigindo uma imersão no contexto organizacional para traduzir desafios de negócio em objetivos técnicos de mineração de dados (Chapman et al., 2000).

Corroborando essa perspectiva, Wirth e Hipp (2000) caracterizam a Compreensão do Negócio como a fase inicial do processo, cujo foco está em entender os objetivos e requisitos do projeto sob a perspectiva do negócio, convertendo esse entendimento em uma definição do problema de mineração de dados e em um plano preliminar de projeto orientado ao alcance desses objetivos. A partir da aplicação prática da metodologia em um projeto real, os autores destacam que essa etapa é determinante para que o processo de mineração de dados possa ser executado de forma confiável e replicável por diferentes equipes, inclusive por profissionais com menor experiência técnica, reforçando a importância de alinhar claramente os objetivos de negócio aos objetivos técnicos desde o início do projeto.

Nesta fase, determinam-se os critérios de sucesso sob uma perspectiva de negócio, antes mesmo de se considerar qualquer métrica técnica. No contexto de um projeto voltado à predição da evasão discente, por exemplo, esta etapa envolve compreender o impacto do fenômeno para a instituição.

Segundo Chapman et al. (2000), negligenciar esta etapa é um erro comum que pode comprometer todo o processo. Uma formulação inadequada do problema pode levar ao desenvolvimento de modelos que, embora tecnicamente precisos, são irrelevantes para a tomada de decisão. Portanto, a Compreensão do Negócio estabelece as fundações sobre as quais todas as fases subsequentes, desde a coleta de dados até a implantação do modelo, serão construídas, garantindo que a solução tecnológica desenvolvida responda efetivamente ao desafio organizacional proposto.

2.4.2 Compreensão dos dados (*Data understanding*)

Após a compreensão do negócio, a metodologia CRISP-DM avança para a fase de Compreensão dos Dados (*Data Understanding*). Esta etapa marca a primeira interação prática com os dados brutos e é fundamental para avaliar se o problema delineado na fase anterior pode ser efetivamente solucionado com as informações disponíveis. As atividades centrais aqui são a coleta, a descrição e a exploração inicial dos dados (Chapman et al., 2000).

A etapa começa com a coleta dos dados de diversas fontes. Em um contexto educacional, isso pode incluir bases de dados acadêmicas (histórico de notas, frequência, curso), registros demográficos (idade, gênero) e informações socioeconômicas. Em seguida, realiza-se uma análise exploratória de dados (EDA), que utiliza estatísticas descritivas e

técnicas de visualização, como histogramas e gráficos de correlação, para obter uma primeira percepção sobre a estrutura e o conteúdo do conjunto de dados.

O objetivo da EDA nesta fase não é construir modelos, mas sim identificar padrões iniciais, anomalias e relações entre variáveis que possam ser relevantes. Por exemplo, pode-se investigar se há uma correlação aparente entre as notas do primeiro semestre e a futura evasão, ou verificar a distribuição dos alunos por faixa de renda. Além disso, um dos focos mais importantes desta etapa é a verificação da qualidade dos dados, identificando problemas como valores ausentes, dados inconsistentes ou erros de digitação, que precisarão ser tratados na fase seguinte.

Como destacam Romero e Ventura (2020), compreender a natureza dos dados é uma condição essencial para que os modelos de EDM sejam consistentes e confiáveis. Portanto, esta fase funciona como um diagnóstico, fornecendo os insumos necessários para planejar a etapa de Preparação dos Dados e garantindo que as hipóteses iniciais sejam fundamentadas em evidências concretas extraídas dos próprios dados.

2.4.3 Preparação dos dados (*Data preparation*)

A fase de Preparação dos Dados (*Data Preparation*) sucede a exploração inicial e é onde os dados brutos são refinados para se tornarem adequados à modelagem. Esta etapa abrange todas as atividades necessárias para construir o conjunto de dados final que servirá de entrada para os algoritmos de mineração de dados. É amplamente reconhecida como uma das fases mais críticas e que consomem mais tempo em um projeto de ciência de dados, podendo ocupar de 60% a 80% do esforço total (Chapman et al., 2000).

As tarefas nesta fase são diversas e podem ser agrupadas em atividades de limpeza, construção e formatação. A limpeza de dados envolve o tratamento de valores ausentes (*missing values*), a correção de dados inconsistentes e a remoção de registros duplicados. A construção de dados inclui a engenharia de atributos (*variável engineering*), que consiste na criação de novas variáveis, mais informativas, a partir das já existentes. No contexto da evasão, em vez de usar apenas o status final de um aluno, poderiam ser criadas variáveis derivadas como "taxa de reprovação no primeiro ano" ou "desempenho relativo à média da turma", que podem ter um poder preditivo maior.

A relevância dessa etapa decorre do fato de que a qualidade dos dados influencia diretamente a qualidade dos modelos preditivos. Dados ausentes, inconsistentes, duplicados ou mal estruturados podem comprometer a aprendizagem dos algoritmos e gerar padrões

artificiais, reduzindo a confiabilidade dos resultados. Por isso, a preparação dos dados não deve ser compreendida apenas como uma etapa técnica, mas como uma fase decisiva para assegurar a consistência, a validade e a possibilidade de interpretação dos modelos produzidos.

Ao final desta fase, o resultado esperado é um conjunto de dados limpo, consistente e estruturado no formato adequado para as ferramentas de modelagem, reduzindo o risco de que os padrões identificados pelos algoritmos sejam artefatos decorrentes de dados ausentes, inconsistentes ou de baixa qualidade (Chapman et al., 2000).

2.4.4 Modelagem (*Modeling*)

Após a preparação e organização dos dados, o projeto avança para a fase de Modelagem (*Modeling*), que é o cerne do processo de mineração de dados. Nesta etapa, aplicam-se diversas técnicas e algoritmos de aprendizado de máquina (*machine learning*) para identificar padrões e construir modelos preditivos que atendam aos objetivos definidos na fase de Compreensão do Negócio.

A seleção da técnica de modelagem apropriada é um dos primeiros passos. A escolha depende da natureza do problema: para prever uma categoria (como "evade" ou "não evade"), utilizam-se algoritmos de classificação; para prever um valor numérico (como a nota final de um aluno), aplicam-se técnicas de regressão. Como ressalta Chapman et al. (2000), raramente um único algoritmo é testado. É uma prática comum aplicar diferentes técnicas, pois cada uma possui pressupostos e abordagens distintas, o que pode levar a resultados variados. Além disso, cada modelo exige calibração e ajuste de seus parâmetros para otimizar o desempenho.

No contexto da EDM, a modelagem busca transformar dados sobre o processo de aprendizagem e os comportamentos dos estudantes. Para a predição da evasão, que é um problema de classificação, a literatura consolidou o uso de diversos algoritmos. Baker e Yacef (2009), em sua revisão sobre o estado da arte em EDM, já apontavam que técnicas como regressão logística, árvores de decisão e redes neurais artificiais são frequentemente empregadas para prever o desempenho acadêmico e o risco de abandono. Mais recentemente, abordagens baseadas em conjuntos de modelos (*ensemble methods*), como *Random Forest* e *Gradient Boosting*, também ganharam destaque por sua alta capacidade preditiva.

2.4.5 Avaliação (*Evaluation*)

Após a construção dos modelos na fase anterior, a etapa de Avaliação (*Evaluation*) é dedicada a analisar os resultados de uma perspectiva técnica e, essencialmente, de negócio. O objetivo é determinar qual modelo melhor atende aos objetivos institucionais e possui o desempenho mais robusto para ser levado à fase de implantação (Chapman et al., 2000).

A avaliação ocorre em dois níveis. O primeiro é a avaliação técnica, na qual o desempenho dos modelos é medido quantitativamente. Para problemas de classificação, como a predição de evasão, são empregadas métricas como acurácia (proporção de previsões corretas), precisão (percentual de predições positivas que estavam corretas) e *recall* (percentual de casos positivos reais que foram corretamente identificados). A análise de ferramentas como a matriz de confusão e a curva ROC (*Receiver Operating Characteristic*) também é comum para entender como o modelo se comporta em diferentes cenários e limiares de decisão.

O segundo nível é a avaliação de negócio. Este é o ponto em que se verifica se os resultados respondem de forma satisfatória à questão de negócio original. Um modelo pode ter alta acurácia, mas ser inútil na prática. Por exemplo, um modelo que prevê a evasão com 95% de acurácia, mas falha em identificar os alunos que realmente evadem (baixo *recall*), não cumpre o objetivo de permitir uma intervenção precoce. Portanto, o critério final de sucesso não é apenas estatístico, mas também a sua aderência aos objetivos institucionais.

É fundamental que a avaliação não seja vista como um evento único no final do projeto. Como reforçam Kalita et al. (2025), a validação deve ser um processo contínuo e iterativo. Isso assegura a confiabilidade e a generalização dos modelos, verificando se eles são robustos o suficiente para performar bem em dados novos e em contextos reais de aplicação. Ao final desta fase, uma decisão é tomada sobre qual modelo seguir para a implantação ou se é necessário retornar a fases anteriores para refinar o processo.

2.4.6 Implantação (*Deployment*)

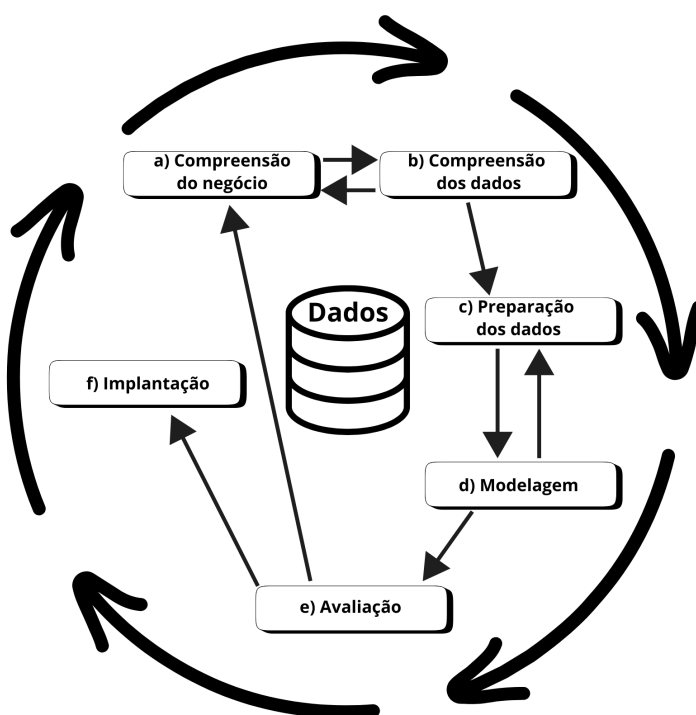
A fase de Implantação é a culminação do processo de mineração de dados, na qual o conhecimento adquirido e os modelos validados são integrados aos processos de negócio da organização para serem utilizados de forma prática. A criação de um modelo, por mais preciso que seja, não gera valor se não for efetivamente utilizado para apoiar a tomada de decisão (Chapman et al., 2000).

A forma de implantação pode variar amplamente, desde a geração de um relatório final com os dados descobertos até a integração do modelo preditivo em um sistema de informação já existente. Em universidades públicas, por exemplo, a implantação de um modelo para prever a evasão pode se materializar através do desenvolvimento de *dashboards* interativos para gestores, da criação de relatórios de risco para coordenadores de curso ou da implementação de sistemas de alerta integrados aos ambientes acadêmicos.

Nessa etapa, a análise de dados deixa de se restringir ao campo técnico e passa a se transformar em conhecimento acionável, com potencial para subsidiar políticas institucionais de permanência estudantil. A implantação também envolve o planejamento de mecanismos de monitoramento e manutenção, a fim de garantir que o modelo continue preciso, relevante e adequado ao contexto organizacional ao longo do tempo (Chapman et al., 2000).

A dinâmica inter-relacionada entre essas seis fases, desde a Compreensão do Negócio até a Implantação, é visualmente sintetizada na Figura 1. A figura destaca não apenas a sequência lógica das etapas, mas, principalmente, a natureza cíclica e iterativa do processo, representada pelas setas que indicam a possibilidade de retornar a fases anteriores. Essa flexibilidade, que permite, por exemplo, que os resultados da Avaliação (e) levem a um novo Compreensão do Negócio (a), é um dos principais motivos que consolidam o CRISP-DM.

Figura 1 - Fases do modelo de referência CRISP-DM.



Fonte: Adaptado de Shearer (2000).

A partir do referencial teórico apresentado, observa-se que a mineração de dados educacionais, aliada às técnicas de aprendizado de máquina, constitui uma abordagem consolidada para a antecipação de fenômenos complexos como a evasão acadêmica, permitindo que instituições de ensino superior transitem de uma postura reativa para uma atuação preventiva. O modelo CRISP-DM, discutido nesta seção, oferece a estrutura processual que orienta o presente estudo, articulando de forma sistemática as etapas que vão desde a compreensão do problema institucional até a implantação de soluções aplicáveis à gestão acadêmica. É sobre essa base teórico-metodológica que se assentam os Procedimentos metodológicos descritos no capítulo seguinte, no qual são detalhados os dados, as técnicas e os critérios adotados para a predição da evasão nos cursos de graduação da IFES investigada.

3 PROCEDIMENTOS METODOLÓGICOS

A metodologia científica constitui o conjunto de princípios e procedimentos sistemáticos que orientam a construção do conhecimento, garantindo validade, coerência e rigor às investigações. No campo das ciências sociais aplicadas, desempenha papel essencial na articulação entre teoria e prática, possibilitando que o conhecimento produzido contribua efetivamente para a solução de problemas concretos e socialmente relevantes (Gil, 2022; Lakatos; Marconi, 2017). Assim, a metodologia delinea as estratégias de investigação que permitem ao pesquisador compreender, interpretar e propor soluções para fenômenos complexos, assegurando o alinhamento entre os objetivos do estudo e as técnicas empregadas.

3.1 CARACTERIZAÇÃO DA PESQUISA

A presente pesquisa caracteriza-se como aplicada, uma vez que se volta à geração de conhecimento direcionado à resolução de um problema específico: a evasão discente nos cursos de graduação de uma instituição federal de ensino superior. A pesquisa aplicada, conforme Gil (2008), busca resultados que possam ser utilizados de forma prática por organizações ou pela sociedade. De acordo com Lakatos e Marconi (2017), diferencia-se da pesquisa básica por não se limitar à produção de saber teórico, mas por orientar suas ações à transformação da realidade. No contexto desta dissertação, esse caráter aplicado manifesta-se na investigação de um problema institucional concreto: a ausência de mecanismos sistemáticos de identificação precoce de estudantes em risco de evasão nos cursos de graduação da UFFS, *campus* Chapecó. Conforme apontado pelo Relatório de Auditoria Interna da instituição (UFFS, 2025), a universidade ainda não dispõe de política institucional consolidada nem de sistema formal e contínuo de monitoramento da evasão, o que limita a capacidade de intervenção preventiva.

A investigação adota uma abordagem quantitativa, uma vez que se fundamenta na coleta, tratamento e análise estatística de dados institucionais, com aplicação de algoritmos de aprendizado de máquina e avaliação por métricas objetivas de desempenho. Conforme Gil (2022), a pesquisa quantitativa caracteriza-se pelo uso de técnicas estatísticas para descrever, explicar ou prever fenômenos a partir de dados mensuráveis, o que se alinha plenamente ao desenvolvimento deste estudo. Os dados analisados compreendem variáveis acadêmicas como notas, frequência e reprovações, variáveis socioeconômicas como origem escolar, tipo de cota e participação em programas assistenciais, e variáveis comportamentais como o uso do

Restaurante Universitário e da Biblioteca, todos tratados por meio de técnicas estatísticas descritivas e algoritmos de aprendizado supervisionado. Sampieri, Collado e Lucio (2013) reforçam que a abordagem quantitativa é especialmente adequada quando o objetivo é identificar padrões, estabelecer relações entre variáveis e desenvolver modelos preditivos com base em evidências empíricas, propósito central desta dissertação.

O delineamento da pesquisa é descritivo, pois busca observar, registrar e analisar as características do fenômeno sem manipulá-lo. Tal tipo de investigação procura responder a questões do tipo “como” e “quem”, possibilitando a descrição sistemática das relações entre variáveis (Gil, 2022; Sampieri; Collado; Lucio, 2013).

Como procedimento metodológico, a pesquisa foi desenvolvida a partir de um estudo de caso na Universidade Federal da Fronteira Sul (UFFS), *Campus* Chapecó, utilizando dados acadêmicos institucionais como campo empírico para a construção e avaliação dos modelos preditivos. A delimitação a uma instituição específica justifica-se pela necessidade de acesso a dados reais, pela relevância do contexto institucional da UFFS para o problema investigado e pela possibilidade de que os resultados obtidos subsidiem diretamente a gestão acadêmica da própria instituição. Conforme Yin (2015), o estudo de caso constitui uma estratégia adequada para investigar fenômenos contemporâneos em seu contexto real, especialmente quando os limites entre o fenômeno e o contexto não estão claramente definidos. Gil (2022), por sua vez, caracteriza o estudo de caso pela análise detalhada de um ou poucos objetos, permitindo seu amplo conhecimento. Nesse sentido, sua utilização possibilitou analisar o problema em um contexto institucional concreto e desenvolver uma solução direcionada à realidade da UFFS.

3.2 COLETA, ANÁLISE E INTERPRETAÇÃO DE DADOS

A coleta, análise e interpretação dos dados seguiram as fases da metodologia CRISP-DM (Cross Industry Standard Process for Data Mining), adotada como referência metodológica para a construção e validação do modelo preditivo. Os dados foram obtidos junto à Secretaria Especial de Tecnologia da Informação (SETI) e à Diretoria de Registro Acadêmico (DRA) da UFFS, provenientes de quatro sistemas institucionais: o SIGAA (Sistema Integrado de Gestão de Atividades Acadêmicas), com os dados acadêmicos; o SAS (Sistema de Análise Socioeconômica), com os registros de participação no Serviço de Assistência ao Estudante; o SCI (Sistema de Credenciamento Institucional), responsável pelo gerenciamento do acesso dos estudantes ao Restaurante Universitário; e o *Pergamum*, sistema de gestão de bibliotecas, com o histórico de empréstimos. As etapas de compreensão do

negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação são descritas a seguir.

3.2.1 Aspectos éticos e proteção dos dados

A pesquisa utilizou dados administrativos secundários provenientes dos sistemas institucionais da Universidade Federal da Fronteira Sul. A extração e a disponibilização das bases foram realizadas pela Secretaria Especial de Tecnologia da Informação (SETI) e pela Diretoria de Registro Acadêmico (DRA), sem que o pesquisador tivesse acesso direto aos sistemas institucionais ou aos registros originais dos estudantes.

Antes de sua disponibilização, os identificadores diretos presentes nas bases, o número do Cadastro de Pessoas Físicas (CPF) e o número de matrícula, foram automaticamente substituídos por códigos aleatórios durante o processo institucional de extração. Dessa forma, o conjunto de dados recebido pelo pesquisador não continha os valores originais desses identificadores, nomes, endereços eletrônicos, telefones ou outras informações que possibilitassem a identificação direta dos estudantes.

O pesquisador não recebeu tabela de correspondência, chave de reversão ou outro mecanismo que permitisse relacionar os códigos utilizados nas bases aos titulares dos registros. Os identificadores transformados foram utilizados exclusivamente para possibilitar a integração dos dados provenientes dos diferentes sistemas institucionais e o agrupamento dos registros pertencentes à mesma trajetória acadêmica.

Considerando que a pesquisa utilizou exclusivamente dados administrativos secundários previamente desidentificados pela instituição, sem interação direta com os estudantes, aplicação de questionários, realização de entrevistas ou qualquer forma de intervenção, não houve submissão ao Comitê de Ética em Pesquisa. Durante todas as etapas do estudo, os dados foram utilizados exclusivamente para os objetivos acadêmicos estabelecidos nesta dissertação, com acesso restrito e apresentação dos resultados de forma agregada, sem identificação individual dos estudantes.

3.2.2 Metodologia CRISP-DM

Após a obtenção das bases institucionais previamente desidentificadas e a definição dos procedimentos éticos e de proteção dos dados, a organização e a execução das etapas de mineração de dados seguiram a metodologia CRISP-DM, consolidada como referência

internacional em projetos de ciência de dados (Chapman *et al.*, 2000; Shearer, 2000). Essa metodologia é composta por seis fases principais, apresentadas a seguir, que se inter-relacionam de forma iterativa e cíclica.

a) Compreensão do Negócio (*Business Understanding*)

A evasão discente foi identificada pela UFFS como um dos maiores desafios institucionais. Relatórios internos apontam que, entre 2015 e 2023, a taxa média de sucesso na graduação foi de apenas 33,11%, revelando que somente um terço dos estudantes concluiu seus cursos no tempo previsto (UFFS, 2025). Além disso, cerca de 70% dos casos de evasão correspondem a abandono não formalizado, detectado somente após o término do semestre letivo, o que evidencia a ausência de mecanismos institucionais de identificação precoce de estudantes em risco (UFFS, 2025). Diante desse cenário, definiu-se como objetivo de negócio a identificação antecipada de estudantes com maior probabilidade de evasão nos cursos de graduação do *campus* Chapecó, de modo a subsidiar ações preventivas de permanência por parte da gestão acadêmica. Como critério de sucesso, estabeleceu-se a obtenção de modelos preditivos com desempenho suficiente para uso institucional, priorizando a capacidade de identificar corretamente os estudantes em risco. Do ponto de vista da mineração de dados, o objetivo de negócio foi traduzido em um problema de classificação supervisionada, no qual cada matrícula é classificada quanto à probabilidade de evasão a partir de dados acadêmicos, socioeconômicos e comportamentais. Como decisões de escopo do projeto, definiu-se a integração de dados provenientes de quatro sistemas institucionais não interoperáveis (SIGAA, SAS, SCI e *Pergamum*), permitindo incorporar variáveis de uso de serviços assistenciais como o Restaurante Universitário e a Biblioteca, e a análise de uma série histórica de onze anos (2014 a 2024), possibilitando capturar padrões longitudinais de evasão em diferentes grupos de ingresso.

b) Compreensão dos Dados (*Data Understanding*)

Os dados provenientes dos quatro sistemas institucionais descritos na seção 3.2 foram integrados em um único conjunto analítico composto por 235.632 registros acadêmicos, referentes a 8.749 matrículas de graduação do *campus* Chapecó, distribuídos em 92 variáveis originais e cobrindo o período de 2014.1 a 2024.2. A etapa de compreensão dos dados envolveu a elaboração de um Relatório de Qualidade de Dados (DQR), com estatísticas

descritivas completas para cada variável, conforme a metodologia proposta por Kelleher, Mac Namee e D'Arcy (2015). A análise do DQR permitiu identificar padrões iniciais, problemas de qualidade, valores atípicos, variáveis com cardinalidade irregular e relações entre variáveis relevantes para a modelagem, conforme destacam Romero e Ventura (2020). Com base nesses achados, foi elaborado um Plano de Qualidade de Dados (DQP), documento que registra as estratégias de tratamento definidas para cada problema identificado e que orientou as decisões executadas na fase subsequente de preparação dos dados.

c) Preparação dos Dados (*Data Preparation*)

A fase de preparação dos dados compreendeu a execução das estratégias definidas no Plano de Qualidade de Dados, a codificação das variáveis e a engenharia de novos atributos, resultando na Analytics Base Table (ABT) utilizada na modelagem. As atividades de limpeza incluíram a remoção de variáveis sem variação ou redundantes, o tratamento de valores atípicos por *clamp transformation*, a imputação de valores ausentes pela mediana acompanhada da criação de indicadores de ausência (*missing indicators*) e a consolidação, em nível de matrícula e semestre, das variáveis provenientes dos sistemas SAS, SCI e *Pergamum*. Nessa fase também foi definida a variável alvo evasão, construída a partir da coluna status do sistema acadêmico, classificando como classe positiva exclusivamente os registros com encerramento definitivo do vínculo, procedimento detalhado na seção 4.4.1. A codificação das variáveis categóricas combinou *binary encoding*, para variáveis com dois valores distintos, e *label encoding*, para as demais, enquanto colunas de cardinalidade excessivamente alta foram substituídas por variáveis derivadas de maior valor analítico, como a *area_disciplina*, criada a partir do agrupamento dos códigos de disciplina em áreas de conhecimento (Han; Kamber; Pei, 2011; García; Luengo; Herrera, 2015). Por fim, a engenharia de variáveis consolidou informações distribuídas em múltiplos atributos e gerou representações com maior potencial preditivo, como a idade no ingresso e os indicadores semestrais de uso do Restaurante Universitário, da Biblioteca e do recebimento de auxílios socioeconômicos. O detalhamento completo dessas transformações, com os respectivos quadros de codificação, é apresentado nas seções 4.3 a 4.6.

d) Modelagem (*Modeling*)

Todos os modelos foram encapsulados em *pipelines* individuais, compostos por uma etapa de imputação pela mediana, destinada ao tratamento de valores ausentes residuais, e, para os algoritmos sensíveis à escala das variáveis, Regressão Logística, SVM Linear e KNN, por uma etapa adicional de normalização (StandardScaler). Ambas as etapas foram ajustadas exclusivamente sobre os dados de treino em cada *fold*, evitando a incorporação de informações do conjunto de teste durante o processo de ajuste. As estratégias principais de tratamento de valores ausentes, definidas no Plano de Qualidade de Dados, foram executadas na fase de preparação dos dados, com a criação de indicadores de ausência (*missing indicators*) preservados como variáveis preditoras. A avaliação foi realizada por validação cruzada estratificada com agrupamento por matrícula (StratifiedGroupKFold, $n=5$, $\text{shuffle}=\text{True}$, $\text{random_state}=42$), garantindo que todas as linhas de uma mesma matrícula permanecessem exclusivamente em treino ou em teste, evitando o vazamento de informação entre registros do mesmo estudante. Além disso, as colunas status, saida e situacao_sigla foram removidas do conjunto de variáveis por conterem informação direta ou indireta sobre o desfecho da evasão. Esses cuidados são coerentes com a discussão de Kelleher, Mac Namee e D'Arcy (2015) sobre a necessidade de estruturar os dados de forma compatível com a modelagem preditiva e de utilizar apenas informações efetivamente disponíveis no momento da predição.

e) Avaliação (*Evaluation*)

Os resultados obtidos pela validação cruzada estratificada com agrupamento por matrícula (StratifiedGroupKFold, 5 folds) indicaram que o XGBoost obteve o melhor desempenho entre os seis modelos avaliados, com AUC-ROC de 0,8878 e F1-score macro de 0,8037, seguido pela Regressão Logística (AUC-ROC = 0,8681), SVM Linear (0,8680), *Random Forest* (0,8587), Árvore de Decisão (0,8258) e KNN (0,7997). O XGBoost também apresentou o maior *recall* macro entre os modelos avaliados (0,8038), característica relevante para o contexto de alerta precoce discutido anteriormente. Esses valores situam-se na faixa de $\geq 0,70$ como modelo forte, conforme a escala interpretativa de Kelleher, Mac Namee e D'Arcy (2015), sendo consistentes com estudos similares de predição de evasão na literatura nacional e internacional (Romero; Ventura, 2020; Kalita et al., 2025).

f) Implantação (*Deployment*)

A fase de implantação, no contexto desta dissertação, compreendeu a documentação completa do modelo preditivo desenvolvido, incluindo os scripts de preparação dos dados, codificação de variáveis, engenharia de variáveis e treinamento dos modelos, todos estruturados de forma reproduzível e acompanhados de relatórios técnicos detalhados. Conforme Kelleher, Mac Namee e D'Arcy (2015), é fundamental registrar e documentar todas as transformações aplicadas ao conjunto de dados de treinamento, de modo que o mesmo *pipeline* possa ser replicado sobre novos dados no momento da implantação efetiva do modelo em ambiente institucional.

Experiências semelhantes demonstram que sistemas de alerta precoce baseados em mineração de dados são instrumentos valiosos para subsidiar políticas de permanência e decisões administrativas no ensino superior (Schmitt; Bernardi; Kantorski, 2019). A viabilidade técnica dessa integração foi demonstrada pelos resultados obtidos neste estudo, que evidenciam capacidade preditiva suficiente para uso institucional, com AUC-ROC de 0,8878 para o modelo XGBoost.

Com o objetivo de tornar essa correspondência explícita e facilitar a leitura do trabalho, o Quadro 2 sintetiza a rastreabilidade entre as seis fases do CRISP-DM, seus respectivos referenciais teóricos e metodológicos e as seções do Capítulo 4 e do Capítulo 5 em que cada fase é operacionalizada e discutida.

Quadro 2 - Rastreabilidade das fases do CRISP-DM na estrutura da dissertação

Fase do CRISP-DM	Referencial teórico	Descrição metodológica	Seção de resultados
a) Compreensão do Negócio	2.4.1	3.2.2 (a)	1.1 a 1.3
b) Compreensão dos Dados	2.4.2	3.2.2 (b)	4.1, 4.2
c) Preparação dos Dados	2.4.3	3.2.2 (c)	4.3, 4.4, 4.5, 4.6
d) Modelagem	2.4.4	3.2.2 (d)	4.7.1
e) Avaliação	2.4.5	3.2.2 (e)	4.7.2, 4.8, 4.9
f) Implantação	2.4.6	3.2.2 (f)	Capítulo 5 (PTT), Apêndice A

Fonte: elaborado pelo autor (2026).

4 APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

Conforme Kelleher, Mac Namee e D'Arcy (2015), a etapa que abrange as fases de apresentação, tratamento e análise dos dados tem como objetivos principais: familiarizar-se com as características das variáveis do conjunto de dados, identificar problemas de qualidade, corrigir eventuais inconsistências e registrar os problemas remanescentes em um plano de qualidade para tratamento posterior. Han, Kamber e Pei (2011) reforçam que a qualidade dos dados é um pré-requisito indispensável para a obtenção de modelos preditivos confiáveis, estimando que até 80% do esforço de um projeto de mineração de dados seja dedicado às etapas de preparação e limpeza dos dados. Os autores ressaltam que, embora o relatório de qualidade de dados seja uma coleção de estatísticas descritivas e visualizações, trata-se de uma ferramenta poderosa e central para a compreensão dos dados antes da modelagem.

Neste trabalho, o conjunto de dados analisado é composto por 235.632 registros acadêmicos referentes a 8.749 estudantes matriculados na Universidade Federal da Fronteira Sul (UFFS), *campus* Chapecó, distribuídos em 92 variáveis, provenientes de quatro sistemas institucionais da UFFS: SIGAA (Sistema Integrado de Gestão de Atividades Acadêmicas), SAS (Sistema de Análise Socioeconômica), SCI (Sistema de Credenciamento Institucional) e o *Pergamum*, sistema informatizado de gestão utilizado para gerenciar o acervo da biblioteca. O processo de exploração e tratamento seguiu como referência a metodologia proposta por Kelleher, Mac Namee e D'Arcy (2015), com a elaboração de um Relatório de Qualidade de Dados (DQR) e de um Plano de Qualidade de Dados (DQP).

4.1 RELATÓRIO DE QUALIDADE DE DADOS

O Relatório de Qualidade de Dados (DQR - *Data Quality Report*) é gerado a partir de estatísticas descritivas calculadas para cada variável do conjunto de dados. Para variáveis contínuas, o relatório inclui contagem de registros, percentual de valores ausentes, cardinalidade, mínimo, primeiro quartil (Q1), mediana, média, terceiro quartil (Q3), máximo e desvio padrão. Para variáveis categóricas, incluem-se percentual de ausentes, cardinalidade, moda e segunda moda com seus respectivos percentuais (Kelleher, Mac Namee e D'Arcy, 2015).

4.1.1 Variáveis contínuas

A Tabela 1 apresenta o DQR para as variáveis contínuas principais do conjunto de dados do SIGAA.

Tabela 1 - Relatório de qualidade de dados: variáveis contínuas

Variável	Contagem	Ausentes	Mínimo	Q1	Mediana	Média	Q3	Máximo	Desvio Padrão	Tipo de Alerta
media_final	208.865	11,4%	0,0	6,0	7,5	6,40	8,7	10,0	3,19	Valores Ausentes
numero_faltas	204.500	13,2%	0,0	0,0	4,0	10,27	12,0	720,0	16,19	Outlier e valores ausentes

Fonte: elaborado pelo autor (2026).

A variável *media_final* apresentou 11,4% de valores ausentes (26.767 registros), o que, segundo Kelleher, Mac Namee e D'Arcy (2015), configura um problema de qualidade decorrente de dados válidos, exigindo registro no plano de qualidade e tratamento posterior. A distribuição da variável é assimétrica à esquerda, com média (6,40) inferior à mediana (7,5), indicando concentração de notas acima da tendência central. Para *numero_faltas*, verificou-se valor máximo de 720, extremamente discrepante em relação ao terceiro quartil ($Q3 = 12,0$) e ao *threshold* superior calculado pelo método IQR ($upper = Q3 + 1,5 \times IQR = 30,0$), o que caracteriza um *outlier* a ser tratado. Também com a variável *numero_faltas* foi identificado 13,2% de valores ausentes (31.132), que será tratado posteriormente.

4.1.2 Variáveis categóricas

O Quadro 3 apresenta o relatório de qualidade das principais variáveis categóricas utilizadas no estudo. A coluna "Ausentes" indica o percentual de valores faltantes; "Card." representa a cardinalidade da variável, isto é, a quantidade de categorias distintas observadas; "Moda" corresponde à categoria mais frequente; "Moda %" indica a proporção dessa categoria no conjunto de dados; "2ª Moda" apresenta a segunda categoria mais frequente; e "Alerta" sintetiza possíveis pontos de atenção identificados na análise exploratória.

Quadro 3 - Relatório de qualidade de dados: variáveis categóricas

Variável	Ausentes	Card.	Moda	Moda %	2ª Moda	Alerta
situacao_sigla	0%	16	APR	61,3%	REPNF	OK
status	0%	5	CANCELADO	48,3%	CONCLUÍDO	OK
campus	0%	1	Chapecó	100%	—	Card=1
curso	0%	14	ADMINISTRAÇÃO	13,8%	PEDAGOGIA	OK
grau_academico	0%	2	BACHARELADO	56,2%	LICENCIATURA	OK
turno	0%	6	Noturno	47,8%	Matutino	OK
forma_ingresso	0%	15	ENEM	75,3%	TRANS. INTERNA	OK
tipo_cota	21,6%	19	L5	35,5%	L1	Valores Ausentes
genero	0%	2	F	55,2%	M	OK
estado_civil	0%	6	Solteiro(a)	66,9%	Não Informado	26,7% NI
etnia	0%	7	Branco	70,2%	Pardo	OK
uf_nascimento	0,15%	29	SC	54,3%	RS	OK
nacionalidade	0%	16	BRASILEIRA	95,9%	HAITIANA	OK
ingresso	0%	22	2016.1	10,7%	2015.1	OK
periodo	5,96%	25	2019.1	6,4%	2020.1	Valores Ausentes

Fonte: elaborado pelo autor (2026).

A análise da cardinalidade revelou que a variável *campus* possui apenas um valor distinto ("Chapecó"), indicando que o conjunto de dados é referente exclusivamente a este *campus*. Conforme Kelleher, Mac Namee e D'Arcy (2015), variáveis com cardinalidade igual a 1 devem ser removidas do conjunto analítico, pois não contribuem com informação preditiva. A variável *tipo_cota* apresentou 21,6% de ausentes, cujos registros, após investigação, foram interpretados como estudantes ingressantes pela modalidade de ampla concorrência. A variável *estado_civil*, apesar de não apresentar ausentes, possui 26,7% dos registros com o valor "Não Informado", que foi mantido como nível válido.

4.1.3 Colunas de uso do RU e sistema de análise socioeconômica (SAS)

O conjunto de dados continha 22 colunas referentes ao uso do Restaurante Universitário (RU) e 22 colunas relacionadas à participação no Sistema de Análise Socioeconômica (SAS), abrangendo os semestres de 2014.1 a 2024.2. As variáveis do RU registravam a quantidade acumulada de acessos de cada estudante ao restaurante em cada semestre. Já as variáveis do SAS indicavam, para cada período, se o estudante era ou não beneficiário de algum auxílio. As colunas ru_2014_1, ru_2020_2, ru_2021_1 e ru_2021_2 apresentaram cardinalidade igual a 1 (único valor: 0,0), indicando o semestre de início das atividades do RU na UFFS *Campus* Chapecó e da suspensão das atividades presenciais durante a pandemia de COVID-19.

As 22 colunas do SAS apresentaram 75,7% de valores ausentes em todos os semestres, excedendo o limite de 60% estabelecido por Kelleher, Mac Namee e D'Arcy (2015) como critério para remoção de variáveis. Adicionalmente, as colunas de 2014.1 a 2018.2 possuem cardinalidade igual a 1 (único valor: não beneficiário), valor esse justificável, pois o sistema SAS só foi implantado no final do ano de 2018.

4.1.4 Colunas de empréstimos da biblioteca

O sistema *Pergamum*, utilizado para gerenciamento do acervo da Biblioteca da UFFS, forneceu 22 colunas relacionadas ao total de empréstimos físicos por semestre para cada usuário, cobrindo o período de 2014.1 a 2024.2. O arquivo continha 21.809 registros de usuários, incluindo estudantes de todos os seis campi da UFFS, docentes e técnicos administrativos. Após o cruzamento com os dados do SIGAA referentes ao campus Chapecó, por matrícula e CPF, identificaram-se 4.500 matrículas de graduação com histórico de empréstimos, correspondentes a 51,4% das 8.749 matrículas do campus Chapecó no período analisado. As 4.249 matrículas restantes não possuíam registro no *Pergamum*, sendo essa condição interpretada como ausência de uso da biblioteca.

A Tabela 2 apresenta o relatório de qualidade das variáveis provenientes do sistema *Pergamum*. A coluna “Contagem” indica o número de registros válidos disponíveis para cada variável, enquanto a coluna “Sem registro/sem uso” representa o percentual de matrículas sem correspondência no *Pergamum* ou sem histórico de utilização da biblioteca. As colunas “Mín”, “Q1”, “Mediana”, “Média”, “Q3” e “Máx” descrevem a distribuição dos valores observados, correspondendo, respectivamente, ao valor mínimo, primeiro quartil, mediana,

média aritmética, terceiro quartil e valor máximo. Por fim, a coluna “Alerta” sintetiza possíveis pontos de atenção identificados na análise exploratória dos dados, especialmente situações em que a ausência de registro foi interpretada como ausência de uso do serviço, e não como erro de qualidade dos dados.

Tabela 2 - Relatório de qualidade de dados: variáveis do *Pergamum*

Variável	Contagem	Sem registro/sem uso	Mín	Q1	Mediana	Média	Q3	Máx	Alerta
bib_total_empr estimos	4.500	48,6%	1	5	24	74,1	89	1.278	Sem registro, interpretado como ausência de uso
bib_usou_algu ma vez bib	8.749	0%	0	—	—	0,51	—	1	OK

Fonte: elaborado pelo autor (2026).

A análise revelou que 4.500 matrículas, correspondentes a 51,4% das 8.749 matrículas analisadas, apresentaram histórico de empréstimos no *Pergamum*. As 4.249 matrículas restantes, equivalentes a 48,6%, não possuíam correspondência no sistema, sendo essa condição interpretada como ausência de uso da biblioteca, e não como erro de qualidade dos dados. Os empréstimos variaram de 1 a 1.278 entre as matrículas com histórico de uso, com distribuição assimétrica à direita, evidenciada pela diferença entre a média de 74,1 e a mediana de 24, indicando que a maioria realizou poucos empréstimos, enquanto um grupo menor fez uso intenso da biblioteca.

4.2 IDENTIFICAÇÃO DE PROBLEMAS DE QUALIDADE

A partir da análise dos DQR, três categorias de problemas de qualidade foram identificadas, conforme a taxonomia de Kelleher, Mac Namee e D'Arcy (2015): valores ausentes, cardinalidade irregular e *outliers*.

4.2.1 Valores ausentes

Kelleher, Mac Namee e D'Arcy (2015) orientam que, quando o percentual de valores ausentes de uma variável supera 60%, a quantidade de informação remanescente é tão reduzida que a remoção da variável é recomendada. Neste estudo, as 22 variáveis do SAS ultrapassaram o limiar de 75,7%, sendo candidatas à remoção. As variáveis *media_final*

(11,4%), *numero_faltas* (13,2%), *tipo_cota* (21,6%) e *periodo* (5,96%) apresentaram valores ausentes tratáveis por imputação ou criação de *missing indicators*. No caso da variável *bib_total_emprestimos*, a ausência de correspondência no Pergamum foi interpretada como ausência de uso da biblioteca, e não como valor faltante.

4.2.2 Cardinalidade irregular

A variável *campus* apresentou cardinalidade 1, configurando uma variável sem variação e, portanto, sem qualquer poder preditivo. As colunas *ru_2014_1*, *ru_2020_2*, *ru_2021_1* e *ru_2021_2* também apresentaram cardinalidade 1 entre os valores não-ausentes. Kelleher, Mac Namee e D'Arcy (2015) recomendam investigar se o problema decorre de erro na geração do conjunto analítico.

4.2.3 Outliers

A variável '*numero_faltas*' apresentou valor máximo de 720, superior ao limite de 30, calculado pelo método do Intervalo Interquartil (IQR), caracterizando a presença de valores extremos (*outliers*). Apesar disso, optou-se por manter tais observações, uma vez que podem representar situações reais associadas ao histórico acadêmico dos estudantes e fornecer informações relevantes para o modelo preditivo. Han, Kamber e Pei (2011) definem *outliers* como observações que se distanciam do comportamento geral dos dados, podendo representar ruídos, erros ou ocorrências relevantes do fenômeno analisado. No presente estudo, optou-se por tratar valores extremos por meio da transformação *clamp*, procedimento discutido por Kelleher, Mac Namee e D'Arcy (2015), no qual valores abaixo de um limite inferior ou acima de um limite superior são substituídos pelos respectivos limites definidos.

Para a definição desses limites, adotou-se o método baseado no intervalo interquartil (IQR), calculado pelas expressões $lower = Q1 - 1,5 \times IQR$ e $upper = Q3 + 1,5 \times IQR$. Esse procedimento foi escolhido por ser menos sensível à influência de valores extremos do que métodos baseados na média e no desvio padrão, especialmente em distribuições assimétricas. A opção pelo tratamento dos *outliers* integra a etapa de pré-processamento dos dados, cuja importância para a qualidade dos modelos de mineração de dados é destacada por García, Luengo e Herrera (2015).

4.3 PLANO DE QUALIDADE DE DADOS

Conforme Kelleher, Mac Namee e D'Arcy (2015), os problemas de qualidade decorrentes de dados válidos devem ser registrados em um Plano de Qualidade de Dados (DQP), junto com as estratégias de tratamento propostas. O Quadro 4 consolida todos os problemas identificados e as respectivas estratégias adotadas neste trabalho.

Quadro 4 - Plano de qualidade de dados

Variável	Problema identificado	Tipo	Estratégia adotada
campus	Cardinalidade = 1 (único valor: Chapecó)	Dados válidos	Removida do conjunto. Sem variação, sem poder preditivo
situacao_descricao	Duplicata textual de situacao_sigla	Redundante	Removida. Mantida apenas situacao_sigla
numero_faltas	<i>Outlier</i> : máx = 720; upper IQR = 30	Dados válidos	<i>Clamp transformation</i> : lower=0, upper=30 (1.615 valores ajustados)
media_final	11,4% de valores ausentes	Dados válidos	<i>Missing indicator</i> criado; imputação pela mediana (7,5)
numero_faltas	13,2% de valores ausentes	Dados válidos	<i>Missing indicator</i> criado; imputação pela mediana (4,0)
tipo_cota	21,6% de valores ausentes	Dados válidos	Ausência interpretada como ampla concorrência; preenchido com AC (ampla concorrência)
periodo	5,96% de valores ausentes	Dados válidos	<i>Missing indicator</i> criado; imputação pela moda (2019.1)
ru_2014_1, ru_2020_2, ru_2021_1, ru_2021_2	Cardinalidade = 1 (único valor: 0.0)	Dados válidos	Removidas. Contexto pandemia e início do funcionamento do RU
Colunas SAS (22 semestres)	75,7% ausentes; 10 delas com cardinalidade = 1	Crítico	Criadas variáveis agregadas: sae_beneficiario, sae_n_semestres; 22 colunas originais removidas
Colunas da biblioteca. Sistema Pergamum (22 semestres)	48,6% dos estudantes sem registro no Pergamum; arquivo inclui discentes, docentes e TAEs de outros campi	Dados válidos	Cruzamento por matrícula (4.368) e CPF (132); 4.249 estudantes sem correspondência receberam valor 0; criadas variáveis bib_total_emprestimos e bib_usou_alguna_vez_bib; 22 colunas originais removidas
data_nascimento	Data bruta (card=4.849)	Engenharia	Convertida para idade_no_ingresso; coluna original removida
id_cpf, id_matricula	Identificadores únicos	Engenharia	Separados do ABT; não usados como variáveis preditivas

Fonte: adaptado de Kelleher, Mac Namee e D'Arcy (2015).

4.3.1 Tratamento dos dados

Com base no DQP, foram aplicados os seguintes tratamentos, implementados em Python com as bibliotecas pandas e numpy:

4.3.1.1 Remoção de variáveis sem variação

As variáveis campus, ru_2014_1, ru_2020_2, ru_2021_1, ru_2021_2 e as 10 colunas do SAS de 2014.1 a 2018.2 foram removidas por apresentarem cardinalidade igual a 1 entre os valores não-ausentes. Adicionalmente, a variável situacao_descricao foi removida por ser duplicata textual de situacao_sigla, eliminando redundância no conjunto analítico.

4.3.1.2 Consolidação das variáveis do sistema de análise socioeconômica

As 12 colunas do SAS remanescentes (2019.1 a 2024.2), apesar de excederem o limiar de 60% de ausentes, continham informação relevante sobre participação em programas assistenciais. Para preservar essa informação, foram criadas duas variáveis agregadas: sae_beneficiario, variável binária indicando se o estudante foi beneficiário em algum auxílio em ao menos um semestre (valor 1 para 1.276 estudantes únicos; 14,6% do total), e sae_n_semestres, que contabiliza o número de semestres em que o estudante foi beneficiário de algum auxílio durante os três primeiros semestres de cada matrícula. A restrição temporal à janela inicial de três semestres aplica-se pelo princípio de prevenção ao vazamento de dados temporais: utilizar o total acumulado ao longo de toda a trajetória incorporaria ao modelo informações sobre semestres posteriores ao momento em que a predição seria realizada na prática, comprometendo a validade do modelo para uso prospectivo (Kelleher, Mac Namee e D'Arcy, 2015). Ambas as variáveis foram calculadas por estudante único, evitando duplicação decorrente da estrutura do *dataset*, em que cada estudante aparece repetido uma vez por disciplina cursada. As 12 colunas originais foram então removidas.

4.3.1.3 Tratamento de outliers

Aplicou-se a *clamp transformation* à variável numero_faltas, com *thresholds* calculados pelo método IQR: $lower = \max(0, Q1 - 1,5 \times IQR) = 0,0$ e $upper = Q3 + 1,5 \times IQR = 30,0$. Um total de 1.615 registros tiveram seus valores limitados ao *threshold* superior.

Conforme Kelleher, Mac Namee e D'Arcy (2015), essa transformação é recomendada quando há suspeita de que *outliers* estão comprometendo o desempenho do modelo, devendo-se avaliar seu impacto por comparação de desempenho.

4.3.1.4 Imputação de valores ausentes

Para as variáveis `media_final`, `numero_faltas` e `periodo` foram criadas variáveis binárias indicadoras de ausência (*missing indicators*), preservando a informação sobre a ausência antes da imputação. Conforme Kelleher, Mac Namee e D'Arcy (2015), a imputação pela mediana é a abordagem mais comum para variáveis contínuas, por ser mais robusta que a média na presença de distribuições assimétricas. García, Luengo e Herrera (2015) corroboram essa recomendação, destacando que métodos de imputação baseados em medidas de tendência central são amplamente utilizados na prática por sua simplicidade e eficácia em datasets com baixo percentual de ausentes. Os valores ausentes de `media_final` foram substituídos pela mediana (7,5) e os de `numero_faltas` pela mediana (4,0). Para a variável `tipo_cota` (21,6% ausentes), os valores foram preenchidos com o nível AC "ampla concorrência", interpretação confirmada pela análise do *dataset*. Já para a variável `periodo`, por representar um semestre letivo e possuir natureza temporal ordinal/categórica, optou-se pela imputação pela moda, mantendo-se um valor válido e recorrente no conjunto de dados. Dessa forma, os valores ausentes de `periodo` foram substituídos por 2019.1, valor modal da variável.

4.3.1.5 Consolidação das variáveis do RU

Analogamente ao tratamento das variáveis do SAS, criaram-se duas variáveis agregadas a partir das colunas de uso do Restaurante Universitário: `ru_total_refeicoes`, que soma o total de refeições utilizadas durante os três primeiros semestres de cada matrícula (média de 69,7 refeições por estudante), e `ru_usou_alguma_vez`, variável binária indicando uso do RU em pelo menos um semestre (58,7% dos estudantes). Ambas as variáveis foram calculadas por estudante único e propagadas ao dataset via mapeamento. As 18 colunas RU originais foram então removidas, mantendo apenas as duas variáveis agregadas.

4.3.1.6 Consolidação das variáveis da biblioteca

Os dados de empréstimos provenientes do *Pergamum* foram consolidados em duas variáveis agregadas por estudante: *bib_total_emprestimos*, que representa o total de empréstimos realizados pelo estudante na biblioteca da UFFS durante os três primeiros semestres de cada matrícula, e *bib_usou_alguma_vez_bib*, variável binária indicando uso da biblioteca em ao menos um semestre. O cruzamento com o SIGAA foi realizado em duas etapas: primeiramente por matrícula (4.368 estudantes identificados) e, complementarmente, por CPF para os casos em que a matrícula não correspondeu (132 estudantes recuperados), totalizando 4.500 estudantes com histórico de empréstimos. Os 4.249 estudantes sem correspondência no *Pergamum* receberam valor 0, interpretado como ausência de uso do serviço. Ambas as variáveis foram calculadas por estudante único e propagadas ao dataset via mapeamento, seguindo a mesma abordagem adotada para as variáveis do RU e do SAS (Kelleher, Mac Namee e D’Arcy, 2015). As 22 colunas originais do *Pergamum* foram então removidas do conjunto analítico.

4.4 CODIFICAÇÃO E ENGENHARIA DE VARIÁVEIS

A conversão de variáveis categóricas em representações numéricas é uma etapa indispensável na preparação de dados para algoritmos de *Machine Learning*, uma vez que a maioria dos modelos preditivos opera exclusivamente sobre valores numéricos (Han; Kamber; Pei, 2011). Segundo García, Luengo e Herrera (2015), o pré-processamento adequado das variáveis é determinante para a qualidade dos modelos resultantes, podendo impactar significativamente o desempenho preditivo.

4.4.1 Definição da variável alvo

A variável alvo *evasao* foi criada a partir da coluna *status* do sistema acadêmico da UFFS. Conforme a literatura de *Educational Data Mining* (Baker; Yacef, 2009), a definição precisa do alvo de predição é fundamental para a validade do modelo. Optou-se por classificar como evasão exclusivamente os registros com status CANCELADO, que representa o encerramento definitivo do vínculo acadêmico. Registros com status TRANCADO foram excluídos da classe positiva por representarem uma situação temporária, passível de reversão. Incluí-los como evasão introduziria falsos positivos no modelo, prejudicando sua capacidade

preditiva. O Quadro 5 apresenta a codificação adotada para a variável-alvo, indicando a correspondência entre os valores originais da coluna status e as classes binárias utilizadas no modelo. O status CANCELADO foi codificado como 1, representando a classe positiva de evasão, enquanto os demais status foram codificados como 0, correspondendo à classe de não evasão.

Quadro 5 - Variável alvo

Valor original	Código
CANCELADO	1 - evasão (classe positiva)
CONCLUÍDO	0 - não evasão
FORMADO	0 - não evasão
ATIVO - FORMANDO	0 - não evasão
TRANCADO	0 - não evasão

Fonte: elaborado pelo autor (2026).

A distribuição resultante foi de 48,3% de evasão (113.738 registros) e 51,7% de não evasão (121.894 registros), caracterizando um *dataset* com classes relativamente balanceadas, favorável ao treinamento dos modelos.

4.5 ESTRATÉGIAS DE CODIFICAÇÃO ADOTADAS

As estratégias de codificação adotadas neste estudo tiveram como finalidade tornar o *Analytics Base Table* (ABT) compatível com os algoritmos de aprendizado de máquina, preservando, ao mesmo tempo, a coerência analítica das informações originais. Kelleher, Mac Namee e D'Arcy (2015) destacam que a preparação dos dados é uma etapa central do processo de modelagem preditiva, pois as estratégias aplicadas às variáveis influenciam diretamente a validade dos modelos e devem ser realizadas de forma consistente ao longo de todo o fluxo analítico. Nessa perspectiva, a codificação das variáveis categóricas e a criação de atributos derivados buscaram adequar o conjunto de dados às exigências técnicas dos algoritmos, sem perder de vista a fidelidade ao fenômeno investigado.

4.5.1 Remoção de colunas de alta cardinalidade

As colunas `codigo_disciplina` (1.472 valores únicos), `nome_disciplina` (1.148) e `cidade_nascimento` (907) foram removidas do *dataset*. Conforme Han, Kamber e Pei (2011), variáveis com cardinalidade excessivamente alta tendem a introduzir ruído e prejudicar o desempenho dos modelos, além de tornar o *label encoding* inadequado, pois atribui uma ordem numérica artificial a categorias sem relação ordinal. Em substituição, foram criadas três novas variáveis que preservam a informação em nível de abstração adequado para a modelagem, conforme descrito na seção 4.6.

4.5.2 Binary encoding

O procedimento foi aplicado às variáveis `sexo` e `grau_academico`, por apresentarem exatamente dois valores distintos em cada caso. Na variável `sexo`, a categoria F (Feminino) foi codificada como 0 e M (Masculino) como 1. Já na variável `grau_academico`, BACHARELADO foi representado por 1 e LICENCIATURA por 0. Essa escolha é coerente com a discussão de Kelleher, Mac Namee e D'Arcy (2015) sobre a transformação de variáveis categóricas em variáveis binárias 0/1, de modo a torná-las compatíveis com a modelagem preditiva.

4.5.3 Label encoding

Aplicado às demais colunas categóricas com múltiplos valores. Os valores foram ordenados alfabeticamente e codificados com inteiros sequenciais a partir de 0. García, Luengo e Herrera (2015) ressaltam que o *label encoding* é adequado quando o algoritmo utilizado não pressupõe uma relação de magnitude entre os códigos, como é o caso das árvores de decisão e do *Random Forest*, que trabalham com particionamentos binários das variáveis, tornando a ordem dos rótulos irrelevante.

O processo de codificação das variáveis categóricas consistiu na substituição dos valores textuais por códigos numéricos, de modo a permitir sua utilização nos algoritmos de aprendizado de máquina. A variável `curso` foi codificada de 0 a 13, contemplando os cursos de Administração, Agronomia, Ciência da Computação, Ciências Sociais, Educação Especial Inclusiva, Enfermagem, Engenharia Ambiental e Sanitária, Filosofia, Geografia, História, Letras - Português e Espanhol, Matemática, Medicina e Pedagogia.

A variável turno recebeu códigos de 0 a 5, correspondentes às categorias Matutino, Matutino e Vespertino, Vespertino e Noturno, Noturno, Vespertino e Vespertino e Noturno. A variável forma_ingresso foi codificada de 0 a 14, incluindo as modalidades, Convênio PEC-G, ENEM, Judicial, PARFOR, PIN, Pró-Haiti, Processo Seletivo Simplificado, Pró-Imigrante, Retorno de Aluno-Abandono, Retorno de Graduado, SiSU, Transferência Compulsória (Ex Officio), Transferência Externa, Transferência Interna e Vestibular. Já a variável tipo_cota foi codificada de 0 a 18, abrangendo as categorias A1, A2, A2-I, AC, L1, L10, L13, L14, L2, L5, L6, L9, LB_EP, LB_PCD, LB_PPI, LB_Q, LI_EP, LI_PCD e LI_PPI.

Também foram codificadas variáveis de caracterização sociodemográfica. A variável estado_civil recebeu códigos de 0 a 5, contemplando Casado(a), Divorciado(a), Não Informado, Separado Judicialmente, Solteiro(a) e Viúvo(a). A variável etnia foi codificada de 0 a 6, incluindo Amarelo, Branco, Indígena, Negro, Não Informado, Pardo e Remanescente de quilombo. A variável uf_nascimento foi codificada de 0 a 28, abrangendo a categoria estrangeiro (“--”) e as unidades federativas AC, AL, AM, AP, BA, CE, DF, ES, GO, IG, MA, MG, MS, MT, PA, PB, PE, PI, PR, RJ, RN, RO, RR, RS, SC, SE, SP e TO. Por fim, as variáveis nacionalidade e pais_nascimento também foram transformadas em códigos numéricos, contemplando, respectivamente, as nacionalidades e países na base de dados.

4.6 NOVAS VARIÁVEIS CRIADAS POR ENGENHARIA

A criação de novas variáveis por meio da engenharia de atributos teve como objetivo ampliar a capacidade informativa do *Analytics Base Table* (ABT), transformando variáveis originais em representações mais adequadas à modelagem preditiva. Conforme discutem Kelleher, Mac Namee e D'Arcy (2015), a preparação dos dados não se restringe às etapas de limpeza e codificação, abrangendo também a construção de atributos que representem de maneira mais apropriada o fenômeno estudado e favoreçam a aplicação de algoritmos de aprendizado de máquina. Nesse contexto, as novas variáveis desenvolvidas neste estudo buscaram reduzir a cardinalidade de determinadas variáveis, consolidar informações distribuídas em múltiplos atributos e gerar representações com maior potencial preditivo para a identificação da evasão discente.

4.6.1 Área de conhecimento: variável `area_disciplina`

Em substituição às colunas `codigo_disciplina` e `nome_disciplina`, foi criada a variável `area_disciplina`, a partir do agrupamento dos prefixos dos códigos de disciplina da UFFS. Considerando que o prefixo alfanumérico identifica a área de conhecimento da disciplina, essa transformação possibilitou reduzir a cardinalidade da variável de 1.472 para 10 categorias, conforme o Quadro 6, preservando uma representação mais agregada e analiticamente útil da estrutura curricular. Tal procedimento é compatível com a abordagem de Kelleher, Mac Namee e D’Arcy (2015), ao privilegiar a construção de variáveis derivadas mais adequadas à modelagem preditiva, com redução de complexidade e manutenção de informação relevante.

Quadro 6 - Área de conhecimento

Valor original	Código
GAC — Adaptação Curricular	0
GCA — Ciências Agrárias	1
GCB — Ciências Biológicas	2
GEX — Ciências Exatas	3
GCH — Ciências Humanas	4
GCS — Ciências Sociais	5
GEN — Engenharias	6
GLA — Linguagens	7
GSA — Saúde	8
Outros	9

Fonte: elaborado pelo autor (2026).

4.6.2 Região geográfica de nascimento

Em substituição parcial à coluna `cidade_nascimento`, foi criada a variável `regiao_nascimento`, conforme o Quadro 7, agrupando a UF de nascimento de cada estudante nas cinco macrorregiões brasileiras, acrescidas das categorias Exterior e Não Informado. Essa abordagem reduz a cardinalidade de 29 categorias para 7 regiões, mantendo a informação geográfica relevante para a análise de perfil dos estudantes da UFFS.

Quadro 7 - Região de nascimento

Valor original	Código
Centro-Oeste (GO, MT, MS, DF)	0
Exterior (IG)	1
Não Informado (--)	2
Nordeste (BA, SE, AL, PE, PB, RN, CE, PI, MA)	3
Norte (AM, PA, AC, RO, RR, AP, TO)	4
Sudeste (SP, RJ, MG, ES)	5
Sul (SC, RS, PR)	6

Fonte: elaborado pelo autor (2026).

Foi criada também a variável *distancia_chapeco_km*, que representa a distância geodésica em quilômetros entre o centróide de cada UF de nascimento e a cidade de Chapecó-SC, sede do *campus* da UFFS onde os dados foram coletados. O cálculo foi realizado por meio da fórmula de Haversine (Sinnott, 1984), que computa a distância entre dois pontos na superfície esférica da Terra a partir de suas coordenadas geográficas de latitude e longitude, conforme a equação:

$$d = 2R \cdot \arctan2(\sqrt{a}, \sqrt{1-a})$$

onde $a = \sin^2(\Delta lat/2) + \cos(lat_1) \cdot \cos(lat_2) \cdot \sin^2(\Delta lon/2)$ e $R = 6.371 \text{ km}$

As coordenadas de referência utilizadas foram latitude $-27,1009343^\circ$ e longitude $-52,615699^\circ$ para Chapecó-SC. Para cada UF, adotou-se o centróide geográfico aproximado da capital estadual como ponto representativo. Estudantes nascidos no exterior receberam o valor fixo de 5.000 km, representando uma distância internacional simbólica. Registros sem informação de UF receberam a mediana geral. Os valores resultantes variaram de 238 km (SC) a 5.000 km (exterior).

4.6.3 Engenharia de variável - idade no ingresso

A variável *data_nascimento*, com cardinalidade de 4.849 valores distintos, foi convertida na variável *idade_no_ingresso*, calculada como a diferença entre o ano de ingresso do estudante e o ano de nascimento. A variável resultante, calculada por estudante único e propagada ao dataset via mapeamento, apresentou média de 23,7 anos, mediana de 21 anos e amplitude de 16 a 66 anos, compatível com o perfil típico de estudantes universitários.

4.6.4 Visão geral das transformações

O Quadro 8 apresenta o resumo completo de todas as transformações e variáveis criadas no processo de preparação do dataset para modelagem.

Quadro 8 - Resumo das transformações de codificação e engenharia de variáveis

Coluna original	Estratégia	Card.	Resultado / mapeamento
codigo_disciplina	Removida	1.472	Substituída por area_disciplina
nome_disciplina	Removida	1.148	Substituída por area_disciplina
cidade_nascimento	Removida	907	Substituída por regioao_nascimento e distancia_chapeco_km
genero	Binary encoding	2	F=0, M=1
grau_academico	Binary encoding	2	BACHARELADO=1, LICENCIATURA=0
situacao_sigla	Removida	16	removida por risco de vazamento de informação sobre desempenho/desfecho
status	Removida	5	Usada apenas para criar a variável alvo evasao; removida do conjunto preditor
curso	Label encoding	14	ADMINISTRAÇÃO=0 ... PEDAGOGIA=13
turno	Label encoding	6	Matutino=0 ... Vespertino e Noturno=5
forma_ingresso	Label encoding	15	CONVÊNIO PEC-G=0 ... VESTIBULAR=14
tipo_cota	Label encoding	19	A1=0 ... LI_PPI=18
estado_civil	Label encoding	6	Casado(a)=0 ... Viúvo(a)=5
etnia	Label encoding	7	Amarelo=0 ... Remanescente quilombo=6
uf_nascimento	Label encoding	29	--=0, AC=1 ... TO=28
nacionalidade	Label encoding	16	ARGENTINA=0 ... VENEZUELANA=15
pais_nascimento	Label encoding	13	ARGENTINA=0 ... VENEZUELA=12
area_disciplina	Nova variável	10	Agrupamento por prefixo do código (GCH/GEX/etc.)
regiao_nascimento	Nova variável	7	Sul=6, Sudeste=5, Norte=4, Nordeste=3, C.Oeste=0, Exterior=1
distancia_chapeco_km	Nova variável	28	Distância Haversine de cada UF até Chapecó-SC (km)
evasao	Variável alvo	2	CANCELADO=1 (evasão), demais=0
saida	Removida	2	Removida por vazamento

Fonte: elaborado pelo autor (2026).

4.7 RESULTADOS DA MODELAGEM PREDITIVA

Os modelos preditivos foram treinados e avaliados sobre o *dataset* da Universidade Federal da Fronteira Sul (UFFS), campus Chapecó, composto por 235.632 registros acadêmicos referentes a 8.749 matrículas de graduação, cobrindo o período de 2014.1 a 2024.2. A base final de modelagem foi composta por 28 variáveis preditoras e uma variável alvo, denominada evasão. Essa variável foi definida como 1 para registros associados a matrículas canceladas, totalizando 113.738 registros (48,3%), e como 0 para os demais status acadêmicos, totalizando 121.894 registros (51,7%).

As 28 variáveis preditoras utilizadas na modelagem foram: *media_final*, *numero_faltas*, *bib_total_emprestimos*, *ru_total_refeicoes*, *bib_usou_alguna_vez*, *ru_usou_alguna_vez*, *media_final_ausente*, *numero_faltas_ausente*, *ingresso*, *idade_no_ingresso*, *turno*, *genero*, *tipo_cota*, *area_disciplina*, *estado_civil*, *etnia*, *grau_academico*, *distancia_chapeco_km*, *regiao_nascimento*, *uf_nascimento*, *pais_nascimento*, *forma_ingresso*, *nacionalidade*, *curso*, *periodo*, *periodo_ausente*, *sae_beneficiario* e *sae_n_semestres*. A variável *evasao* foi utilizada exclusivamente como variável alvo, não integrando o conjunto de variáveis preditoras.

Foram aplicados seis algoritmos de classificação supervisionada: Regressão Logística (RL), Árvore de Decisão (AD), *Random Forest* (RF), XGBoost, Máquina de Vetores de Suporte Linear (SVM Linear) e K-Vizinhos Mais Próximos (KNN). A avaliação foi realizada por validação cruzada estratificada com agrupamento por matrícula (*StratifiedGroupKFold*, *n_splits=5*, *shuffle=True*, *random_state=42*), garantindo que todas as linhas de uma mesma matrícula ficassem exclusivamente em treino ou em teste, nunca nos dois simultaneamente. A imputação de valores ausentes e a normalização foram aplicadas dentro do *pipeline* de cada modelo, calculadas exclusivamente sobre os dados de treino de cada *fold*.

O Quadro 9 apresenta a configuração dos modelos preditivos avaliados no estudo, incluindo as etapas de pré-processamento aplicadas e os principais hiperparâmetros utilizados em cada algoritmo. A coluna "Modelo" identifica o classificador supervisionado empregado, enquanto a coluna "Pipeline de pré-processamento e hiperparâmetros" descreve a sequência de procedimentos executados antes e durante a modelagem. O termo *SimpleImputer(median)* indica que valores ausentes foram imputados pela mediana de cada variável numérica, calculada exclusivamente sobre o conjunto de treino em cada dobra e aplicada ao conjunto de teste, evitando vazamento de informação. O *StandardScaler* corresponde à padronização das variáveis, utilizada nos modelos sensíveis à escala dos dados, como Regressão Logística,

SVM Linear e KNN. Na sequência, são apresentados os algoritmos aplicados e seus respectivos hiperparâmetros, entre os quais profundidade máxima das árvores, número de estimadores, parâmetros de regularização, taxa de aprendizado, número de vizinhos e número máximo de iterações. Essa organização permite garantir maior transparência e reprodutibilidade ao processo de modelagem preditiva.

Quadro 9 - Configuração dos modelos

Modelo	Pipeline de pré-processamento e hiperparâmetros
Regressão Logística	SimpleImputer(median) → StandardScaler → LogisticRegression(C=0.5, class_weight='balanced', max_iter=1000, random_state=42)
Árvore de Decisão	SimpleImputer(median) → DecisionTreeClassifier(max_depth=5, min_samples_leaf=100, min_samples_split=200, ccp_alpha=0.0001, class_weight='balanced', random_state=42)
Random Forest	SimpleImputer(median) → RandomForestClassifier(n_estimators=200, max_depth=7, min_samples_leaf=50, min_samples_split=100, max_samples=0.8, oob_score=True, class_weight='balanced', n_jobs=-1, random_state=42)
XGBoost	SimpleImputer(median) → XGBClassifier(n_estimators=200, max_depth=4, learning_rate=0.1, min_child_weight=5, subsample=0.8, colsample_bytree=0.8, reg_lambda=2.0, scale_pos_weight=dinâmico, n_jobs=-1, random_state=42, eval_metric='logloss')
SVM (Linear)	SimpleImputer(median) → StandardScaler → CalibratedClassifierCV(LinearSVC(C=0.5, class_weight='balanced', max_iter=2000, random_state=42), cv=3, method='sigmoid')
KNN	SimpleImputer(median) → StandardScaler → KNeighborsClassifier(n_neighbors=31, weights='uniform', metric='manhattan', n_jobs=-1)

Fonte: elaborado pelo autor (2026).

4.7.1 Resultados dos modelos

A Tabela 3 apresenta o desempenho dos seis modelos avaliados, com destaque para os maiores valores de cada métrica.

Tabela 3 - Desempenho dos modelos preditivos

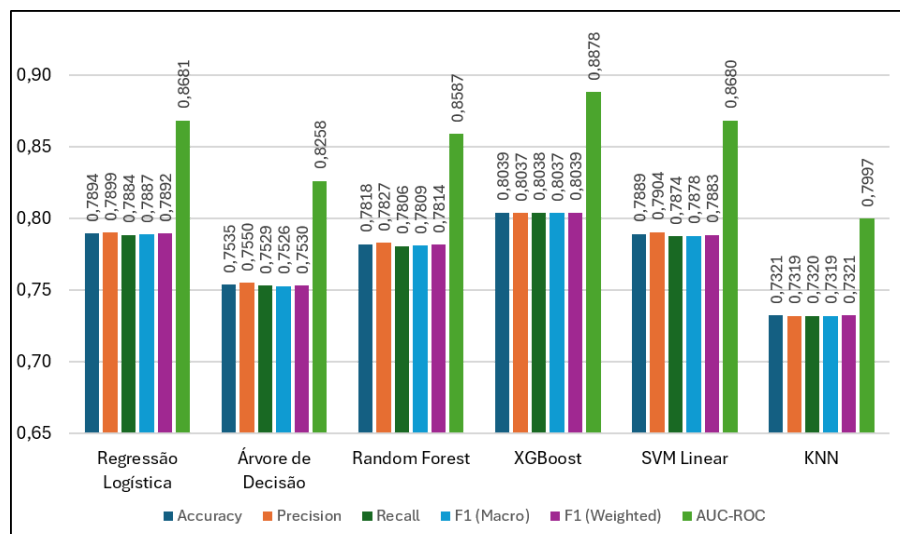
Modelo	<i>Accuracy</i>	<i>Precision</i> (Macro)	<i>Recall</i> (Macro)	F1 (Macro)	F1 (<i>Weighted</i>)	AUC-ROC
Regressão Logística	0,7894	0,7899	0,7884	0,7887	0,7892	0,8681
Árvore de Decisão	0,7535	0,7550	0,7529	0,7526	0,7530	0,8258
<i>Random Forest</i>	0,7818	0,7827	0,7806	0,7809	0,7814	0,8587
XGBoost	0,8039	0,8037	0,8038	0,8037	0,8039	0,8878
SVM (Linear)	0,7889	0,7904	0,7874	0,7878	0,7883	0,8680
KNN	0,7321	0,7319	0,7320	0,7319	0,7321	0,7997

Fonte: elaborado pelo autor (2026).

O XGBoost destacou-se como o melhor modelo em todas as métricas avaliadas, com AUC-ROC de 0,8878, Accuracy de 0,8039 e F1-score macro de 0,8037. A Regressão Logística obteve o segundo melhor desempenho em termos de AUC-ROC (0,8681), seguida de perto pelo SVM Linear (0,8680) e pelo *Random Forest* (0,8587). A Árvore de Decisão apresentou desempenho intermediário (AUC-ROC = 0,8258), enquanto o KNN registrou o menor desempenho entre os modelos avaliados (AUC-ROC = 0,7997), resultado esperado para datasets com elevada dimensionalidade, nos quais a distância euclidiana perde eficácia discriminativa, fenômeno conhecido como maldição da dimensionalidade (Hastie; Tibshirani; Friedman, 2009).

A Figura 2 apresenta o comparativo de todas as métricas para os seis modelos.

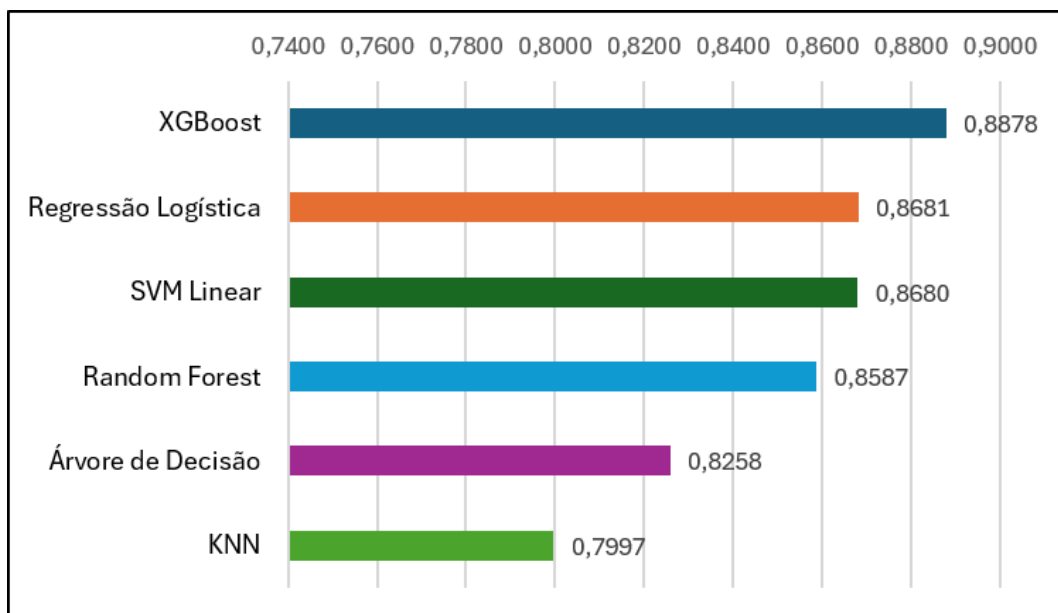
Figura 2 - Comparativo de desempenho dos 6 modelos



Fonte: elaborado pelo autor (2026).

A Figura 3 apresenta o ranking dos modelos de classificação com base nos valores médios de AUC-ROC obtidos durante a validação cruzada. Esse gráfico permite uma visualização mais direta do desempenho relativo dos algoritmos.

Figura 3 - Ranking por AUC-ROC



Fonte: elaborado pelo autor (2026).

4.7.2 Matriz de confusão do modelo XGBoost

Considerando o desempenho apresentado pelos modelos avaliados, procedeu-se a uma análise mais detalhada do XGBoost, que obteve os melhores resultados globais entre os seis algoritmos testados. Embora métricas como acurácia, precisão, *recall*, F1-score e AUC-ROC permitam sintetizar diferentes dimensões da capacidade preditiva de um classificador, a análise da matriz de confusão possibilita observar diretamente a distribuição dos acertos e dos diferentes tipos de erro produzidos pelo modelo. Conforme Kelleher, Mac Namee e D'Arcy (2015), essa matriz relaciona as classes reais às classes previstas, permitindo identificar verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos.

Essa distinção assume especial importância no contexto da predição de evasão acadêmica, uma vez que os diferentes tipos de erro não possuem necessariamente o mesmo impacto para a gestão institucional. Hastie, Tibshirani e Friedman (2009) destacam que a avaliação de modelos de classificação deve considerar não apenas seu desempenho agregado, mas também a natureza e o custo dos erros cometidos. Em sistemas voltados à identificação

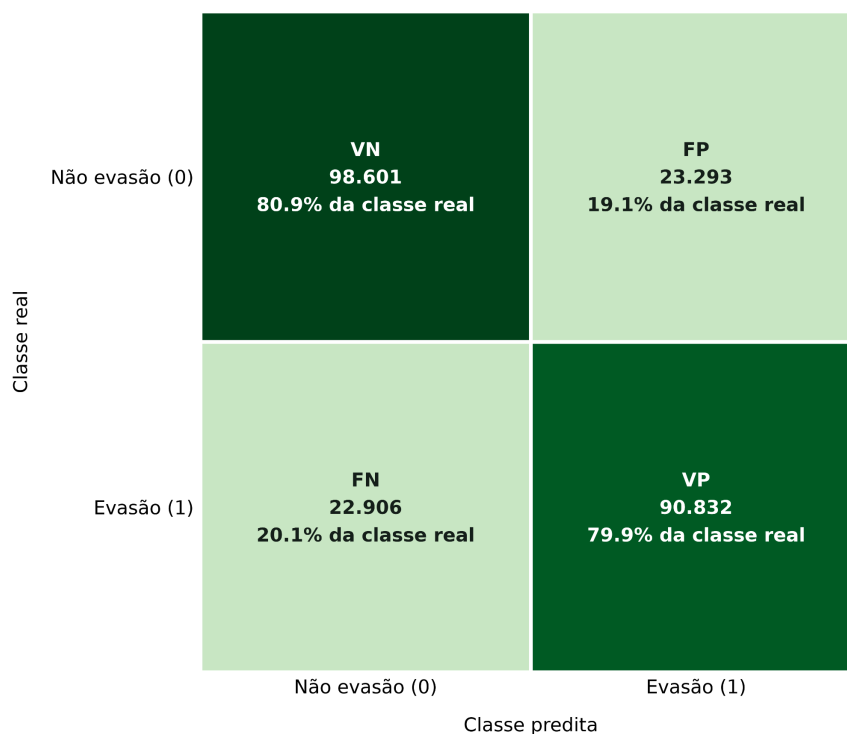
precoce de estudantes em risco, essa questão torna-se particularmente relevante, pois a classificação incorreta pode afetar diretamente a capacidade institucional de direcionar ações de acompanhamento e permanência.

Nesse sentido, em aplicações de Mineração de Dados Educacionais orientadas à intervenção, os falsos negativos merecem atenção especial, uma vez que correspondem a situações em que estudantes efetivamente pertencentes à classe de risco não são identificados pelo modelo. Por outro lado, os falsos positivos representam estudantes classificados como potenciais casos de evasão sem que esse desfecho efetivamente ocorra, podendo resultar na mobilização de recursos institucionais para casos que não demandariam intervenção. Essa assimetria entre os custos dos erros, amplamente discutida na literatura de análise ROC, em que a escolha do ponto de operação de um classificador depende explicitamente do custo relativo atribuído a falsos positivos e falsos negativos (Fawcett, 2006), reforça a importância de interpretar conjuntamente a matriz de confusão e métricas como *recall*, precisão e F1-score.

Para o XGBoost, a matriz de confusão foi construída a partir das predições fora da amostra produzidas nas cinco partições da validação cruzada estratificada e agrupada por matrícula, preservando, portanto, o mesmo protocolo utilizado na avaliação comparativa dos modelos.

A Figura 4 apresenta a distribuição das classificações obtidas entre as classes de não evasão (0) e evasão (1), consolidando as predições fora da amostra das cinco partições da validação cruzada. Do total de 235.632 registros avaliados, 121.894 pertenciam efetivamente à classe de não evasão e 113.738 à classe de evasão. Entre os casos reais de não evasão, o modelo classificou corretamente 98.601 (80,9%, verdadeiros negativos) e incorreu em 23.293 falsos positivos (19,1%), ou seja, estudantes que não evadiram mas foram sinalizados pelo modelo como estudantes em risco. Entre os casos reais de evasão, o XGBoost identificou corretamente 90.832 (79,9%, verdadeiros positivos), enquanto 22.906 (20,1%) foram classificados como falsos negativos, estudantes que efetivamente evadiram mas não foram capturados pelo modelo.

Figura 4 - Matriz de confusão do modelo XGBoost



Fonte: elaborado pelo autor (2026).

Esses resultados são consistentes com as métricas agregadas reportadas na Tabela 3, que indicam acurácia de 80,39%, precisão macro de 80,37%, *recall* macro de 80,38%, F1-score macro de 80,37% (F1 ponderado de 80,39%) e AUC-ROC de 0,8878, o melhor desempenho global entre os seis modelos avaliados. A proximidade entre as taxas de acerto das duas classes (80,9% para não evasão e 79,9% para evasão) indica que o modelo não favorece desproporcionalmente nenhuma delas, resultado relevante diante do desbalanceamento moderado entre as classes reais (121.894 vs. 113.738 registros). Ainda assim, a taxa de falsos negativos de 20,1% indica que aproximadamente um quinto dos estudantes que efetivamente evadiram não seria identificado pelo modelo caso este fosse aplicado como sistema de alerta precoce, o que reforça a discussão anterior sobre o custo assimétrico dos erros de classificação nesse contexto.

4.8 ANÁLISE POR MODELO

Entre os seis modelos avaliados, o XGBoost obteve o melhor desempenho geral, com AUC-ROC de 0,8878 e F1-score macro de 0,8037. Chen e Guestrin (2016) descrevem o

boosting sequencial como um processo que constrói modelos de forma adaptativa, corrigindo iterativamente os erros dos modelos anteriores, mecanismo que confere ao algoritmo alta capacidade preditiva, ainda que com maior sensibilidade ao ruído nos dados de treinamento em comparação a métodos de *bagging* como o *Random Forest*. Os resultados confirmam essa vantagem: o XGBoost superou todos os demais algoritmos avaliados, incluindo o *Random Forest*, que também pertence à família dos métodos de ensemble baseados em árvores, mas constrói seus estimadores de forma independente, por meio de amostragem *bootstrap*.

A Regressão Logística e o SVM Linear posicionaram-se, respectivamente, em segundo e terceiro lugar, com AUC-ROC de 0,8681 e 0,8680, uma diferença de apenas 0,0001 entre os dois modelos. Kelleher, Mac Namee e D'Arcy (2015) e Hastie, Tibshirani e Friedman (2009) descrevem ambos como classificadores lineares que separam as classes por meio de hiperplanos no espaço de variáveis, distinguindo-se pelo critério de otimização: maximização da verossimilhança na Regressão Logística e maximização da margem de separação no SVM. A proximidade dos resultados é consistente com essa equivalência estrutural. Notavelmente, ambos os modelos lineares superaram o *Random Forest* neste estudo, resultado que aponta para a presença de um componente linear relevante na relação entre as variáveis preditoras e a evasão neste conjunto de dados.

O *Random Forest* obteve o quarto melhor resultado, com AUC-ROC de 0,8587 e F1-score macro de 0,7809. Kelleher, Mac Namee e D'Arcy (2015) observam que métodos de *ensemble* baseados em *bagging* tendem a reduzir a variância das predições por meio da agregação de múltiplas árvores treinadas em amostras *bootstrap*, tornando o modelo mais estável frente a variações nos dados de treinamento. Essa propriedade teórica, contudo, não se traduziu em vantagem competitiva neste estudo: o *Random Forest* ficou atrás não apenas do XGBoost, mas também dos dois modelos lineares avaliados, sugerindo que a redução de variância, isoladamente, não compensou a menor capacidade do algoritmo de capturar a estrutura discriminativa presente nos dados em comparação ao boosting sequencial do XGBoost.

A Árvore de Decisão registrou AUC-ROC de 0,8258, o quinto resultado entre os seis modelos avaliados, uma distância de 0,0329 em relação ao Random Forest e de 0,0620 em relação ao XGBoost. Kelleher, Mac Namee e D'Arcy (2015) apontam que modelos individuais de árvore tendem a apresentar alta variância, o que limita sua capacidade de generalização em comparação aos métodos de *ensemble*, diferença que se manifesta empiricamente nesses resultados. Ainda assim, o desempenho da Árvore de Decisão permaneceu relativamente próximo dos demais modelos, indicando que as variáveis

preditoras disponíveis sustentam previsões razoáveis mesmo com um modelo de maior instabilidade estrutural.

O KNN registrou o menor desempenho entre todos os modelos avaliados, com AUC-ROC de 0,7997 e F1-score macro de 0,7319. Conforme descrito por Bellman (1961) e detalhado por Hastie, Tibshirani e Friedman (2009), à medida que o número de variáveis aumenta, as distâncias euclidianas tornam-se progressivamente menos discriminativas, limitação diretamente aplicável ao presente estudo. Kelleher, Mac Namee e D'Arcy (2015) acrescentam que a natureza *lazy* do algoritmo o torna computacionalmente custoso na fase de predição para grandes volumes de dados, característica que limita sua adequação a aplicações operacionais em tempo real. Ainda assim, com AUC-ROC de 0,7997, o KNN permanece relativamente próximo dos demais modelos avaliados, resultado que evidencia o poder preditivo das variáveis selecionadas mesmo para o algoritmo de menor desempenho relativo neste estudo.

4.9 ANÁLISE INTEGRADA DOS ACHADOS DA MODELAGEM PREDITIVA

A evasão acadêmica constitui um dos fenômenos mais complexos e onerosos do ensino superior brasileiro. Para a Universidade Federal da Fronteira Sul (UFFS), *campus* Chapecó, onde a taxa de sucesso na graduação é de apenas 33,11% entre 2015 e 2023 (UFFS, 2025), o desafio da retenção estudantil assume caráter estratégico. Este estudo desenvolveu e avaliou modelos preditivos de evasão com base em técnicas de mineração de dados e aprendizado de máquina, aplicados a um conjunto de dados institucional integrado composto por 235.632 registros de 8.749 matrículas de graduação, abrangendo o período de 2014.1 a 2024.2. O objetivo central foi demonstrar que dados já disponíveis nos sistemas institucionais da UFFS são suficientes para construir modelos com capacidade discriminativa robusta, capazes de subsidiar políticas institucionais de identificação precoce e retenção estudantil.

No que concerne ao atendimento dos objetivos específicos propostos, pode-se afirmar que todos foram satisfatoriamente alcançados. O primeiro objetivo, analisar o fenômeno da evasão nos cursos de graduação das IFES, foi cumprido por meio da revisão da literatura, que abrangeu os fatores associados ao abandono, as políticas públicas de permanência estudantil e as principais abordagens de predição da evasão adotadas em estudos nacionais e internacionais. O segundo objetivo, realizar a análise exploratória e o tratamento dos dados acadêmicos e demográficos, foi plenamente atendido mediante a extração, limpeza e harmonização de registros provenientes de quatro sistemas institucionais: SIGAA, SCI, SAS e

Pergamum, com elaboração de relatório e plano de qualidade de dados, tratamento de valores ausentes, codificação de variáveis categóricas e engenharia de atributos, resultando em um conjunto analítico com 235.632 registros referentes a 8.749 matrículas. O terceiro objetivo, desenvolver um modelo preditivo com base em técnicas de aprendizado supervisionado, foi cumprido com o treinamento e a comparação de seis algoritmos: Regressão Logística, Árvore de Decisão, *Random Forest*, XGBoost, SVM Linear e KNN. O quarto objetivo, validar o desempenho do modelo preditivo e verificar sua aplicabilidade como ferramenta de apoio à gestão acadêmica, foi alcançado por meio da avaliação sistemática das métricas, com o XGBoost emergindo como modelo de melhor desempenho (AUC-ROC = 0,8878), e da análise de importância das variáveis, que evidenciou a primazia do desempenho acadêmico como fator preditivo dominante e a viabilidade de integração do modelo aos fluxos de gestão institucional da UFFS. O quinto objetivo, apresentar o PTT com o modelo validado, foi concretizado por meio de um relato técnico que sistematiza a experiência de aplicação dos modelos preditivos no contexto da UFFS *campus* Chapecó e propõe diretrizes para a implementação de um sistema de alerta precoce integrado aos sistemas institucionais, com potencial de replicação em outras IFES.

4.9.1 Interpretação das variáveis preditivas do XGBoost

A análise de importância de variáveis do XGBoost, calculada pelo critério de ganho médio por divisão (*feature importance by gain*), revelou que o desempenho acadêmico domina amplamente a estrutura preditiva do modelo. A variável *media_final* foi a mais importante, com importância normalizada de aproximadamente 0,33, ou seja, sozinha responde por cerca de um terço do ganho informacional total do modelo. As variáveis *media_final_ausente* (indicador de ausência de média registrada, importância $\approx 0,115$) e *numero_faltas_ausente* (indicador de ausência de registro de faltas, importância $\approx 0,105$) ocupam o segundo e terceiro lugares, seguidas por *numero_faltas* ($\approx 0,048$) e *area_disciplina* ($\approx 0,047$). As demais 23 variáveis apresentam importâncias individuais inferiores a 0,03, com contribuições decrescentes que incluem variáveis como *ingresso*, *periodo*, *bib_total_emprestimos* e indicadores de uso do Restaurante Universitário.

A primazia da média final na estrutura preditiva do XGBoost é coerente com a teoria clássica de integração acadêmica proposta por Tinto (1975), segundo a qual o desempenho nas disciplinas e o compromisso com os objetivos educacionais são os fatores de maior peso na decisão de permanência ou abandono. Para Tinto, estudantes com baixa integração

acadêmica, evidenciada por queda de rendimento e dificuldades de adaptação às exigências curriculares, apresentam risco significativamente maior de evasão. Nierotka, Bonamino e Carrasqueira (2023), ao acompanharem longitudinalmente um grupo de estudantes de ensino superior público, confirmam que o desempenho acadêmico é o preditor mais robusto de permanência, mesmo após controle por variáveis socioeconômicas e demográficas.

A segunda e terceira posições no ranking, ocupadas pelos indicadores de ausência de média (`media_final_ausente`) e de ausência de faltas (`numero_faltas_ausente`), merecem interpretação cuidadosa. Esses indicadores foram criados durante o pré-processamento para sinalizar matrículas para as quais os sistemas institucionais não apresentaram registros de desempenho ou frequência em determinado semestre, situação que pode ocorrer por abandono de fato antes do encerramento do período letivo, por períodos de trancamento ou por inconsistências nos dados de origem. A elevada importância atribuída pelo XGBoost a esses indicadores de ausência sugere que a própria ausência de informação acadêmica constitui um sinal preditivo relevante, estudantes que não geram registros de desempenho ou frequência apresentam probabilidade significativamente maior de evasão. Esse achado tem implicação prática direta: sistemas de alerta precoce devem monitorar não apenas as métricas de desempenho negativo, mas também a ausência de geração de dados acadêmicos como um indicador precoce de desengajamento.

O número de faltas (`numero_faltas`), quarta variável em importância, aprofunda a dimensão comportamental da predição. A correlação de Pearson negativa observada entre a média final e a evasão ($r = -0,423$) e positiva entre o número de faltas e a evasão ($r = +0,326$), as duas correlações lineares mais fortes identificadas na análise exploratória, são consistentes com a dominância dessas variáveis no modelo. O Relatório de Auditoria Interna da UFFS (2025) corrobora esses achados ao identificar, entre os principais fatores associados à evasão na instituição, as dificuldades de acompanhamento das atividades acadêmicas. O fato de o modelo preditivo capturar objetivamente essas dificuldades por meio de variáveis já registradas nos sistemas institucionais representa um resultado de alta relevância operacional.

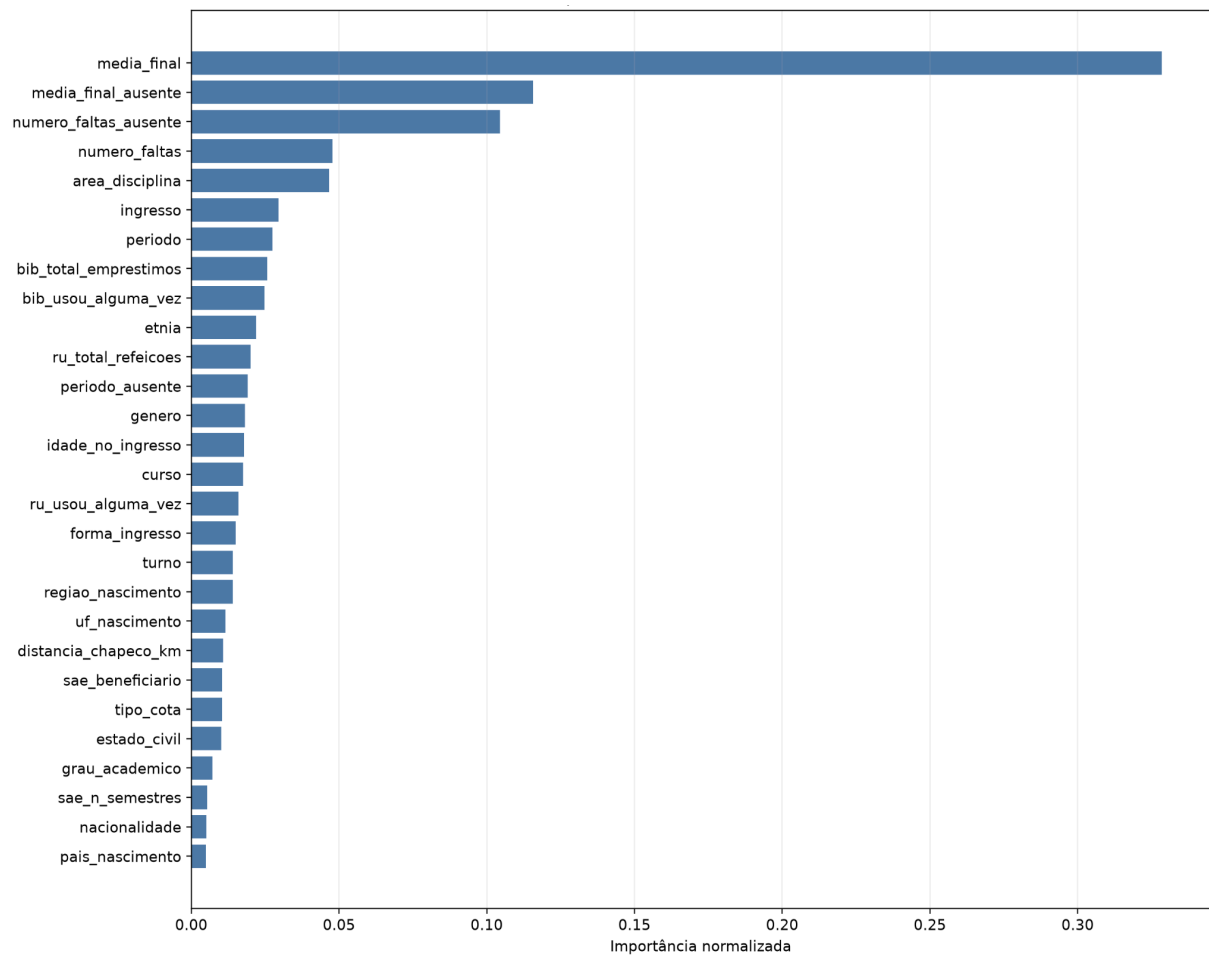
A variável `area_disciplina`, que classifica as disciplinas cursadas por área de conhecimento, ocupa a quinta posição. Sua presença entre as variáveis mais relevantes indica que o risco de evasão varia sistematicamente entre áreas, resultado que dialoga com a literatura sobre evasão diferencial nos cursos de ciências exatas e tecnológicas (Garcia; Gomes, 2022), onde as taxas de abandono tendem a ser mais elevadas do que nas ciências humanas e sociais.

Os indicadores de uso da Biblioteca (*bib_total_emprestimos* e *bib_usou_alguna_vez*) e do Restaurante Universitário (*ru_total_refeicoes* e *ru_usou_alguna_vez*) figuram em posições intermediárias no ranking de importância do XGBoost, com valores aproximados entre 0,02 e 0,025. Embora suas contribuições individuais sejam modestas em comparação às variáveis acadêmicas, sua presença entre as variáveis com maior poder preditivo, superando variáveis demográficas como gênero, etnia e distância até a UFFS, confirma a hipótese que motivou a integração dos quatro sistemas institucionais neste estudo: indicadores de engajamento comportamental e de uso de serviços institucionais possuem informação preditiva genuína sobre o risco de evasão, complementando os sinais fornecidos pelo desempenho acadêmico. Queiroga et al. (2020) demonstram resultado análogo ao mostrar que variáveis de comportamento, como frequência de acesso a plataformas digitais de ensino, têm poder preditivo relevante mesmo quando controladas por métricas de desempenho.

As variáveis socioeconômicas e demográficas, tipo de cota, gênero, etnia, distância até a UFFS, região de nascimento não figuraram entre as dez variáveis de maior importância no XGBoost, resultado que não implica irrelevância dessas dimensões para a compreensão do fenômeno, mas indica que, em termos preditivos, o desempenho acadêmico e os indicadores comportamentais capturam de forma mais direta e imediata o risco de abandono. Conforme a hipótese de mediação proposta por Paula e Picanço (2024), a origem socioeconômica pode operar de forma indireta sobre a evasão, aumentando a probabilidade de reprovações e baixo rendimento acadêmico, que por sua vez elevam o risco de abandono, o que explicaria a dominância das variáveis acadêmicas no ranking preditivo.

A Figura 5 apresenta a importância normalizada das variáveis utilizadas pelo modelo XGBoost na predição da evasão estudantil.

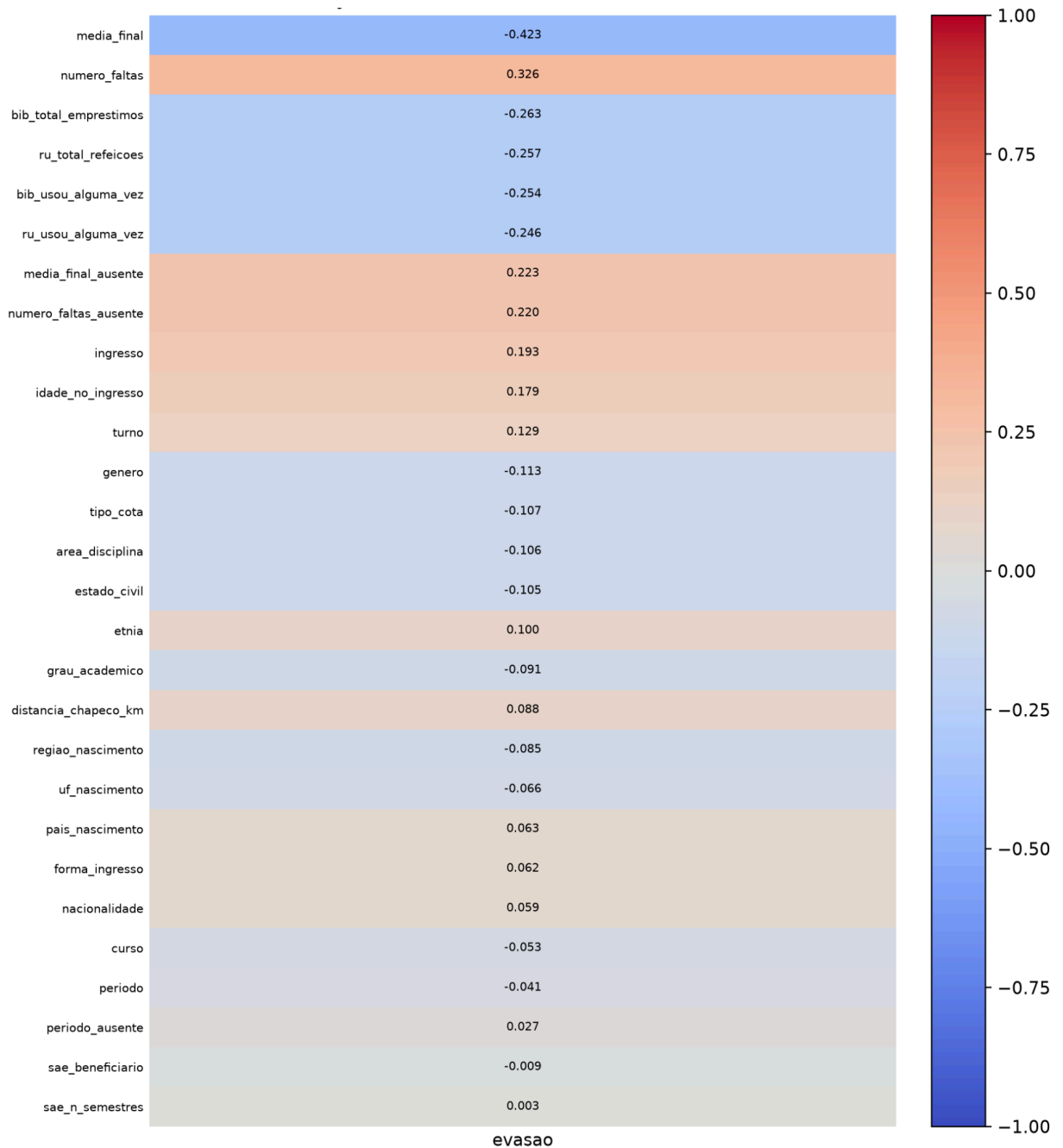
Figura 5 - Importância das variáveis pelo XGBoost



Fonte: elaborado pelo autor (2026).

A Figura 6 apresenta os coeficientes de correlação de Pearson entre as variáveis do conjunto de treinamento e a evasão estudantil. Essa medida varia entre -1 e 1 e indica a direção e a intensidade da associação linear entre duas variáveis: valores positivos representam relações diretas, enquanto valores negativos indicam relações inversas (Pearson, 1895). Na figura, a média final apresenta a maior correlação negativa com a evasão ($r = -0,423$), enquanto o número de faltas possui a maior correlação positiva ($r = 0,326$). Ressalta-se que a correlação identifica associações lineares, mas não permite estabelecer relações de causalidade.

Figura 6 - Coeficientes de correlação de Pearson entre as variáveis



Fonte: elaborado pelo autor (2026).

4.9.2 Contribuição original do estudo

Este estudo apresenta três contribuições originais para a literatura de predição de evasão em IFES brasileiras, que o diferenciam dos trabalhos precedentes no campo.

A primeira contribuição é a integração de dados provenientes de quatro sistemas institucionais, SIGAA, SCI, SAS e *Pergamum*, em um único conjunto analítico. A literatura

de EDM tem enfatizado, de forma crescente, a importância de bases de dados multidimensionais que transcendam os registros acadêmicos tradicionais (Romero; Ventura, 2020). Ao incorporar dados de uso do Restaurante Universitário e da Biblioteca como variáveis preditoras, este estudo demonstra empiricamente que indicadores de engajamento comportamental possuem poder preditivo genuíno, resultado que amplia o repertório de variáveis recomendadas para sistemas de alerta precoce em universidades federais brasileiras. A contribuição estende-se ao plano metodológico: o *pipeline* de integração desenvolvido é inteiramente reproduzível e documentado, podendo ser adaptado por outros *campi* da UFFS e por outras IFES que operem sistemas análogos.

A segunda contribuição é a adoção e documentação de um protocolo rigoroso de prevenção ao vazamento de informação (*data leakage*), que incluiu três medidas complementares: a remoção de variáveis que codificam direta ou indiretamente o desfecho da evasão (*status*, *saida*, *situacao_sigla*); a aplicação de validação cruzada estratificada com agrupamento por matrícula (*StratifiedGroupKFold*), impedindo que registros de uma mesma matrícula fossem simultaneamente utilizados nas partições de treino e teste; e a restrição temporal das variáveis de serviços institucionais aos primeiros semestres de cada matrícula, eliminando informações retrospectivas indisponíveis no momento real da predição.

A terceira contribuição é a amplitude temporal do dataset, que cobre onze anos de registros acadêmicos (2014.1 a 2024.2) e inclui 8.749 matrículas de graduação em todos os cursos presenciais do *campus* Chapecó. A maioria dos estudos brasileiros de predição de evasão opera com datasets de menor escala e período mais restrito (Magalhães et al., 2025). A série histórica ampla permite que o modelo capture padrões de evasão em diferentes grupos de ingresso, incluindo os efeitos da pandemia de COVID-19 sobre as trajetórias acadêmicas dos estudantes matriculados entre 2019 e 2021, período de suspensão das atividades presenciais que se refletiu, de forma observável, no tratamento das variáveis de uso do Restaurante Universitário nos semestres de 2020.2 a 2021.2.

4.9.3 Implicações para a gestão acadêmica da UFFS

Os resultados deste estudo têm implicações diretas para a gestão acadêmica da UFFS, especialmente diante do diagnóstico apresentado pelo Relatório de Auditoria Interna (UFFS, 2025) e pelo Acórdão 1868/2024 do TCU (2024): a ausência de sistema formal e contínuo de monitoramento da evasão e a taxa média de sucesso na graduação de 33,11% entre 2015 e 2023. O TCU identificou que 51% das universidades federais brasileiras declararam não

possuir diagnóstico do perfil dos estudantes evadidos ou das causas subjacentes ao abandono, contexto que torna a disponibilidade de um modelo preditivo validado um recurso de elevada relevância institucional.

O modelo XGBoost desenvolvido neste estudo oferece uma ferramenta objetiva e replicável para subsidiar a identificação precoce de estudantes em risco de evasão, com base em dados já disponíveis nos sistemas institucionais. Com AUC-ROC de 0,8878 e recall macro de 0,8038, o modelo apresentou desempenho satisfatório e equilíbrio geral na classificação das classes analisadas. Esses resultados indicam potencial de uso como instrumento de apoio à gestão acadêmica, especialmente em estratégias de alerta precoce, nas quais a antecipação do risco pode contribuir para a definição de intervenções pedagógicas e assistenciais mais adequadas (Baker; Yacef, 2009).

A prioridade das variáveis acadêmicas no modelo tem implicação prática imediata: os primeiros sinais de risco de evasão são identificáveis já nos semestres iniciais da trajetória acadêmica, por meio do acompanhamento do desempenho e da frequência nas disciplinas. Isso sugere que intervenções precoces, como programas de tutoria e nivelamento nas disciplinas de ingresso, especialmente nas áreas de ciências exatas, têm o maior potencial de impacto na redução da evasão, conforme identificado por Gilioli (2016) em análise das IFES brasileiras. A elevada importância dos indicadores de ausência de dados (*media_final_ausente* e *numero_faltas_ausente*) acrescenta uma dimensão operacional adicional: a qualidade e completude dos registros acadêmicos nos sistemas institucionais é, ela própria, uma fonte de informação preditiva que deve ser preservada e monitorada.

A presença de indicadores de engajamento institucional, como o uso do Restaurante Universitário, da Biblioteca e do Sistema de Análise Socioeconômica entre as variáveis com poder preditivo relevante, sugere que esses elementos podem funcionar como marcadores indiretos de vínculo do estudante com a instituição. Em vez de indicar, de forma causal, que tais políticas reduzem a evasão, os resultados apontam que estudantes com maior interação com serviços institucionais tendem a apresentar padrões distintos de risco em relação àqueles sem registro de uso. Essa interpretação é coerente com a dimensão institucional da evasão discutida por Baggi e Lopes (2011), para quem o suporte institucional e o acompanhamento discente são fatores relevantes na compreensão do abandono. Do ponto de vista operacional, esse achado reforça a importância de monitorar não apenas indicadores de desempenho acadêmico, mas também indicadores de presença, acesso a serviços e engajamento comportamental nos sistemas de alerta precoce.

Slade e Prinsloo (2013) alertam, entretanto, que a implantação de sistemas preditivos em contextos educacionais deve ser acompanhada de reflexão ética sobre privacidade, transparência e uso responsável dos dados. Os autores defendem que os modelos preditivos não devem ser utilizados para estigmatizar estudantes ou para reduzir a complexidade do fenômeno da evasão a uma classificação binária de risco, mas como instrumentos de apoio à decisão institucional que ampliam a capacidade de identificar estudantes que necessitam de suporte, respeitando sua autonomia e privacidade. Essa perspectiva é plenamente aplicável ao contexto da UFFS, onde os dados devem ser tratados em conformidade com a Lei Geral de Proteção de Dados Pessoais (LGPD) e utilizados exclusivamente para fins de apoio pedagógico e assistencial, com governança transparente e mecanismos de contestação pelos próprios estudantes.

5 SEÇÃO PROPOSITIVA (PTT)

Esta seção consolida, em formato de relato técnico conclusivo, a recomendação de uso institucional dos modelos de aprendizado de máquina desenvolvidos e validados nesta dissertação, dirigida à Pró-Reitoria de Graduação da Universidade Federal da Fronteira Sul. A proposta decorre diretamente dos resultados da modelagem preditiva apresentada no Capítulo 4 e tem por finalidade demonstrar que os dados já disponíveis nos sistemas institucionais contêm padrões consistentes para a identificação precoce de estudantes em risco de evasão, bem como que o *pipeline* documentado neste trabalho pode ser operacionalizado de forma periódica pela própria equipe técnica da instituição, sem necessidade de desenvolvimento adicional significativo. O detalhamento operacional integral dos procedimentos de implantação consta do Relatório Técnico Conclusivo apresentado como apêndice, na forma do Produto Técnico Tecnológico (PTT) desta dissertação.

O fundamento técnico da proposta está no desempenho obtido pelos seis algoritmos avaliados. O modelo XGBoost apresentou o melhor resultado, com AUC-ROC de 0,8878, seguido pela Regressão Logística (0,8681), pelo SVM Linear (0,8680), pelo Random Forest (0,8587), pela Árvore de Decisão (0,8258) e pelo KNN (0,7997). Todos os modelos situaram-se na faixa de alto desempenho discriminativo, indicando que os dados integrados dos quatro sistemas da UFFS contêm informações relevantes para distinguir registros associados a estudantes com maior ou menor risco de evasão. Esses resultados retrospectivos demonstram viabilidade técnica preliminar para o desenvolvimento de um sistema institucional de alerta precoce, a partir de informações institucionais até então não exploradas de forma sistemática para essa finalidade.

A análise de importância de variáveis, conduzida sobre o modelo XGBoost pelo critério de ganho médio por divisão (feature importance by gain), indicou predomínio das variáveis de desempenho acadêmico na estrutura preditiva. A média final (media_final) foi a variável mais importante, com importância normalizada de aproximadamente 0,33, respondendo isoladamente por cerca de um terço do ganho informacional total do modelo. Na sequência posicionaram-se os indicadores de ausência de registro de média (media_final_ausente, $\approx 0,115$) e de faltas (numero_faltas_ausente, $\approx 0,105$), o número de faltas (numero_faltas, $\approx 0,048$) e a área de conhecimento das disciplinas cursadas (area_disciplina, $\approx 0,047$); as demais variáveis apresentaram importâncias individuais inferiores a 0,03. Do ponto de vista operacional, o resultado evidencia que os sinais de risco de evasão se manifestam de forma mensurável já nos primeiros semestres, por meio de

indicadores rotineiramente acompanhados pelas coordenações de curso e equipes pedagógicas, como a queda da média final, o aumento do número de faltas e a ausência de registros de desempenho acadêmico. A contribuição do modelo consiste em combinar sistematicamente esses indicadores e atribuir a cada estudante uma estimativa quantitativa do risco de abandono, permitindo priorizar com maior precisão os casos que demandam atenção institucional.

Destaca-se, ainda, o papel do uso do Restaurante Universitário como preditor de evasão. A variável correspondente ao total de refeições utilizadas pelo estudante no RU durante os três primeiros semestres da matrícula (*ru_total_refeicoes*) apresentou poder preditivo relevante no modelo XGBoost, com importância intermediária, superando variáveis sociodemográficas como gênero, etnia e distância até a UFFS. O resultado sugere que a frequência ao RU funciona como indicador indireto de presença física no *campus* e de engajamento institucional nos semestres iniciais da graduação: estudantes que utilizam regularmente o serviço tendem a permanecer mais tempo na universidade, desenvolver vínculos com a comunidade acadêmica e apresentar menor risco de abandono. Em termos de gestão, esse resultado reforça a importância de monitorar o acesso aos serviços institucionais e assistenciais nos primeiros semestres, especialmente entre ingressantes oriundos de contextos socioeconômicos mais vulneráveis, como parte de estratégias de acompanhamento e alerta precoce.

Quanto às variáveis socioeconômicas e demográficas, como tipo de cota, gênero, etnia, região de nascimento e distância até a UFFS, os resultados indicam importância preditiva real, porém consideravelmente inferior à das variáveis de desempenho acadêmico. Tal constatação não implica irrelevância dessas dimensões para a compreensão do fenômeno, mas evidencia que, em termos preditivos, o desempenho acadêmico captura de forma mais direta e imediata o risco de abandono do que as características de origem do estudante. Uma interpretação plausível é a de mediação: a vulnerabilidade socioeconômica não conduz diretamente à evasão, mas eleva a probabilidade de dificuldades acadêmicas que, por sua vez, aumentam o risco de abandono, de modo que as variáveis acadêmicas acabam capturando o efeito final dessa cadeia. Para a gestão da UFFS, o resultado reforça a importância do acompanhamento contínuo do desempenho acadêmico como principal mecanismo de identificação precoce de risco, sem prejuízo das políticas de assistência socioeconômica voltadas à mitigação das condições de vulnerabilidade dos estudantes.

Em termos de entrega concreta à Pró-Reitoria de Graduação, esta dissertação disponibiliza um conjunto articulado de recursos técnicos que podem subsidiar o

desenvolvimento e a experimentação institucional de um sistema de alerta precoce. O primeiro é o *dataset* consolidado, resultante da integração dos dados do SIGAA, do SCI, do SAS e do *Pergamum*, com 235.632 registros acadêmicos referentes a 8.749 matrículas do *campus* Chapecó no período de 2014.1 a 2024.2, tratado e documentado conforme os procedimentos descritos no Capítulo 4. O segundo é o *pipeline* de pré-processamento completo, implementado em Python com as bibliotecas *pandas*, *numpy* e *scikit-learn*, executável pela equipe técnica da Secretaria Especial de Tecnologia da Informação da UFFS sobre novos dados extraídos dos sistemas a cada semestre letivo, reproduzindo de forma automatizada todas as etapas de tratamento documentadas. O terceiro é o modelo XGBoost treinado, aplicável sobre o conjunto de dados preparado para gerar, a cada semestre, uma estimativa da probabilidade de evasão de cada matrícula ativa e classificar os estudantes em faixas de risco que orientem a priorização das intervenções pelas coordenações de curso e equipes de assistência estudantil.

Para que esses recursos se convertam em apoio efetivo à gestão acadêmica, recomenda-se que a Pró-Reitoria de Graduação avalie a viabilidade de incorporar a execução periódica do *pipeline* e do modelo ao calendário institucional, gerando os resultados da predição ao final de cada semestre letivo e disponibilizando-os às coordenações de curso e equipes pedagógicas no início do semestre seguinte. O ciclo de atualização semestral asseguraria que as informações utilizadas na identificação dos estudantes em risco reflitam sempre a trajetória acadêmica mais recente de cada matrícula, aumentando a precisão e a relevância dos alertas gerados. Os limiares de corte que definem a classificação em alto, médio ou baixo risco devem ser calibrados pela própria instituição, em função de sua capacidade operacional de acompanhamento, podendo ser ajustados ao longo dos primeiros ciclos de uso com base nos resultados observados.

Ressalva-se que a aplicação dos modelos preditivos deve ser compreendida como recurso de apoio à decisão dos gestores acadêmicos, e não como mecanismo autônomo de classificação de estudantes. O score de probabilidade de evasão é uma estimativa estatística, não uma certeza, devendo ser interpretado pelas coordenações de curso e equipes pedagógicas à luz do conhecimento que detêm sobre a realidade específica de cada estudante e de cada curso. A utilidade do modelo reside na capacidade de processar sistematicamente um volume de dados inviável para a análise humana direta e de sinalizar, com base em padrões históricos, quais estudantes apresentam características semelhantes às de matrículas que resultaram em evasão. A decisão sobre como e quando intervir permanece sob a responsabilidade dos profissionais da instituição, que devem conduzir as abordagens com sensibilidade, respeito à

autonomia dos estudantes e observância à Lei Geral de Proteção de Dados Pessoais. Na eventual implantação institucional, a geração e a análise dos escores deverão ocorrer em ambiente controlado, com acesso restrito aos profissionais formalmente autorizados. Os identificadores deverão ser pseudonimizados durante o processamento técnico, sendo a reidentificação admitida somente na etapa de acompanhamento institucional, exclusivamente pelos agentes competentes e para finalidade pedagógica ou assistencial previamente estabelecida.

Por fim, embora o modelo tenha sido desenvolvido e validado para os dados do *campus* Chapecó, o *pipeline* documentado nesta dissertação possui potencial de adaptação aos demais *campi* da instituição e a outras instituições federais de ensino superior com estrutura de dados similar. A principal condição de replicabilidade é a disponibilidade de dados integrados de sistemas acadêmicos, de assistência estudantil e de serviços institucionais em formato compatível com o pipeline desenvolvido. A replicação em outros contextos deve ser precedida de novo processo de treinamento e validação do modelo, uma vez que os padrões de evasão podem variar significativamente entre *campi*, cursos e perfis discentes distintos. Recomenda-se, adicionalmente, a reavaliação periódica do desempenho preditivo do modelo, sobretudo após mudanças curriculares relevantes ou alterações expressivas no perfil dos ingressantes, de modo a assegurar que os padrões históricos sobre os quais foi treinado permaneçam representativos da realidade institucional.

6 CONSIDERAÇÕES FINAIS

Este capítulo discute os resultados obtidos na modelagem preditiva da evasão acadêmica nos cursos de graduação da Universidade Federal da Fronteira Sul (UFFS), *campus* Chapecó, articulando-os com a fundamentação teórica e com a literatura nacional e internacional de Mineração de Dados Educacionais (EDM). A discussão organiza-se em seis eixos: o desempenho comparativo dos seis modelos avaliados; a interpretação das variáveis preditivas mais relevantes identificadas pelo modelo XGBoost e sua relação com as teorias de evasão; a contribuição original do estudo para o campo; as implicações institucionais dos resultados para a gestão acadêmica da UFFS; as limitações do estudo; e as perspectivas para trabalhos futuros. A análise parte de um pressuposto consolidado na literatura de EDM: modelos preditivos de evasão não devem ser avaliados apenas por sua performance técnica, mas também por sua utilidade institucional e pela coerência entre os padrões identificados pelos algoritmos e o conhecimento acumulado sobre o fenômeno (Quimiz-Moreira et al., 2025). Nessa perspectiva, os resultados são discutidos por sua capacidade de subsidiar intervenções pedagógicas e políticas institucionais de permanência estudantil (Slade; Prinsloo, 2013).

Os resultados deste estudo demonstram que a integração de dados de quatro sistemas institucionais da UFFS, tratados com rigor metodológico para prevenção ao vazamento de informação, permite o desenvolvimento de modelos preditivos de evasão com desempenho consistente com resultados reportados na literatura internacional de EDM. O XGBoost emergiu como o modelo mais discriminativo ($AUC-ROC = 0,8878$), combinando alto poder preditivo, capacidade de capturar interações complexas entre variáveis heterogêneas e fornecimento de importância relativa das variáveis, elemento essencial para a tradução dos resultados em políticas institucionais concretas.

As variáveis de desempenho acadêmico, com destaque para a média final, os indicadores de ausência de dados e o número de faltas, mostraram-se os preditores mais relevantes, confirmando a teoria da integração estudantil de Tinto (1975) e os achados da literatura contemporânea de EDM. Os indicadores de uso do Restaurante Universitário, da Biblioteca e dos beneficiários de auxílios socioeconômicos, emergiram como variáveis complementares de engajamento institucional precoce, resultado original deste estudo que abre perspectivas para o monitoramento de comportamentos não-acadêmicos em sistemas de alerta precoce.

Em termos institucionais, os resultados apontam para a viabilidade técnica e a relevância estratégica de um sistema de alerta precoce à evasão na UFFS, fundamentado em dados já disponíveis nos sistemas institucionais e operacionalizado por meio do *pipeline* documentado neste trabalho. Tal sistema representaria uma resposta concreta às deficiências identificadas pelo TCU (2024) e pela auditoria interna da UFFS: a ausência de mecanismos sistemáticos de monitoramento e identificação precoce de estudantes em risco de abandono. Ao mesmo tempo, reafirma-se que a implementação de qualquer ferramenta preditiva em contexto educacional deve ser orientada por princípios éticos de transparência, privacidade e uso responsável dos dados, de modo a que a tecnologia seja colocada a serviço da permanência estudantil e da equidade no ensino superior público, e não como instrumento de classificação ou estigmatização dos estudantes mais vulneráveis.

Este estudo apresenta limitações que devem ser reconhecidas para a correta interpretação de seus resultados e para orientar pesquisas futuras.

A primeira limitação diz respeito ao escopo geográfico e institucional. Os dados utilizados referem-se exclusivamente ao *campus* Chapecó da UFFS, o que circunscreve a generalização dos resultados. Embora o perfil do estudante da UFFS, predominantemente oriundo da região Sul e de contextos socioeconômicos vulneráveis, seja representativo de uma parcela significativa das universidades federais interiorizadas (Nierotka; Bonamino; Carrasqueira, 2023), a replicação do estudo nos demais *campi* da UFFS e em outras IFES é necessária para verificar a estabilidade dos padrões identificados.

A segunda limitação refere-se à definição operacional da variável alvo. A evasão foi operacionalizada como o status CANCELADO no SIGAA, excluindo situações como transferência interna entre cursos e trancamento prolongado, que podem representar formas distintas de desengajamento acadêmico. Conforme alerta o TCU (2024), a definição restrita de evasão adotada pelo INEP subestima o real alcance do fenômeno nas universidades federais, o que pode resultar em subnotificação de casos efetivos de abandono no conjunto de dados analisado.

A terceira limitação é a ausência de variáveis subjetivas no modelo, como motivação para o curso, qualidade do vínculo com os docentes e percepção de pertencimento à instituição, dimensões que a literatura reconhece como relevantes para a compreensão da evasão (Cunha; Morosini, 2014), mas que não se encontram registradas nos sistemas institucionais analisados e cuja mensuração requereria abordagens metodológicas complementares, como *surveys* ou entrevistas longitudinais.

A quarta limitação refere-se ao horizonte temporal da predição. O modelo foi treinado e avaliado com dados de semestres já concluídos, e sua aplicação prospectiva, sobre matrículas ativas em semestres em curso, requer atualização periódica do *pipeline* de integração e re-treinamento do modelo para incorporar os dados mais recentes. A estabilidade das métricas de desempenho ao longo do tempo, em face de mudanças nos perfis de ingresso e nas políticas institucionais, constitui uma dimensão de avaliação que este estudo não pôde abordar.

Com base nas limitações identificadas e nos resultados obtidos, sugere-se as seguintes direções para pesquisas futuras.

A primeira perspectiva é a extensão do estudo aos demais *campi* da UFFS: Erechim, Cerro Largo, Laranjeiras do Sul, Realeza e Passo Fundo, permitindo uma análise comparativa dos padrões de evasão entre *campi* com perfis distintos de cursos e populações estudantis. Tal extensão verificaria se os preditores identificados no *campus* Chapecó se mantêm estáveis em outros contextos institucionais, contribuindo para a construção de um modelo unificado de alerta precoce para toda a UFFS.

A segunda perspectiva é o desenvolvimento e implantação efetiva do sistema de alerta precoce delineado na Seção Propositiva desta dissertação, integrando o *pipeline* de modelagem aos sistemas da UFFS para gerar relatórios periódicos de risco por matrícula destinados a coordenadores de curso e equipes de assistência estudantil. Quimiz-Moreira et al. (2025) identificam essa lacuna como uma das mais relevantes na literatura: poucos estudos avançam da modelagem preditiva para a implantação de sistemas operacionais de uso institucional.

A terceira perspectiva é a incorporação de técnicas de explicabilidade algorítmica, como SHAP (*SHapley Additive exPlanations*) e LIME (*Local Interpretable Model-agnostic Explanations*), para produzir explicações individualizadas das predições, indo além da importância global de variáveis para oferecer, a gestores e equipes pedagógicas, as razões específicas pelas quais determinada matrícula foi classificada em categoria de risco. Quimiz-Moreira et al. (2025) apontam a explicabilidade como a fronteira mais promissora e menos explorada na literatura de predição de evasão.

A quarta perspectiva é a investigação da dimensão temporal da evasão por meio de modelos de análise de sobrevivência ou de redes neurais recorrentes, capazes de modelar a trajetória acadêmica ao longo dos semestres como uma sequência temporal, abordagem que pode ampliar a capacidade preditiva e a precisão da janela de intervenção em relação aos modelos de classificação binária adotados neste estudo.

A quinta perspectiva é a ampliação do escopo de variáveis por meio da integração de dados externos, como indicadores socioeconômicos do município de origem (IDH, cobertura de internet, mercado de trabalho local) e dados do SISU sobre a ordem de preferência dos candidatos por curso e instituição, variáveis que podem capturar dimensões da motivação inicial e da compatibilidade entre expectativas e realidade institucional que os sistemas internos da UFFS não registram.

A sexta perspectiva refere-se à condução de uma validação prospectiva do modelo preditivo, reservando coortes de ingresso mais recentes como conjunto de teste final, de modo a verificar a estabilidade do desempenho quando o modelo é efetivamente aplicado a estudantes de períodos posteriores ao intervalo utilizado no treinamento. Tal validação demandaria, adicionalmente, a explicitação formal da unidade de observação, da unidade de predição e do horizonte temporal de ocorrência do desfecho, bem como o refinamento da definição operacional da variável-alvo, distinguindo cancelamento, evasão, conclusão, retenção e matrículas ainda ativas (censuradas), de modo a assegurar que o modelo capture padrões antecipatórios do abandono, e não apenas a replicação de uma trajetória acadêmica já consolidada.

A sétima perspectiva envolve a formalização de um protocolo de governança algorítmica para a operacionalização responsável do sistema de alerta precoce proposto, contemplando a calibração das probabilidades preditas, a justificativa metodológica do limiar de classificação em função da capacidade institucional de atendimento e a análise de desempenho desagregada por curso, coorte de ingresso e grupos potencialmente vulneráveis, de modo a identificar eventuais diferenças de desempenho do modelo. Esse protocolo poderá estabelecer regras de acesso aos escores de risco, mecanismos de revisão humana das classificações, salvaguardas contra usos automatizados de caráter punitivo e procedimentos de atualização periódica do modelo, consolidando o PTT como proposta de implementação piloto passível de validação, monitoramento e aprimoramento contínuos antes de eventual adoção em escala institucional.

REFERÊNCIAS

- ALBUQUERQUE, J. G. M.; ABREU, M. T. C. DE; LIMA, I. N. DE. O impacto da Inteligência Artificial na personalização do ensino. **Rebena: Revista Brasileira De Ensino E Aprendizagem**, v. 9, p. 182–192, 2024. Disponível em: <https://rebena.emnuvens.com.br/revista/article/view/242>. Acesso em: 5 abr. 2026.
- ALMAGHRABI, H.; SOH, B.; LI, A.; ALSOLBI, I. SoK: The Impact of Educational Data Mining on Organisational Administration. **Information**, v. 15, n. 11, p. 738, 2024. MDPI AG. Disponível em: <https://doi.org/10.3390/info15110738>. Acesso em: 4 abr. 2026.
- ALVES, A. F.; INÁCIO, C. B. V.; POZZEBON, E.; SILVA, J. B. DA. Aspectos relevantes dos modelos preditivos de inteligência artificial no combate à evasão escolar em cursos de graduação: uma revisão sistemática. **Anais do XXXV Simpósio Brasileiro de Informática na Educação (SBIE 2024)**, p. 1503–1514, 2024. Sociedade Brasileira de Computação - SBC. Disponível em: <https://doi.org/10.5753/sbie.2024.241899>. Acesso em: 7 fev. 2026.
- ANDRIOLA, W. FATORES ASSOCIADOS À EVASÃO DISCENTE NA UNIVERSIDADE FEDERAL DO CEARÁ (UFC) DE ACORDO COM AS OPINIÕES DE DOCENTES E DE COORDENADORES DE CURSOS REICE. **Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación**, v. 7, n. 4, p. 342–356, 2009. Madrid, España: Red Iberoamericana de Investigación Sobre Cambio y Eficacia Escolar. Disponível em: <https://www.redalyc.org/pdf/551/55114094018.pdf>. Acesso em: 12 mar. 2026.
- BAGGI, C. A. DOS S.; LOPES, D. A. Evasão e avaliação institucional no ensino superior: uma discussão bibliográfica. **Avaliação: Revista da Avaliação da Educação Superior (Campinas)**, v. 16, n. 2, p. 355–374, 2011. Disponível em: <https://doi.org/10.1590/S1414-40772011000200007>. Acesso em: 15 jan. 2026.
- BAKER, R. S. J. D.; YACEF, K. The State of Educational Data Mining in 2009: A Review and Future Visions. **JEDM | Journal of Educational Data Mining**, v. 1, n. 1, p. 3–17, 2009. Disponível em: <https://jedm.educationaldatamining.org/index.php/JEDM/article/view/8>. Acesso em: 15 mar. 2026.
- BANDEIRA, Y. A.; AQUINO, F. J. A. DE. FUTURO DA EDUCAÇÃO COM INTELIGÊNCIA ARTIFICIAL. **Revista Nova Paideia: Revista Interdisciplinar em Educação e Pesquisa**, v. 7, n. 1, p. 331–615, 2025. Revista Nova Paideia. Disponível em: <https://ojs.novapaideia.org/index.php/RIEP/article/view/437>. Acesso em: 21 mar. 2026.
- BARLEM, J. G. T.; LUNARDI, V. L.; BORDIGNON, S. S.; et al. Opção e evasão de um curso de graduação em enfermagem: percepção de estudantes evadidos. **Revista Gaúcha de Enfermagem**, v. 33, n. 2, p. 132–138, 2012. Disponível em: <https://doi.org/10.1590/S1983-14472012000200019>. Acesso em: 17 abr. 2026.
- BELLMAN, Richard E. **Adaptive control processes: a guided tour**. Princeton: Princeton University Press, 1961. Disponível em: <https://doi.org/10.1515/9781400874668>. Acesso em: 2 mar. 2026.
- BITENCOURT, W. A.; SILVA, D. M.; XAVIER, G. C. Pode a inteligência artificial apoiar ações contra evasão escolar universitária? **Ensaio: Avaliação e Políticas Públicas em**

Educação, 2021. Disponível em: <https://doi.org/10.1590/S0104-403620220003002854>. Acesso em: 2 mar. 2026.

BRASIL. **Lei nº 13.005, de 25 de junho de 2014.** Aprova o Plano Nacional de Educação - PNE e dá outras providências. Brasília, DF: Presidência da República, 2014. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2011-2014/2014/lei/l13005.htm. Acesso em: 10 out. 2025.

BRASIL. **Lei nº 14.129, de 29 de março de 2021.** Dispõe sobre princípios, regras e instrumentos para o Governo Digital. Brasília, DF: Presidência da República, 2021. Disponível em: https://www.planalto.gov.br/ccivil_03/_ato2019-2022/2021/lei/l14129.htm. Acesso em: 10 out. 2025.

BRASIL. **Lei nº 14.934, de 25 de julho de 2024.** Prorroga, até 31 de dezembro de 2025, a vigência do Plano Nacional de Educação, aprovado por meio da Lei nº 13.005, de 25 de junho de 2014. Brasília, DF: Presidência da República, 2024a. Disponível em: <https://www2.camara.leg.br/legin/fed/lei/2024/lei-14934-25-julho-2024-795997-publicacaooriginal-172483-pl.html>. Acesso em: 9 out. 2025.

BRASIL. **Decreto nº 12.198, de 24 de setembro de 2024.** Institui a Estratégia Federal de Governo Digital 2024–2027 e a Infraestrutura Nacional de Dados. Brasília, DF: Presidência da República, 2024b. Disponível em: <https://www2.camara.leg.br/legin/fed/decret/2024/decreto-12198-24-setembro-2024-796286-publicacaooriginal-173095-pe.html>. Acesso em: 1 out. 2025.

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

CHAPMAN, P.; CLINTON, J.; KERBER, R.; et al. **CRISP-DM 1.0 Step-by-step data mining guide.** The CRISP-DM Consortium, 2000.

CHEN, T.; GUESTRIN, C. XGBoost: a Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16**, v. 1, n. 1, p. 785–794, 2016. Disponível em: <https://doi.org/10.1145/2939672.2939785>. Acesso em: 16 dez. 2025.

COLPANI, R. Educação a Distância: Identificação dos Fatores que Contribuíram para a Evasão dos Alunos no Curso de Gestão Empresarial da Faculdade de Tecnologia de Mococa. **EaD em Foco**, v. 8, n. 1, 2018. Fundação CECIERJ. Disponível em: <https://doi.org/10.18264/eadf.v8i1.688>. Acesso em: 3 mar. 2026.

COVER, T.; HART, P. Nearest neighbor pattern classification. **IEEE Transactions on Information Theory**, v. 13, n. 1, p. 21–27, 1967.

CUNHA, E. R.; MOROSINI, M. C. Evasão na educação superior: uma temática em discussão. **Revista Cocar**, v. 7, n. 14, p. 82–89, 2014. Disponível em: <https://periodicos.uepa.br/index.php/cocar/article/view/283>. Acesso em: 1 fev. 2026.

DATA SCIENCE PM. CRISP-DM is Still the Most Popular Framework for Executing Data Science Projects, 2024. Disponível em: <https://www.datascience-pm.com/crisp-dm-still-most-popular/>. Acesso em: 24 nov. 2025.

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

FIX, E.; HODGES, J. L. **Discriminatory analysis**: nonparametric discrimination, consistency properties. Randolph Field, TX: USAF School of Aviation Medicine, 1951.

FRITSCH, R.; ROCHA, C. S. DA; VITELLI, R. F. A evasão nos cursos de graduação em uma instituição de ensino superior privada. **Revista Educação em Questão**, v. 52, n. 38, p. 81–108, 2015. Universidade Federal do Rio Grande do Norte - UFRN. Disponível em: <https://doi.org/10.21680/1981-1802.2015v52n38ID7963>. Acesso em: 19 mar. 2026.

GARCIA, L. M. L. DA S.; GOMES, R. S. Causas da evasão em cursos de ciências exatas: uma revisão da produção acadêmica. **Revista Educar Mais**, v. 6, p. 937–957, 2022. Disponível em: <https://doi.org/10.15536/reducarmais.6.2022.2970>. Acesso em: 17 nov. 2025.

GARCÍA, S.; LUENGO, J.; HERRERA, F. **Data Preprocessing in Data Mining**. Cham: Springer, 2015.

GIL, A. C. **Métodos e técnicas de pesquisa social**. 6 ed. São Paulo: Atlas, 2008.

GIL, A. C. **Como elaborar Projetos de Pesquisa**. 7 ed. São Paulo: Atlas, 2022.

GILIOLI, R. DE S. P. **Evasão em instituições federais de ensino superior no Brasil**: expansão da rede, Sisu e desafios. Brasília, DF: Câmara dos Deputados, Consultoria Legislativa, 2016.

GOLDSCHMIDT, R.; PASSOS, E. **Data Mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

GOMES, M.; HIRATA, G. Determinantes da evasão no ensino superior: uma abordagem de riscos competitivos. **Pesquisa e Planejamento Econômico**, v. 52, n. 03, p. 9–37, 2022. Disponível em: <http://doi.org/10.38116/ppe52n3art1>. Acesso em: 14 fev. 2026.

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

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. 2 ed. New York: Springer, 2009.

HEATON, J. Ian Goodfellow, Yoshua Bengio, and Aaron Courville: Deep learning. **Genetic Programming and Evolvable Machines**, v. 19, n. 1–2, p. 305–307, 2017. Disponível em: <https://link.springer.com/article/10.1007/s10710-017-9314-z>. Acesso em: 10 jul. 2026.

HEGGLER, J. M.; SZMOSKI, R. M.; MIQUELIN, A. F. AS DUALIDADES ENTRE O USO DA INTELIGÊNCIA ARTIFICIAL NA EDUCAÇÃO E OS RISCOS DE VIESES ALGORÍTMICOS. **Educação & Sociedade**, v. 46, 2025. Centro de Estudos Educação e Sociedade. Disponível em: <https://doi.org/10.1590/ES.289323>. Acesso em: 8 mar. 2026.

INEP. **Censo da Educação Superior 2022**: notas estatísticas, 2023. Brasília: INEP.

Disponível em:

<https://www.gov.br/inep/pt-br/centrais-de-conteudo/acervo-linha-editorial/publicacoes-institucionais/estatisticas-e-indicadores-educacionais/censo-da-educacao-superior-2022-notas-estatisticas>. Acesso em: 15 fev. 2026.

INSTITUTO SEMESP. **Mapa do Ensino Superior no Brasil 2024**, v. 14, 2024. São Paulo:

Instituto Semesp. Disponível em:

<https://www.semesp.org.br/wp-content/uploads/2024/04/mapa-ensino-superior-brasil-2024.pdf>. Acesso em: 24 set. 2025.

INSTITUTO SEMESP. **Mapa do Ensino Superior no Brasil 2025**, v. 15, 2025. São Paulo:

Instituto Semesp. Disponível em:

<https://www.semesp.org.br/wp-content/uploads/2025/02/mapa-ensino-superior-brasil-2025.pdf>. Acesso em: 25 set. 2025.

KALITA, E.; OYELERE, S. S.; GAFTANDZHIEVA, S.; et al. Educational data mining: a 10-year review. **Discover Computing**, v. 28, n. 1, 2025. Springer Science and Business Media LLC. Disponível em: <https://doi.org/10.1007/s10791-025-09589-z>. Acesso em: 7 mar. 2026.

KELLEHER, J. D.; MAC NAMEE, B.; D'ARCY, A. **Fundamentals of Machine Learning for Predictive Data Analytics**: Algorithms, Worked Examples, and Case Studies. Cambridge: MIT Press, 2015.

KNOP, M.; COLLARES, A. C. M. A influência da origem social na probabilidade de concluir os diferentes cursos de ensino superior. **Sociedade e Estado**, v. 34, n. 2, p. 351–380, 2019. Disponível em: <https://doi.org/10.1590/s0102-6992-201934020001>. Acesso em: 16 mar. 2026.

KOTSIANTIS, S. B. Supervised Machine Learning: A Review of Classification Techniques. **Informatica**, v. 31, p. 249–268, 2007. Disponível em: <http://www.informatica.si/index.php/informatica/article/viewFile/148/140>. Acesso em: 24 fev. 2026.

KRAJEVSKI, L. C. Universidade e desenvolvimento regional: a experiência da UFFS. **Revista Brasileira de Desenvolvimento Regional**, v. 9, n. 1, p. 175–196, 2021. Ivo Marcos Theis. Disponível em: <https://ojsrevista.furb.br/ojs/index.php/rbdr/article/view/9189>. Acesso em: 17 dez. 2025.

KRAUSE BIERHALZ, C. D.; BRUNO, J. C. Evasão nos cursos de graduação: um estudo de caso em uma universidade federal. **Devir Educação**, v. 9, n. 1, 2025. Devir Educação. Disponível em: <https://doi.org/10.30905/rde.v9i1.949>. Acesso em: 21 mar. 2026.

LADEIRA, P. P.; LIMA, L. M.; KROHLING, R. A. A visualization tool for data analysis on higher education dropout: a case study at UFES. **Computers and Society**, 2022. Disponível em: <https://arxiv.org/abs/2201.12637>. Acesso em: 26 mar. 2026.

LAKATOS, E. M.; MARCONI, M. DE A. **Fundamentos de metodologia científica**. 8 ed. São Paulo: Atlas, 2017.

LAMPROPOULOS, G.; EVANGELIDIS, G. Learning Analytics and Educational Data Mining in Augmented Reality, Virtual Reality, and the Metaverse: A Systematic Literature Review, Content Analysis, and Bibliometric Analysis. **Applied Sciences**, v. 15, n. 2, 2025. Multidisciplinary Digital Publishing Institute. Disponível em: <https://doi.org/10.3390/app15020971>. Acesso em: 13 nov. 2025.

LOBO, M. B. DE C. M. Panorama da evasão no ensino superior brasileiro: aspectos gerais das causas e soluções. **ABMES Cadernos**, n. 25, p. 9–58, 2012. Disponível em: <http://www.abmes.org.br/arquivos/publicacoes/Cadernos25.pdf>. Acesso em: 18 mar. 2026.

MACEDO, F. C. DE; MONTEIRO NETO, A.; VIEIRA, D. J. Universidade e território : ensino superior e desenvolvimento regional no Brasil do século XXI. **Instituto de Pesquisa Econômica Aplicada (Ipea)**, 2022. Disponível em: <http://repositorio.ipea.gov.br/handle/11058/11150>. Acesso em: 9 fev. 2026.

MAGALHÃES, P. V. S.; SOUSA, R. A. DA S.; SANTOS, A. C. B.; AGUIAR, F. J. M.; MILFONT, R. T. P. Modelos de Inteligência Artificial para Predição de Evasão escolar: Uma revisão sistemática. **Revista Eletrônica de Iniciação Científica em Computação**, v. 23, p. 48–54, 2025. Sociedade Brasileira de Computação - SB. Disponível em: <https://doi.org/10.5753/reic.2025.5523>. Acesso em: 2 fev. 2026.

MORAIS, D. M. G. de; SOUZA, D. K. F. de. MODELOS PREDITIVOS DA EVASÃO DE ESTUDANTES: COMPARATIVO ENTRE REGRESSÃO LOGÍSTICA, REDES NEURAIS E ALGORITMOS GENÉTICOS. **South American Development Society Journal**, v. 10, n. 30, p. 182–195, 2024. Disponível em: <https://doi.org/10.24325/issn.2446-5763.v10i30p182-195>. Acesso em: 22 fev. 2026.

MELO GONÇALVES, H. H. EVASÃO NAS INSTITUIÇÕES DE ENSINO SUPERIOR PÚBLICAS DO BRASIL: UMA ANÁLISE QUANTITATIVA. **RECIMA21: Revista Científica Multidisciplinar**, v. 5, n. 8, p. e585432, 2024. Editora RECIMA21 LTDA. Disponível em: <https://doi.org/10.47820/recima21.v5i8.5432>. Acesso em: 15 mar. 2026.

MYSTAKIDIS, Stylianos. **Metaverse. Encyclopedia**, v. 2, n. 1, p. 486-497, 2022. Disponível em: <https://doi.org/10.3390/encyclopedia2010031>. Acesso em: 20 dez. 2025.

NIEROTKA, R. L.; BONAMINO, A. M. C. DE; CARRASQUEIRA, K. Acesso, evasão e conclusão no Ensino Superior público: evidências para uma coorte de estudantes. **Ensaio: Avaliação e Políticas Públicas em Educação**, v. 31, n. 118, 2023. Disponível em: <https://doi.org/10.1590/S0104-40362022003003107>. Acesso em: 15 dez. 2025.

NIEROTKA, R. L.; SALATA, A.; KLITZKE MARTINS, M. FATORES ASSOCIADOS À EVASÃO NO ENSINO SUPERIOR: UM ESTUDO LONGITUDINAL. **Cadernos de Pesquisa**, v. 53, p. e09961, 2023. Disponível em: <https://doi.org/10.1590/198053149961>. Acesso em: 7 dez. 2025.

OECD. **Education at a Glance 2023: OECD Indicators**, 2023. Paris: OECD Publishing. Disponível em: <https://doi.org/10.1787/e13bef63-en>. Acesso em: 27 set. 2025.

PACHECO, R. D.; RIBEIRO, M. A. T.; SILVA, A. P. DA; et al. OS IMPACTOS DA INTELIGÊNCIA ARTIFICIAL NA SALA DE AULA. **REVISTA FOCO**, v. 17, n. 6, p.

e5429, 2024. Disponível em: <https://doi.org/10.54751/revistafoco.v17n6-104>. Acesso em: 8 set. 2026.

PAULA, G. B. DE; PICANÇO, F. DESIGUALDADES APÓS O ACESSO: ORIGEM SOCIAL E EVASÃO DO SISTEMA DE ENSINO SUPERIOR. **Educação & Sociedade**, v. 45, 2024. FapUNIFESP (SciELO). Disponível em: <https://doi.org/10.1590/ES.281915>. Acesso em: 15 mar. 2026.

PEARSON, K. Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material. **Philosophical Transactions of the Royal Society of London. A**, v. 186, p. 343–414, 1895. Londres.

PINTO, S. C. **Os custos da evasão de discentes das universidades brasileiras na modalidade de ensino presencial**: uma perspectiva de custos contábeis e custos econômicos, 2021. Dissertação (Mestrado em Ciências Contábeis), São Leopoldo: Universidade do Vale do Rio dos Sinos. Disponível em: <https://www.repositorio.jesuita.org.br/handle/UNISINOS/10835>. Acesso em: 9 dez. 2025.

QUEIROGA, E.; LOPES, J.; KAPPEL, K.; et al. A Learning Analytics Approach to Identify Students at Risk of Dropout: A Case Study with a Technical Distance Education Course. **Applied Sciences**, v. 10, p. 3998, 2020. Disponível em: <https://doi.org/10.3390/app10113998>. Acesso em: 3 fev. 2026.

QUIMIZ-MOREIRA, M.; DELGADILLO, R.; PARRAGA-ALAVA, J.; MACULAN, N.; MAURICIO, D. Factors, Prediction, Explainability, and Simulating University Dropout Through Machine Learning: A Systematic Review, 2012–2024. **Computation**, v. 13, n. 8, p. 198, 2025. Multidisciplinary Digital Publishing Institute. Disponível em: <https://doi.org/10.3390/computation13080198>. Acesso em: 7 jan. 2026.

ROMERO, C.; VENTURA, S. Educational data mining and learning analytics: An updated survey. **WIREs Data Mining Knowl Discov**, v. 10, p. e1355, 2020. Disponível em: <https://doi.org/10.1002/widm.1355>. Acesso em: 18 jan. 2026.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence**: a modern approach. 3 ed. New Jersey: Pearson, 2016.

SACCARO, A.; FRANÇA, M. T. A.; JACINTO, P. DE A. Fatores Associados à Evasão no Ensino Superior Brasileiro: um estudo de análise de sobrevivência para os cursos das áreas de Ciência, Matemática e Computação e de Engenharia, Produção e Construção em instituições públicas e privadas. **Estudos Econômicos (São Paulo)**, v. 49, p. 337–373, 2019. Disponível em: <https://doi.org/10.1590/0101-41614925amp>. Acesso em: 7 maio 2026.

SAMPIERI, R. H.; COLLADO, C. F.; LUCIO, M. P. B. **Metodologia de pesquisa**. 5 ed. Porto Alegre: Penso, 2013.

SCHMITT, J. A.; BERNARDI, G.; KANTORSKI, G. Z. iPermanency: uma ferramenta para gestão da evasão na EaD baseada em Mineração de Dados Educacionais. **RENOTE**, v. 17, n. 1, p. 306–315, 2019. Porto Alegre. Disponível em: <https://seer.ufrgs.br/index.php/renote/article/view/95798>. Acesso em: 7 maio 2026.

SHEARER, C. The CRISP-DM model: the new blueprint for data mining. **Journal of Data Warehousing**, v. 5, n. 4, p. 13–22, 2000.

SILVA, D. B. DA; FERRE, A. A. DE O.; GUIMARÃES, P. DOS S.; LIMA, R. DE; ESPINDOLA, I. B. Evasão no ensino superior público do Brasil: estudo de caso da Universidade de São Paulo. **Avaliação: Revista da Avaliação da Educação Superior**, v. 27, n. 2, p. 248–259, 2022. Disponível em: <https://doi.org/10.1590/S1414-40772022000200003>. Acesso em: 4 fev. 2026.

SILVA FILHO, R. L. L.; MOTEJUNAS, P. R.; HIPÓLITO, O.; LOBO, M. B. DE C. M. A evasão no ensino superior brasileiro. **Cadernos de Pesquisa**, v. 37, n. 132, p. 641–659, 2007. Disponível em: <https://doi.org/10.1590/S0100-15742007000300007>. Acesso em: 16 mar. 2026.

SILVA, J. J. DA; ROMAN, N. T. Predicting Dropout in Higher Education: a Systematic Review. **Anais do XXXII Simpósio Brasileiro de Informática na Educação (SBIE 2021)**, p. 1107–1117, 2021. Sociedade Brasileira de Computação - SBC. Disponível em: <https://doi.org/10.5753/sbie.2021.217437>. Acesso em: 2 fev. 2026.

SINNOTT, R. W. Virtues of the Haversine. **Sky and Telescope**, v. 68, p. 158, 1984.

SLADE, S.; PRINSLOO, P. Learning Analytics Ethical Issues and Dilemmas. **American Behavioral Scientist**, v. 57, n. 10, p. 1510–1529, 2013. Disponível em: <http://doi.org/10.1177/0002764213479366>. Acesso em: 13 fev. 2026.

SOARES, L. C. C. P.; CARVALHO, R. L. Sistema inteligente para previsão de evasão escolar em um ambiente universitário. **Humanidades & Inovação**, v. 10, n. 9, p. 242–257, 2023. Palmas. Disponível em: <https://revista.unitins.br/index.php/humanidadeseinovacao/article/view/9156>. Acesso em: 18 nov. 2025.

TCU. **Conhecendo o Tribunal**, v. 8, 2022. Brasília, DF: TCU. Disponível em: https://ejm.tjmmg.jus.br/wp-content/uploads/2024/12/Conhecendo_Tribunal_8_Edicao_portugues.pdf. Acesso em: 17 nov. 2025.

TCU. **Acórdão nº 1868/2024 – TCU – Plenário**, 2024. Brasília, DF: Tribunal de Contas da União. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A1868%2520ANOACORDAO%253A2024/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/0. Acesso em: 7 nov. 2025.

TERTULINO, R. Centralized vs. Federated Learning for Educational Data Mining: A Comparative Study on Student Performance Prediction with SAEB Microdata. **Brazilian Journal of Informatics in Education - RBIE**, 2025. Disponível em: <https://doi.org/10.48550/arXiv.2509.00086>. Acesso em: 7 maio 2026.

TINTO, V. Dropout from higher education: a theoretical synthesis of recent research. **Review of Educational Research**, v. 45, n. 1, p. 89–125, 1975. Disponível em: <https://www.jstor.org/stable/1170024>. Acesso em: 7 dez. 2025.

UFFS. Universidade Federal da Fronteira Sul. **Plano de Desenvolvimento Institucional (PDI) 2019–2023**. Chapecó: UFFS, 2019. Disponível em: https://site-antigo-2025.uffs.edu.br/institucional/a_uffs/a_instituicao/plano_de_desenvolvimento_institucional. Acesso em: 8 nov. 2025.

UFFS. Universidade Federal da Fronteira Sul. **Relatório de Auditoria Operacional nº 02/AUDIN/UFFS/2025: evasão nos cursos de graduação da Universidade Federal da Fronteira Sul**. Chapecó: UFFS, 2025. Disponível em: <https://www.uffs.edu.br/uffs/conteudo/AUDITORIA%20INTERNA%20-%20AUDIN/RA%2002-AUDIN-UFFS-2025.pdf>. Acesso em: 4 set. 2025.

VAARMA, M.; LI, H. Predicting student dropouts with machine learning: An empirical study in Finnish higher education. **Technology in Society**, v. 76, p. 102474, 2024. Elsevier BV. Disponível em: <https://doi.org/10.1016/j.techsoc.2024.102474>. Acesso em: 13 nov. 2025.

VALENTE, J. A. Inovação nos processos de ensino e de aprendizagem: o papel das tecnologias digitais. In: **Tecnologia e educação: passado, presente e o que está por vir**. p.17–41, 2018. Campinas: NIED/Unicamp. Disponível em: <https://www.nied.unicamp.br/wp-content/uploads/2018/11/Livro-NIED-2018-final.pdf>. Acesso em: 4 mar. 2026.

WEGNER, R. C.; ALVES JUNIOR, A. J.; FERNANDES, M. P. Evaluation of indicators for higher education: development of tools. **Avaliação: Revista da Avaliação da Educação Superior** (Campinas), v. 29, 2024. FapUNIFESP (SciELO). Disponível em: <https://doi.org/10.1590/1982-57652024v29id27881828>. Acesso em: 15 nov. 2025.

WIRTH, R.; HIPPI, J. CRISP-DM: Towards a standard process model for data mining. **Proceedings of the 4th International Conference on the Practical Applications of Knowledge Discovery and Data Mining**, 2000. Disponível em: https://www.researchgate.net/publication/239585378_CRISP-DM_Towards_a_standard_process_model_for_data_mining. Acesso em: 12 nov. 2025.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining: Practical Machine Learning Tools and Techniques**. 3 ed. Burlington: Morgan Kaufmann, 2011.

YIN, Robert K. **Estudo de caso: planejamento e métodos**. 5. ed. Porto Alegre: Bookman, 2015.

ZAGO, N. Do acesso à permanência no ensino superior: percursos de estudantes universitários de camadas populares. **Revista Brasileira de Educação**, v. 11, n. 32, p. 226–237, 2006. Disponível em: <https://doi.org/10.1590/S1413-24782006000200003>. Acesso em: 7 jan. 2026.

APÊNDICE A - PROTOCOLO DE IMPLEMENTAÇÃO DE UM SISTEMA DE ALERTA PRECOCE DE EVASÃO ACADÊMICA NA UFFS CAMPUS CHAPECÓ



PROTOCOLO DE IMPLEMENTAÇÃO DE UM SISTEMA DE ALERTA PRECOCE DE EVASÃO ACADÊMICA NA UFFS CAMPUS CHAPECÓ

Diretrizes para o uso institucional periódico de modelos de Mineração de Dados e Machine Learning

PROTOCOLO DE IMPLEMENTAÇÃO DE UM SISTEMA DE ALERTA PRECOCE DE EVASÃO ACADÊMICA NA UFFS CAMPUS CHAPECÓ

Relatório técnico apresentado pelo mestrando Jean Carlo Rodio ao Mestrado Profissional em Administração Pública em Rede Nacional (PROFIAP), sob orientação do Prof. Dr. Moacir Francisco Deimling e coorientação do Prof. Dr. Giancarlo Dondoni Salton, como parte dos requisitos para obtenção do título de Mestre em Administração Pública.



SUMÁRIO	Contexto de aplicação da proposta	04
	Público-alvo da proposta	05
	Situação problema	06
	Objetivo da proposta	08
	PTT – Proposta de Intervenção	09
	Autoria da Proposta	11
	Referências	12
	Protocolo de Entrega da Proposta aos Praticantes	13

CONTEXTO DE APLICAÇÃO DA PROPOSTA

A presente proposta de intervenção insere-se no contexto da Universidade Federal da Fronteira Sul (UFFS), campus Chapecó, instituição federal de ensino superior criada em 2009 com a missão de democratizar o acesso à educação superior na Mesorregião Grande Fronteira do Mercosul. Entre 2015 e 2023, a Taxa de Sucesso na Graduação da UFFS manteve-se em torno de 33%, patamar que o Relatório de Auditoria Operacional nº 02/AUDIN/UFFS/2025 classificou como tema de alto risco institucional, associado à ausência de mecanismos sistemáticos de monitoramento e identificação precoce de estudantes em risco de evasão (UFFS, 2025). No mesmo sentido, o Tribunal de Contas da União, por meio do Acórdão nº 1868/2024, recomendou às instituições federais de ensino superior a estruturação de políticas de prevenção à evasão fundamentadas em evidências e sustentadas por sistemas de monitoramento contínuo dos indicadores acadêmicos (TCU, 2024).

A proposta decorre diretamente da dissertação de mestrado intitulada "Aplicação de Técnicas de Mineração de Dados e *Machine Learning* para Predição da Evasão de Estudantes nos Cursos de Graduação de uma IFES", desenvolvida no âmbito do Mestrado Profissional em Administração Pública em Rede Nacional (PROFIAP), campus Chapecó, e posiciona-se no momento institucional em que a UFFS ainda carece de instrumento preditivo consolidado para subsidiar a gestão da permanência estudantil.

Do ponto de vista de seu alinhamento com a agenda internacional de desenvolvimento sustentável, a proposta dialoga diretamente com o Objetivo de Desenvolvimento Sustentável (ODS) 4 - Educação de Qualidade, da Organização das Nações Unidas, ao contribuir para a redução do abandono e para a garantia de conclusão dos estudos superiores. De forma complementar, ao evidenciar o papel de indicadores de engajamento institucional e de assistência estudantil - como o uso do Restaurante Universitário - na predição da evasão, a proposta também se relaciona ao ODS 10 - Redução das Desigualdades, na medida em que pode subsidiar políticas de permanência voltadas a estudantes oriundos de contextos socioeconômicos mais vulneráveis.



a proposta dialoga diretamente com o Objetivo de Desenvolvimento Sustentável (ODS) 4 - Educação de Qualidade, da Organização das Nações Unidas, ao contribuir para a redução do abandono e para a garantia de conclusão dos estudos superiores.

PÚBLICO-ALVO DA PROPOSTA

A proposta de intervenção destina-se, em primeiro lugar, à Pró-Reitoria de Graduação da Universidade Federal da Fronteira Sul (PROGRAD/UFFS), instância institucional com competência e autonomia para deliberar sobre a incorporação do protocolo aqui descrito aos fluxos de gestão da permanência estudantil, configurando-se como praticante-líder desta proposta.

A operacionalização técnica do protocolo depende, ainda, da equipe técnica da Secretaria Especial de Tecnologia da Informação (SETI) da UFFS, responsável pela infraestrutura dos sistemas SIGAA, SCI, SAS e Pergamum, e destina-se, em seus resultados finais, às coordenações de curso e às equipes de assistência estudantil do campus Chapecó, responsáveis por interpretar as estimativas de risco e conduzir, com autonomia pedagógica, as intervenções de acompanhamento.

DADOS



PRATICANTE-LÍDER

Pró-Reitoria de Graduação da UFFS (PROGRAD) - delibera sobre a incorporação institucional do protocolo.



EXECUÇÃO TÉCNICA

Secretaria Especial de Tecnologia da Informação (SETI/UFFS) - executa o *pipeline* e o modelo sobre os dados institucionais.



USUÁRIOS FINAIS

Coordenações de curso e equipes de assistência estudantil do campus Chapecó - interpretam os resultados e conduzem as intervenções.



SITUAÇÃO PROBLEMA

Diagnóstico, análise e linha de intervenção



Diagnóstico

O diagnóstico que fundamenta esta proposta foi conduzido ao longo da dissertação de mestrado que a antecede e pode ser sintetizado em três constatações centrais. Primeiro, a UFFS convive, desde sua criação, com taxas de sucesso na graduação estruturalmente baixas: a média histórica de 33,11% entre 2015 e 2023 significa que menos de um terço dos estudantes ingressantes concluem o curso no tempo previsto (UFFS, 2025). Segundo, essa taxa não é acompanhada de mecanismos institucionais de identificação precoce: o Relatório de Auditoria Operacional nº 02/AUDIN/UFFS/2025 constatou a ausência de política formal e de sistema contínuo de monitoramento da evasão

na instituição, limitação que o TCU (2024), identificou como recorrente em 51% das universidades federais brasileiras. Terceiro, a UFFS dispõe, ainda que de forma não integrada, de um volume expressivo de dados institucionais, acadêmicos, de assistência estudantil e de uso de serviços, que, até a realização deste estudo, não haviam sido explorados de forma sistemática para fins preditivos.

Análise

A análise que subsidiou a linha de intervenção seguiu a metodologia CRISP-DM (*Cross Industry Standard Process for Data Mining*) e integrou registros de quatro sistemas institucionais - SIGAA, SCI, SAS e Pergamum -, resultando em um conjunto analítico de 235.632 registros referentes a 8.749 matrículas de graduação do campus Chapecó, no período de 2014.1 a 2024.2. Sobre esse conjunto de dados foram treinados e comparados seis algoritmos de aprendizado supervisionado - Regressão Logística, Árvore de Decisão, *Random Forest*, XGBoost, SVM Linear e KNN, validados por meio do protocolo *StratifiedGroupKFold*, que assegura a separação estrita entre os conjuntos de treino e teste por matrícula e evita o vazamento de dados temporais entre observações de um mesmo estudante.

O modelo XGBoost apresentou o melhor desempenho global (AUC-ROC = 0,8878), seguido pela Regressão Logística (AUC-ROC = 0,8681) e pelo SVM Linear (AUC-ROC = 0,8680), o que demonstra que os dados já disponíveis nos sistemas da UFFS sustentam modelos com capacidade discriminativa robusta para identificar estudantes em risco de evasão. A análise de importância de variáveis revelou, ainda, que o desempenho acadêmico predomina na estrutura preditiva: a média final nas disciplinas respondeu por aproximadamente um terço do ganho informacional total do modelo, seguida por indicadores de ausência de registro acadêmico, número de faltas e área de conhecimento, ao passo que o uso do Restaurante Universitário nos primeiros semestres emergiu como preditor relevante de engajamento institucional, superando variáveis sociodemográficas como gênero e etnia.

Linha de intervenção

A linha de intervenção escolhida para converter esse diagnóstico em ação institucional consistiu na disponibilização de um *pipeline* de pré-processamento e de um modelo preditivo prontos para operacionalização periódica pela própria equipe técnica da UFFS, em vez de alternativas como a aquisição de sistemas comerciais de terceiros ou a criação de uma unidade dedicada exclusivamente à predição de evasão. Essa escolha justifica-se por três razões: primeiro, porque os recursos técnicos necessários à execução do *pipeline* (Python e bibliotecas de código aberto) já estão disponíveis na SETI/UFFS; segundo, porque a documentação integral do processo de tratamento de dados, codificação e modelagem, produzida ao longo da dissertação, permite a reprodução independente do processo sem custos de licenciamento; terceiro, porque a manutenção do conhecimento técnico internamente à instituição favorece a continuidade do protocolo.

OBJETIVO DA PROPOSTA

Esta proposta tem por objetivo dar materialidade à linha de intervenção definida a partir do diagnóstico e da análise apresentados na seção anterior, por meio da instituição de um protocolo de geração periódica de estimativas de risco de evasão para as matrículas ativas dos cursos de graduação do campus Chapecó da UFFS, a ser incorporado aos fluxos de gestão acadêmica da Pró-Reitoria de Graduação.

De modo específico, pretende-se: a) viabilizar a execução semestral do *pipeline* de pré-processamento e do modelo preditivo XGBoost sobre os dados extraídos dos sistemas SIGAA, SCI, SAS e Pergamum; b) subsidiar as coordenações de curso e as equipes de assistência estudantil com estimativas objetivas de risco de evasão por matrícula, organizadas em faixas de prioridade; c) orientar a calibração institucional dos limiares de corte que definem as faixas de risco, de acordo com a capacidade operacional de acompanhamento da instituição; d) assegurar que a implementação do protocolo observe a Lei Geral de Proteção de Dados Pessoais (LGPD), com pseudonimização dos identificadores durante o processamento técnico, acesso restrito aos profissionais autorizados e reidentificação somente quando necessária à realização do acompanhamento pedagógico ou assistencial.



"O escore de probabilidade de evasão é uma estimativa estatística, não uma certeza, devendo ser interpretado pelas coordenações de curso e equipes pedagógicas à luz do conhecimento que detêm sobre a realidade específica de cada estudante e de cada curso."

PTT – PROPOSTA DE INTERVENÇÃO

Recursos técnicos disponibilizados

Como parte concreta desta proposta, disponibilizam-se três recursos técnicos articulados, resultantes diretamente da dissertação que a fundamenta: a) o dataset consolidado, com 235.632 registros acadêmicos referentes a 8.749 matrículas do campus Chapecó no período de 2014.1 a 2024.2, integrando informações do SIGAA, do SCI, do SAS e do Pergamum, tratado e documentado conforme os procedimentos de qualidade de dados descritos na dissertação; b) o *pipeline* de pré-processamento, implementado em Python com as bibliotecas pandas, numpy e scikit-learn, executável pela equipe técnica da SETI/UFFS sobre novos dados extraídos a cada semestre letivo, reproduzindo de forma automatizada as etapas de tratamento, codificação e engenharia de variáveis documentadas; c) o modelo XGBoost treinado e validado (AUC-ROC = 0,8878), aplicável sobre o conjunto de dados preparado para gerar, a cada semestre, uma estimativa da probabilidade de evasão de cada matrícula ativa.

Protocolo de execução periódica

Recomenda-se que a Pró-Reitoria de Graduação avalie a incorporação da execução do *pipeline* e do modelo ao calendário institucional, segundo o seguinte ciclo: ao final de cada semestre letivo, a SETI/UFFS executa o *pipeline* sobre os dados atualizados dos quatro sistemas institucionais e aplica o modelo XGBoost às matrículas ativas; os resultados são consolidados em relatório de risco por matrícula e disponibilizados às coordenações de curso e às equipes de assistência estudantil no início do semestre subsequente, de modo que as informações utilizadas reflitam sempre a trajetória acadêmica mais recente de cada estudante.

Protocolo de execução periódica

Os limiares de corte que definem a classificação de cada matrícula em alto, médio ou baixo risco devem ser calibrados pela própria instituição, em função de sua capacidade operacional de acompanhamento, podendo ser ajustados ao longo dos primeiros ciclos de uso com base nos resultados observados. Recomenda-se, adicionalmente, que a UFFS reavalie periodicamente o desempenho do modelo, sobretudo após mudanças curriculares relevantes ou alterações expressivas no perfil dos ingressantes, de modo a assegurar que os padrões históricos sobre os quais o modelo foi treinado permaneçam representativos da realidade institucional.

Protocolo de execução periódica

A aplicação do modelo preditivo deve ser compreendida exclusivamente como recurso de apoio à decisão dos gestores acadêmicos, e não como mecanismo autônomo de classificação de estudantes. O tratamento dos dados pessoais envolvidos deve observar integralmente a Lei Geral de Proteção de Dados Pessoais (LGPD), com governança transparente e mecanismos de contestação pelos próprios estudantes, conforme já indicado por Slade e Prinsloo (2013) quanto ao uso ético de analytics em contextos educacionais.

Potencial de replicação

Embora o modelo tenha sido desenvolvido e validado exclusivamente para os dados do Campus Chapecó, o pipeline proposto possui potencial de replicação em outras instituições de ensino superior, públicas ou privadas. Sua aplicação requer novo processo de treinamento e validação com dados próprios de cada instituição, uma vez que os padrões de evasão podem variar conforme o contexto acadêmico, regional e o perfil dos estudantes. Assim, a principal possibilidade de replicação está na utilização da metodologia e do pipeline

desenvolvidos como referência para construção de modelos adaptados a diferentes realidades institucionais.

O quadro abaixo apresenta os modelos avaliados, seus respectivos *pipelines* de pré-processamento e os hiperparâmetros utilizados para replicação.

Modelo	Pipeline de pré-processamento e hiperparâmetros
Regressão Logística	SimpleImputer(median) → StandardScaler → LogisticRegression(C=0.5, class_weight='balanced', max_iter=1000, random_state=42)
Árvore de Decisão	SimpleImputer(median) → DecisionTreeClassifier(max_depth=5, min_samples_leaf=100, min_samples_split=200, ccp_alpha=0.0001, class_weight='balanced', random_state=42)
Random Forest	SimpleImputer(median) → RandomForestClassifier(n_estimators=200, max_depth=7, min_samples_leaf=50, min_samples_split=100, max_samples=0.8, oob_score=True, class_weight='balanced', n_jobs=-1, random_state=42)
XGBoost	SimpleImputer(median) → XGBClassifier(n_estimators=200, max_depth=4, learning_rate=0.1, min_child_weight=5, subsample=0.8, colsample_bytree=0.8, reg_lambda=2.0, scale_pos_weight=dinâmico, n_jobs=-1, random_state=42, eval_metric='logloss')
SVM (Linear)	SimpleImputer(median) → StandardScaler → CalibratedClassifierCV(LinearSVC(C=0.5, class_weight='balanced', max_iter=2000, random_state=42), cv=3, method='sigmoid')
KNN	SimpleImputer(median) → StandardScaler → KNeighborsClassifier(n_neighbors=31, weights='uniform', metric='manhattan', n_jobs=-1)

AUTORIA DA PROPOSTA

Discente

Jean Carlo Rodio

Servidor Técnico-Administrativo em Educação (TAE) da Universidade Federal da Fronteira Sul (UFFS) campus Chapecó e mestrando do Mestrado Profissional em Administração Pública em Rede Nacional (PROFIAP), campus Chapecó. Contato: jean.rodio@uffs.edu.br

Orientador

Prof. Dr. Moacir Francisco Deimling

Docente do Mestrado Profissional em Administração Pública em Rede Nacional (PROFIAP), Universidade Federal da Fronteira Sul (UFFS), campus Chapecó.

Coorientador

Prof. Dr. Giancarlo Dondoni Salton

Docente na Universidade Federal da Fronteira Sul (UFFS), campus Chapecó.

Data de finalização do documento: [20/07/2026]

REFERÊNCIAS

CHAPMAN, P.; CLINTON, J.; KERBER, R. et al. **CRISP-DM 1.0 Step-by-step data mining guide**. The CRISP-DM Consortium, 2000.

CHEN, T.; GUESTRIN, C. XGBoost: a Scalable Tree Boosting System. **Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16**, v. 1, n. 1, p. 785-794, 2016. Disponível em: 10.1145/2939672.2939785. Acesso em: 16 dez. 2025.

SLADE, S.; PRINSLOO, P. Learning Analytics Ethical Issues and Dilemmas. **American Behavioral Scientist**, v. 57, n. 10, p. 1510-1529, 2013. Disponível em: <http://doi.org/10.1177/0002764213479366>. Acesso em: 13 fev. 2026.

TCU. **Acórdão no 1868/2024 – TCU – Plenário**, 2024. Brasília, DF: Tribunal de Contas da União. Disponível em: https://pesquisa.apps.tcu.gov.br/documento/acordao-completo/*/NUMACORDAO%253A1868%2520ANOACORDAO%253A2024/DTRELEVANCIA%2520desc%252C%2520NUMACORDAOINT%2520desc/0. Acesso em: 7 nov. 2025.

UFFS. Universidade Federal da Fronteira Sul. **Relatório de Auditoria Operacional no 02/AUDIN/UFFS/2025: evasão nos cursos de graduação da Universidade Federal da Fronteira Sul**. Chapecó: UFFS, 2025. Disponível em: <https://www.uffs.edu.br/uffs/conteudo/AUDITORIA%20INTERNA%20-%20AUDIN/RA%2002-AUDIN-UFFS-2025.pdf>. Acesso em: 4 set. 2025.

PROTOCOLO DE ENTREGA DA PROPOSTA AOS PRATICANTES

A

Pró-Reitoria de Graduação (PROGRAD)
Universidade Federal da Fronteira Sul (UFFS)

Pelo presente, encaminhamos o produto técnico-tecnológico intitulado "**Protocolo de Implementação de um Sistema de Alerta Precoce de Evasão Acadêmica na UFFS Campus Chapecó**", derivado da dissertação de mestrado "Aplicação de Técnicas de Mineração de Dados e Machine Learning para Predição da Evasão de Estudantes nos Cursos de Graduação de uma IFES", de autoria de Jean Carlo Rodio.

Os documentos citados foram desenvolvidos no âmbito do Mestrado Profissional em Administração Pública em Rede Nacional (Profiap), instituição associada Universidade Federal da Fronteira Sul (UFFS).

A solução técnico-tecnológica é apresentada sob a forma de relatório técnico conclusivo e seu propósito é subsidiar, por meio da geração periódica de estimativas de risco de evasão, a identificação precoce de estudantes em risco de abandono nos cursos de graduação do campus Chapecó da UFFS.

Solicitamos, por gentileza, que ações voltadas à implementação desta proposição sejam informadas à Coordenação Local do Profiap, por meio do endereço **sec.profiap@uffs.edu.br**

Chapecó, SC, 06 de agosto de 2026.

Registro de recebimento



Documento assinado digitalmente
PEDRO ADALBERTO AGUIAR CASTRO
Data: 06/08/2026 13:48:40-0300
Verifique em <https://validar.it.gov.br>

Pedro Adalberto Aguiar Castro
Diretor de Registro Acadêmico
Pró-Reitoria de Graduação - UFFS

Discente: Jean Carlo Rodio, mestrando do PROFIAP
Orientador: Dr. Moacir Francisco Deimling, Professor da UFFS
Coorientador: Dr. Giancarlo Dondoni Salton, Professor da UFFS
UFFS – Universidade Federal da Fronteira Sul
Data de finalização do documento (20/07/2026)

ATRIBUTOS QUALIFICADORES DO PTT

IMPACTO

A implementação do protocolo tem potencial de impacto direto sobre a gestão da permanência estudantil na UFFS, ao oferecer um instrumento preditivo validado ($AUC-ROC = 0,8878$), capaz de identificar corretamente cerca de 80% dos estudantes que efetivamente evadem ($recall\ macro = 0,8038$), com base em dados já disponíveis nos sistemas institucionais. O impacto potencial mais direto é a possibilidade de priorizar, entre milhares de matrículas ativas, os casos que demandam atenção das coordenações de curso e das equipes de assistência estudantil, tarefa inviável por meio de análise humana direta em razão do volume de dados envolvido.

EVIDENCIAÇÃO DE IMPLEMENTAÇÃO E IMPACTO DA PROPOSTA

A implementação da solução permitirá que a gestão acadêmica do campus Chapecó da UFFS identifique, de forma periódica, estudantes com maior risco de evasão. As estimativas geradas pelo modelo poderão subsidiar o planejamento de ações preventivas, como contato individualizado, orientação acadêmica, encaminhamento aos setores de assistência estudantil e acompanhamento das dificuldades acadêmicas e socioeconômicas identificadas.

O impacto potencial da proposta consiste no aprimoramento da capacidade institucional de atuação preventiva, possibilitando que as equipes responsáveis direcionem esforços e recursos aos estudantes que apresentam maior necessidade de acompanhamento. A solução também poderá contribuir para a sistematização das informações, o fortalecimento da tomada de decisão baseada em dados e o aperfeiçoamento das políticas institucionais de permanência estudantil.

Após sua implementação, os resultados poderão ser acompanhados por meio de indicadores como: número de estudantes identificados pelo modelo, quantidade de estudantes atendidos, ações de acompanhamento realizadas, permanência dos estudantes sinalizados e evolução das taxas de evasão nos cursos de graduação. Essas informações permitirão avaliar periodicamente a efetividade da solução e realizar os ajustes necessários.