



**UNIVERSIDADE FEDERAL DE SANTA CATARINA
CAMPUS ARARANGUÁ
PROGRAMA DE PÓS-GRADUAÇÃO EM TECNOLOGIAS DA INFORMAÇÃO E
COMUNICAÇÃO**

Markson Rebelo Marcolino

**Análise Preditiva do Desempenho Acadêmico em Ambientes Virtuais de
Aprendizagem: Uma Abordagem com Aprendizado de Máquina Otimizado**

Araranguá
2025

Markson Rebelo Marcolino

**Análise Preditiva do Desempenho Acadêmico em Ambientes Virtuais de
Aprendizagem: Uma Abordagem com Aprendizado de Máquina Otimizado**

Dissertação apresentada ao Programa de Pós-Graduação em Tecnologias da Informação e Comunicação (PPGTIC) da Universidade Federal de Santa Catarina (UFSC) como requisito parcial para a obtenção do título de Mestre em Tecnologias da Informação e Comunicação. Área de concentração: Tecnologia Computacional.
Orientador: Prof. Dr. Cristian Cechinel.

Araranguá
2025

Ficha catalográfica gerada por meio de sistema automatizado gerenciado pela BU/UFSC.
Dados inseridos pelo próprio autor.

Rebelo Marcolino, Markson

Análise Preditiva do Desempenho Acadêmico em Ambientes Virtuais de Aprendizagem : Uma Abordagem com Aprendizado de Máquina Otimizado / Markson Rebelo Marcolino ; orientador, Cristian Cechinel, 2025.

94 p.

Dissertação (mestrado) - Universidade Federal de Santa Catarina, Campus Araranguá, Programa de Pós-Graduação em Tecnologias da Informação e Comunicação, Araranguá, 2025.

Inclui referências.

1. Tecnologias da Informação e Comunicação. 2. Previsão de Evasão Escolar. I. Cechinel, Cristian. II. Universidade Federal de Santa Catarina. Programa de Pós-Graduação em Tecnologias da Informação e Comunicação. III. Título.

Markson Rebelo Marcolino

Análise Preditiva do Desempenho Acadêmico em Ambientes Virtuais de Aprendizagem: Uma Abordagem com Aprendizado de Máquina Otimizado

O presente trabalho em nível de Mestrado foi avaliado e aprovado, em 16 de junho de 2025, pela banca examinadora composta pelos seguintes membros:

Eliane Pozzebon, Dr.

Membro titular interno – UFSC

Emanuel Marques Queiroga, Dr.

Membro titular externo – IFSUL

Rodrigo Lins Rodrigues, Dr.

Membro externo – UFRPE

Certificamos que esta é a versão original e final do trabalho de conclusão que foi julgado adequado para obtenção do título de Mestre em Tecnologias da Informação e Comunicação pelo Programa de Pós-Graduação em Tecnologias da Informação e Comunicação da Universidade Federal de Santa Catarina.



Documento assinado digitalmente

Giovani Mendonça Lunardi

Data: 26/08/2025 14:58:01-0300

CPF: ***.394.559-**

Verifique as assinaturas em <https://v.ufsc.br>

Coordenação do Programa de Pós-Graduação



Documento assinado digitalmente

Cristian Cechinel

Data: 25/08/2025 06:59:36-0300

CPF: *** 628.449-**

Verifique as assinaturas em <https://v.ufsc.br>

Cristian Cechinel, Dr.

Orientador

Araranguá
2025

*Dedico esta dissertação à minha família, pelo amor incondicional e pelo apoio em
cada etapa desta jornada.*

*À minha esposa, por sua paciência, compreensão e incentivo constantes, que me
motivaram a seguir em frente mesmo nos momentos mais desafiadores.*

*Ao meu orientador, Professor Cristian Cechinel, pela orientação, dedicação e por
acreditar no meu potencial, guiando-me com sabedoria e inspiração ao longo deste
caminho.*

Agradecimentos

A realização deste trabalho não teria sido possível sem o apoio e incentivo de pessoas especiais que contribuíram de diversas formas ao longo desta jornada.

À minha esposa, pelo apoio incondicional, paciência e incentivo constante em toda essa trajetória, sempre me motivando e ajudando no que podia. Seu suporte foi essencial para que eu pudesse seguir em frente e concluir esta etapa.

À minha família, pelo amor e apoio inabalável desde o início da minha trajetória acadêmica. Cada palavra de incentivo e cada gesto de carinho foram fundamentais para que eu superasse os desafios e conquistasse mais este objetivo.

Ao meu orientador, Professor Dr. Cristian Cechinel, pela confiança no meu potencial, pela orientação precisa e pelo suporte contínuo ao longo de todo o desenvolvimento deste trabalho. Sua dedicação e conhecimento foram essenciais para que eu pudesse aprimorar minhas habilidades e aprofundar minha pesquisa.

À equipe do *LANSE* pelo apoio no projeto, em especial ao Thiago Reis Porto, pelo auxílio e colaboração constantes.

Aos professores do *PPGTIC*, que contribuíram para minha formação acadêmica e científica. Em especial, ao Professor Dr. Juarez Bento da Silva, cuja empatia, dedicação e sabedoria sempre se destacaram, inspirando seus alunos e incentivando o pensamento crítico e a busca pelo conhecimento.

A todos e todas as pessoas amigas que, de alguma forma ou outra, contribuíram ou me desafiaram a alcançar a graça do mestrado ofertado pelo Programa de Pós-Graduação em Tecnologias da Informação e Comunicação da Universidade Federal de Santa Catarina, meu muito obrigado.

À Universidade Federal de Santa Catarina, por oferecer um ensino público, gratuito e de qualidade, permitindo que tantos estudantes possam buscar conhecimento e transformar suas vidas.

A todos aqueles que, de alguma forma, contribuíram para que este trabalho se tornasse realidade, meus mais sinceros agradecimentos.

*“Não há ensino sem pesquisa e pesquisa sem ensino.
Esses quehaceres se encontram um no corpo do outro.”*
(Paulo Freire)

Resumo

A evasão estudantil e o baixo desempenho acadêmico continuam sendo desafios persistentes na educação, frequentemente ocorrendo em taxas substanciais e apresentando dificuldades consideráveis para identificação e intervenção oportunas. Sistemas de gerenciamento de aprendizagem, como o Moodle, geram extensos conjuntos de dados que refletem as interações dos alunos e os padrões de matrícula, apresentando oportunidades para a aplicação de análise preditiva.

Este estudo busca avançar na área de previsão de evasão e baixo desempenho acadêmico por meio da aplicação de inteligência artificial com metodologias de aprendizado de máquina. Em particular, empregamos o algoritmo CatBoost, treinado em registros de atividade estudantil da plataforma Moodle. Para mitigar os desafios impostos por um conjunto de dados limitado e desbalanceado, utilizamos técnicas sofisticadas de balanceamento de dados, como o Adaptive Synthetic Sampling (ADASYN), e realizamos otimização multiobjetivo de hiperparâmetros usando o Algoritmo Genético de Ordenação Não Dominada II (NSGA-II).

Comparamos modelos treinados com dados semanais de logs com um único modelo treinado com dados de todas as semanas. O modelo proposto, treinado com os dados de todas as semanas, demonstrou desempenho superior, apresentando melhorias significativas nas pontuações F1 e recall, especialmente para a classe minoritária de alunos em risco. Por exemplo, o modelo obteve uma pontuação F1 média ao longo de múltiplas semanas de aproximadamente 0,8 no teste holdout. Esses achados destacam o potencial das abordagens de aprendizado de máquina direcionadas para facilitar a identificação precoce de alunos em risco, permitindo intervenções oportunas e melhorando os resultados educacionais.

Palavras-chave: Previsão de Evasão Escolar, Aprendizado de Máquina na Educação, NSGA-II, CatBoost, Logs do Moodle.

Abstract

Student attrition and academic failure remain pervasive challenges in education, often occurring at substantial rates and posing considerable difficulties for timely identification and intervention. Learning management systems such as Moodle generate extensive datasets reflecting student interactions and enrollment patterns, presenting opportunities for predictive analytics.

This study seeks to advance the field of dropout and failure prediction through the application of artificial intelligence with machine learning methodologies. In particular, we employed the CatBoost algorithm, trained on student activity logs from the Moodle platform. To mitigate the challenges posed by a limited and imbalanced dataset, we employed sophisticated data balancing techniques, such as Adaptive Synthetic Sampling (ADASYN), and conducted multi-objective hyperparameter optimization using the Non-dominated Sorting Genetic Algorithm II (NSGA-II).

We compared models trained on weekly log data against a single model trained on all weeks' data. The proposed model trained with all weeks' data demonstrated superior performance, showing significant improvements in F1-scores and recall, particularly for the minority class of at-risk students. For example, the model got an average F1-score across multiple weeks of approximately 0.8 in the holdout test. These findings underscore the potential of targeted machine learning approaches to facilitate early identification of at-risk students, thereby enabling timely interventions and improving educational outcomes.

Keywords: Student Dropout Prediction, Machine Learning in Education, NSGA-II, CatBoost, Moodle Logs.

Lista de ilustrações

Figura 1 – Visão geral da metodologia Fonte: Elaborado pelo autor	47
Figura 2 – Distribuição das Categorias de Cursos nos Conjuntos de Treinamento e Teste (<i>Experimento1</i>) Fonte: Elaborado pelo autor	51
Figura 3 – Distribuição das Categorias de Cursos nos Conjuntos de Treinamento e Teste (<i>Experimento2</i>) Fonte: Elaborado pelo autor	52
Figura 4 – Frente de Pareto 3D da busca NSGA-II mostrando a convergência das populações em termos de acurácia, F1-score macro e F1-score para alunos reprovados. Fonte: Elaborado pelo autor	61
Figura 5 – Comparação de Metodologia do Modelo - F1-score (Aprovados) Fonte: Elaborado pelo autor	64
Figura 6 – Comparação de Metodologia do Modelo - F1-score (Reprovados) Fonte: Elaborado pelo autor	65
Figura 7 – Comparação de Metodologia do Modelo - AUC Fonte: Elaborado pelo autor	66
Figura 8 – Valores SHAP para importância das características. Vermelho representa valores mais altos da característica, e azul representa valores mais baixos. O eixo x mostra o impacto na saída do modelo. Fonte: Elaborado pelo autor	67

Lista de tabelas

Tabela 1 – Resumo dos Trabalhos Relacionados Fonte: Elaborado pelo autor .	42
Tabela 2 – Resumo das Variáveis Utilizadas no Modelo	48
Tabela 3 – Resumo da Distribuição do Conjunto de Dados para os Experimentos 1 e 2 Fonte: Elaborado pelo autor	51
Tabela 4 – Resultados dos Testes KS e Mann-Whitney U para os Experimentos 1 e 2 Fonte: Elaborado pelo autor	54
Tabela 5 – Comparação de Desempenho no Conjunto de Teste Holdout. Novos modelos possuem um modelo distinto para cada semana (abordagem antiga). Fonte: Elaborado pelo autor	62
Tabela 6 – Comparação de Desempenho no Conjunto de Teste Holdout. Modelo único vs múltiplos modelos por semana	63
Tabela 7 – Comparação de Métricas de Desempenho entre Estudos	71

Lista de abreviaturas e siglas

ADASYN	<i>Adaptive Synthetic Sampling</i> – Técnica de oversampling que ajusta a geração de amostras sintéticas com base na distribuição dos dados.
AUC	<i>Area Under the Curve</i> – Área sob a curva ROC, utilizada para avaliar modelos de classificação.
AUC-PR	<i>Area Under the Precision-Recall Curve</i> – Área sob a curva de Precisão-Recall, útil para modelos em dados desbalanceados.
AVA	<i>Ambiente Virtual de Aprendizagem</i> – Plataformas online utilizadas no ensino digital.
AutoML	<i>Automated Machine Learning</i> – Técnicas automatizadas para ajuste de modelos de aprendizado de máquina.
BEED	<i>Bachelor of Elementary Education</i> – Licenciatura em Educação Primária.
BSED-Math	<i>Bachelor of Secondary Education in Mathematics</i> – Licenciatura em Matemática.
CatBoost	<i>Categorical Boosting</i> – Algoritmo baseado em Gradient Boosting otimizado para variáveis categóricas.
CNN	<i>Convolutional Neural Network</i> – Rede Neural Convolucional, usada principalmente para análise de imagens e séries temporais.
EBC	<i>Explainable Boosting Classifier</i> – Classificador baseado em Boosting explicável.
EDM	<i>Elementos da Mineração de Dados Educacionais</i>
FDC	<i>Função de Distribuição Acumulada</i>
FWER	<i>Family-Wise Error Rate</i> - Taxa de erro por família
GPA	<i>Grade Point Average</i> – Média de notas acadêmicas.
I_SMOTE	<i>Improved Synthetic Minority Over-sampling Technique</i> – Versão aprimorada do SMOTE para balanceamento de classes.
KS	<i>Kolmogorov-Smirnov</i> – Teste estatístico não paramétrico usado para comparar distribuições.

U-Test	<i>Mann-Whitney U Test</i> – Teste estatístico não paramétrico para comparar medianas de duas amostras independentes.
k-NN	<i>k-Nearest Neighbors</i> – Algoritmo baseado na proximidade de vizinhos para classificação e regressão.
LANSE	<i>Learning Analytics em Nuvem para Sistemas Educacionais</i> .
LightGBM	<i>Light Gradient Boosting Machine</i> – Algoritmo baseado em Gradient Boosting otimizado para eficiência computacional.
LIME	<i>Local Interpretable Model-Agnostic Explanations</i> – Método para explicar modelos de aprendizado de máquina.
LMS	<i>Learning Management System</i> – Sistema de Gerenciamento de Aprendizagem, como o Moodle.
ML	<i>Machine Learning</i> – Aprendizado de Máquina.
MORL	<i>Multi-Objective Reinforcement Learning</i> – Aprendizado por reforço multiobjetivo.
Moodle	<i>Modular Object-Oriented Dynamic Learning Environment</i> – Plataforma de ensino online.
NGBoost	<i>Natural Gradient Boosting</i> – Algoritmo baseado em boosting e inferência Bayesiana.
NSGA-II	<i>Non-dominated Sorting Genetic Algorithm II</i> – Algoritmo genético multiobjetivo.
PCA	<i>Principal Component Analysis</i> – Análise de Componentes Principais, usada para redução de dimensionalidade.
PI	<i>Permutation Importance</i> – Método para medir a importância das características em modelos de ML.
PDP	<i>Partial Dependence Plot</i> – Técnica para visualizar o efeito de uma variável sobre a predição do modelo.
RNA	<i>Rede Neural Artificial</i> .
1DCNN	<i>One-Dimensional Convolutional Neural Network</i> – Rede neural convolucional unidimensional.
SHAP	<i>SHapley Additive Explanations</i> – Técnica para explicar a importância das características nos modelos.

SMOTE	<i>Synthetic Minority Over-sampling Technique</i> – Técnica de oversampling usada para balanceamento de classes.
SMOTEENN	<i>SMOTE + Edited Nearest Neighbors</i> – Combinação de oversampling (SMOTE) com remoção de ruídos.
SVM	<i>Support Vector Machine</i> – Máquina de Vetores de Suporte, usada para classificação e regressão.
STEM	<i>Science, Technology, Engineering, and Mathematics</i> – Ciências, Tecnologia, Engenharia e Matemática.
XAI	<i>Explainable Artificial Intelligence</i> – Inteligência Artificial Explicável.
XGBoost	<i>eXtreme Gradient Boosting</i> – Algoritmo de boosting altamente eficiente.
LSTM	<i>Long Short-Term Memory</i> – Rede neural recorrente otimizada para séries temporais.

Sumário

1	INTRODUÇÃO	16
1.1	PROBLEMA DE PESQUISA	17
1.2	JUSTIFICATIVA	18
1.3	OBJETIVOS	18
1.3.1	Objetivo Geral	19
1.3.2	Objetivos Específicos	19
1.4	INTERDISCIPLINARIDADE E ADERÊNCIA AO PPGTIC	19
1.5	ESTRUTURA DA DISSERTAÇÃO	20
2	FUNDAMENTAÇÃO TEÓRICA	21
2.1	LEARNING ANALYTICS	21
2.2	MOODLE	23
2.3	EVASÃO ESTUDANTIL E PREVISÃO DE BAIXO DESEMPENHO ACADÊMICO	24
2.4	FUNDAMENTOS DE MACHINE LEARNING	25
2.5	MACHINE LEARNING NA EDUCAÇÃO: APLICAÇÕES E DESAFIOS	27
2.6	MODELOS DE MACHINE LEARNING PARA PREVISÃO DE EVASÃO ESCOLAR	28
2.7	O CatBoost e Sua Aplicação na Predição de Reprovação	29
2.8	FONTES DE DADOS E ENGENHARIA DE CARACTERÍSTICAS	30
2.9	MÉTRICAS DE AVALIAÇÃO DE MODELOS PREDITIVOS	30
2.9.1	Acurácia na Previsão de Evasão Escolar	30
2.9.2	Recall na Previsão de Evasão Escolar	32
2.9.3	F1-Score na Previsão de Evasão Escolar	33
2.9.4	AUC-ROC na Previsão de Evasão Escolar	35
2.10	TÉCNICAS DE BALANCEAMENTO E OTIMIZAÇÃO DE MODELOS	36
2.11	OTIMIZAÇÃO MULTIOBJETIVO COM NSGA-II	38
2.12	TESTES ESTATÍSTICOS PARA COMPARAÇÃO DE DISTRIBUIÇÕES	39
2.13	DESAFIOS E LIMITAÇÕES	41
2.14	CONSIDERAÇÕES SOBRE O ESTADO DA ARTE E CONTRIBUIÇÕES DESTE TRABALHO	45
3	METODOLOGIA	47
3.1	DESCRIÇÃO E PRÉ-PROCESSAMENTO DOS DADOS	48
3.2	ESTRATÉGIA DE OTIMIZAÇÃO	52

3.3	MEDIÇÕES ESTATÍSTICAS E DE DISTÂNCIA PARA AVALIAÇÃO DO MODELO	53
3.4	TREINAMENTO E AVALIAÇÃO DO MODELO	54
3.5	ENGENHARIA DE CARACTERÍSTICAS	55
3.6	REPRODUTIBILIDADE DO ESTUDO	55
4	A PLATAFORMA LANSE	57
5	RESULTADOS	59
5.1	MODELOS EXISTENTES PARA COMPARAÇÃO	59
5.2	MODELO PROPOSTO	60
5.3	MÉTRICAS DE DESEMPENHO	62
5.3.1	Comparação de Desempenho do Modelo	63
5.3.1.1	F1-Score para Alunos Aprovados	63
5.3.1.2	F1-Score para Alunos Reprovados	64
5.4	Comparação do AUC Score	64
5.5	Importância das Características	65
6	DISCUSSÃO	68
6.1	ANÁLISE DE ERROS E CONSIDERAÇÕES PEDAGÓGICAS	69
6.2	IMPORTÂNCIA DAS VARIÁVEIS: INTERPRETAÇÃO COM SHAP	69
6.3	DISCUSSÃO SOBRE A ESTRATÉGIA DE MODELAGEM: MODELO ÚNICO VS MÚLTIPLOS MODELOS	70
6.4	COMPARAÇÃO DE MÉTRICAS COM ESTUDOS EXISTENTES	71
6.5	QUESTÕES DE PESQUISA	74
6.6	LIMITAÇÕES E CONSIDERAÇÕES SOBRE GENERALIZAÇÃO	76
6.7	CONTRIBUIÇÕES CIENTÍFICAS	76
7	CONSIDERAÇÕES FINAIS	77
7.1	CONCLUSÃO	77
7.2	TRABALHOS FUTUROS	78
7.3	ARTIGO PUBLICADO	79

1 INTRODUÇÃO

A evasão e a reprovação acadêmica são desafios persistentes no ensino superior, afetando diretamente a trajetória dos estudantes e comprometendo a eficácia dos cursos ofertados pelas instituições. Essas taxas variam consideravelmente, de 30% a 80%, conforme a região, a instituição e a natureza do curso (JENKINS, 2002; DUNICAN, 2002; VIHAVAINEN; PAKSULA; LUUKKAINEN, 2011; BENNEDSEN; CASPERSEN, 2007; WATSON; LI, 2014; SHYNARBEEK et al., 2022). Disciplinas como programação, oferecidas geralmente nos primeiros anos de cursos de graduação, apresentam maiores taxas de abandono, dada a exigência de pensamento abstrato e resolução de problemas complexos (CHEAH, 2020). Desafios semelhantes ocorrem em outras áreas STEM, como matemática e engenharia, em que a carga cognitiva elevada constitui uma barreira significativa à aprendizagem bem-sucedida (GUPTA; ZHENG, 2020).

Para enfrentar esses problemas, diferentes estratégias pedagógicas têm sido propostas, especialmente aquelas que promovem o engajamento ativo dos alunos e a oferta de feedback formativo contínuo. Neste trabalho, adotamos a definição de engajamento proposta por Fredricks, Blumenfeld e Paris (2004), que o compreende como um construto multidimensional composto por três componentes interdependentes: engajamento comportamental (participação em atividades acadêmicas), engajamento emocional (reações afetivas em relação aos colegas, professores e à instituição) e engajamento cognitivo (esforço e investimento mental para compreender o conteúdo). Estudos recentes reforçam a importância de considerar tanto fatores acadêmicos quanto não acadêmicos, como condições socioeconômicas e circunstâncias pessoais, na predição de desempenho e na compreensão das causas da evasão (SHOHAG; BAKAUL, 2021).

Paralelamente, a área de *Learning Analytics* tem avançado como uma abordagem promissora para identificar estudantes em risco, sobretudo em contextos de ensino remoto e híbrido. Com o uso de Sistemas de Gerenciamento da Aprendizagem (LMS), como o Moodle, torna-se possível monitorar padrões de acesso e comportamento dos estudantes em tempo real, permitindo a detecção precoce de sinais de risco e favorecendo a adoção de intervenções educativas oportunas (QUEIROGA et al., 2020).

Neste contexto, o projeto **LANSE** (Learning Analytics em Nuvem para Sistemas Educacionais) introduziu uma plataforma analítica baseada em nuvem para auxiliar na identificação de estudantes em risco de evasão ou reprovação. A plataforma utiliza dados extraídos do Moodle e modelos de aprendizado de máquina (*Machine Learning* – ML) treinados semanalmente para detectar padrões de comportamento associados ao insucesso escolar (CECHINEL et al., 2024).

Entretanto, os modelos originais do LANSE enfrentam limitações relacionadas à generalização para novos cursos e à alta correlação entre ausência de acesso e reprovação. Diante disso, este trabalho visa propor melhorias nos modelos preditivos existentes, explorando abordagens alternativas de modelagem, otimização e representação temporal dos dados de log.

Com base nesses desafios, este estudo é guiado pelas seguintes questões de pesquisa:

- **Questão 1:** *Quão eficazmente os modelos de aprendizado de máquina, treinados com dados limitados de logs estudantis do Moodle, podem generalizar para prever resultados em conjuntos de dados maiores e diversificados, incluindo cursos não vistos anteriormente?*
- **Questão 2:** *Quais características nos logs estudantis são mais preditivas do sucesso ou engajamento, e essas características permanecem igualmente importantes em diferentes cursos?*
- **Questão 3:** *O algoritmo de otimização multiobjetivo NSGA-II pode efetivamente melhorar a acurácia preditiva e a confiabilidade dos modelos de aprendizado de máquina na identificação de alunos em risco em cenários de predição de evasão?*
- **Questão 4:** *Como a acurácia preditiva e a estabilidade de um modelo único treinado com dados de todas as semanas se comparam com modelos separados treinados para cada semana?*

1.1 PROBLEMA DE PESQUISA

Altas taxas de evasão e reprovação são desafios significativos no domínio educacional, afetando uma ampla variedade de cursos em diversas disciplinas. Essas taxas podem variar drasticamente, de 30% a 80%, dependendo da região, instituição e tipo específico de curso (JENKINS, 2002; DUNICAN, 2002; VIHAVAINEN; PAKSULA; LUUKKAINEN, 2011; BENNEDSEN; CASPERSEN, 2007; WATSON; LI, 2014; SHY-NARBEK et al., 2022). Em cursos como programação, geralmente oferecidos nos primeiros anos da graduação, os alunos frequentemente enfrentam dificuldades com o pensamento abstrato e a resolução de problemas complexos, resultando em taxas mais altas de evasão (CHEAH, 2020).

Desafios semelhantes são observados em disciplinas que historicamente apresentam maiores taxas de reprovação e evasão, especialmente nas áreas de Ciência, Tecnologia, Engenharia e Matemática (STEM), como matemática, programação e engenharia. Embora fatores como condições socioeconômicas e habilidades cognitivas e metacognitivas também desempenhem um papel, a elevada carga cognitiva e a

complexidade conceitual desses conteúdos representam barreiras significativas para o aprendizado bem-sucedido (GUPTA; ZHENG, 2020; CHEAH, 2020).

1.2 JUSTIFICATIVA

Diversas estratégias pedagógicas têm sido propostas para enfrentar esses desafios, enfatizando o envolvimento ativo dos alunos em diversas atividades educacionais. Pesquisas sugerem que essas estratégias são mais eficazes quando combinadas com feedback contínuo e formativo dos instrutores, o que pode ajudar a identificar e resolver dificuldades dos alunos precocemente.

Estudos anteriores destacam a importância de considerar fatores acadêmicos e não acadêmicos, como condições socioeconômicas e circunstâncias pessoais, ao prever o desempenho dos alunos e as taxas de evasão (SHOHAG; BAKAUL, 2021). Ao integrar essas influências mais amplas, é possível obter uma compreensão mais abrangente das dificuldades enfrentadas pelos alunos.

Ao mesmo tempo, há estudos que destacam o potencial da análise de aprendizagem (*learning analytics*) na identificação de alunos em risco, particularmente em contextos de educação online ou a distância (QUEIROGA et al., 2020). Esses estudos enfatizam como a análise de padrões de engajamento e atividades dos alunos dentro dos Sistemas de Gerenciamento de Aprendizagem (LMS) pode permitir a detecção precoce dos riscos de evasão. A identificação precoce é crucial para reduzir as taxas de reprovação, pois possibilita intervenções oportunas e direcionadas antes que os alunos desistam ou abandonem completamente os cursos.

1.3 OBJETIVOS

Nos últimos anos, a modelagem preditiva usando ML emergiu como uma ferramenta poderosa para identificar alunos em risco em diversos cursos. Esses modelos analisam uma ampla gama de dados, incluindo interações em ambientes virtuais de aprendizagem, avaliações contínuas e informações demográficas (KUH et al., 2014; AGUIAR et al., 2018; JAIN; SOLANKI, 2019; MACARINI et al., 2019; QUEIROGA et al., 2020).

Ao utilizar esses dados, os modelos preditivos podem fornecer insights sobre quais alunos estão em risco e por que estão enfrentando dificuldades, permitindo que educadores implementem intervenções proativas e personalizadas. Por exemplo, alguns estudos demonstraram a eficácia de sistemas de alerta precoce baseados em aprendizado de máquina na melhoria da retenção de alunos em múltiplas disciplinas (KRÜGER; BRITTO; BARDDAL, 2023).

1.3.1 Objetivo Geral

Desenvolver e avaliar modelos preditivos baseados em aprendizado de máquina para identificar estudantes em risco de evasão ou reprovação, utilizando dados de log do Moodle, com foco na melhoria da generalização para novos cursos, na seleção de atributos relevantes e na aplicação de técnicas de balanceamento e otimização multiobjetivo.

1.3.2 Objetivos Específicos

- Desenvolver um novo modelo de previsão de evasão baseado no CatBoost, incorporando técnicas de categorização de logs e otimização multiobjetivo utilizando o algoritmo NSGA-II.
- Comparar o desempenho do modelo proposto com os modelos existentes implementados na plataforma LANSE, utilizando os mesmos dados de treinamento e novos dados como teste *holdout*.
- Avaliar a eficácia de um modelo único treinado com dados de todas as semanas do curso, em comparação com a abordagem tradicional de treinar um modelo separado para cada semana.
- Incorporar ao modelo características relacionadas à estrutura temporal dos cursos, como a duração total do curso e a posição relativa da semana, a fim de melhorar a capacidade de generalização em diferentes contextos.
- Aplicar técnicas de balanceamento de dados, como o ADASYN, para lidar com o desbalanceamento entre alunos aprovados e reprovados e melhorar a detecção de alunos em risco.
- Otimizar hiperparâmetros dos modelos de forma simultânea com múltiplos objetivos (ex.: acurácia, F1-score para aprovados e reprovados), utilizando o NSGA-II como estratégia evolutiva.

1.4 INTERDISCIPLINARIDADE E ADERÊNCIA AO PPGTIC

O Programa de Pós-Graduação em Tecnologias da Informação e Comunicação (PPGTIC) tem como área de concentração Tecnologia e Inovação, sendo estruturado em três linhas de pesquisa: Tecnologia Computacional, Tecnologia Educacional e Tecnologia, Gestão e Inovação. Este estudo está inserido na linha de pesquisa Tecnologia Computacional, que busca explorar métodos, técnicas e ferramentas computacionais para solucionar problemas complexos em diversas áreas do conhecimento.

A pesquisa apresentada nesta dissertação possui uma abordagem interdisciplinar ao combinar conceitos de aprendizado de máquina, análise de dados educacionais e otimização computacional para prever a evasão de alunos em cursos online. A interseção entre essas áreas permite um avanço significativo na análise preditiva aplicada à educação, oferecendo soluções inovadoras que podem ser utilizadas tanto por educadores quanto por gestores acadêmicos.

Além disso, a pesquisa está alinhada aos objetivos do PPGTIC ao desenvolver um modelo computacional capaz de aprimorar a gestão da educação baseada em dados, utilizando técnicas avançadas de inteligência artificial. Ao explorar abordagens como aprendizado de máquina supervisionado, algoritmos evolutivos e técnicas de balanceamento de dados, este trabalho contribui para a inovação na tecnologia computacional aplicada à análise da aprendizagem, fortalecendo o uso de ferramentas digitais para a tomada de decisões estratégicas no contexto educacional.

1.5 ESTRUTURA DA DISSERTAÇÃO

O restante do trabalho está estruturado da seguinte maneira. O Capítulo 2 apresenta a fundamentação teórica relacionada à temática da pesquisa, abordando conceitos centrais como *Learning Analytics*, evasão estudantil, fundamentos e aplicações de técnicas de *Machine Learning*, modelos preditivos utilizados na educação, e a utilização do algoritmo CatBoost. Além disso, são discutidas as fontes de dados, engenharia de características, métricas de avaliação, técnicas de balanceamento e otimização de modelos, bem como os testes estatísticos utilizados para comparação de distribuições. O Capítulo 3 descreve os procedimentos metodológicos adotados para o desenvolvimento do estudo, incluindo o pré-processamento dos dados, as estratégias de otimização, a construção e avaliação dos modelos preditivos, além de considerações sobre a reprodutibilidade do experimento. O Capítulo 4 apresenta a plataforma LANSE, onde os modelos foram aplicados como apoio à tomada de decisão educacional. Os resultados obtidos, bem como a comparação com modelos existentes e a análise de desempenho preditivo, são apresentados no Capítulo 5. No Capítulo 6, realiza-se uma discussão crítica dos achados à luz da literatura, retomando as questões de pesquisa. Por fim, o Capítulo 7 apresenta as considerações finais, sugestões de trabalhos futuros e informações sobre a publicação derivada deste estudo.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo apresenta a fundamentação teórica necessária para compreender a previsão de evasão estudantil e o baixo desempenho acadêmico por meio de modelos de aprendizado de máquina. São discutidos os principais avanços na área, abordando modelos preditivos, fontes de dados relevantes e os desafios enfrentados na construção de sistemas eficazes para identificar alunos em risco.

2.1 LEARNING ANALYTICS

O campo de Learning Analytics (LA) tem se consolidado como uma abordagem promissora para compreender e melhorar o processo de aprendizagem a partir da análise de dados educacionais. Segundo Ferguson (2012) e Papamitsiou e Economides (2014), o LA pode ser definido como a medição, coleta, análise e apresentação de dados sobre os aprendizes e seus contextos, com o objetivo de compreender e otimizar a aprendizagem e os ambientes nos quais ela ocorre. O campo possui raízes interdisciplinares, combinando elementos da mineração de dados educacionais (EDM), inteligência de negócios, análise da web e ciência da aprendizagem.

Estudos mais recentes, como os de Motz et al. (2022), destacam que embora essa definição enfatize a melhoria da aprendizagem e dos ambientes educacionais, muitos trabalhos da área ainda se concentram predominantemente em aspectos técnicos e metodológicos, sem avaliar diretamente os impactos sobre o processo de aprendizagem. Em contrapartida, Donche (2023) evidenciam uma crescente adoção do LA no ensino superior europeu, especialmente em níveis micro e meso, com aplicações voltadas à personalização do feedback, monitoramento do desempenho estudantil e suporte à tomada de decisão pedagógica.

As técnicas de LA são frequentemente aplicadas à análise de registros provenientes de ambientes virtuais de aprendizagem (AVAs ou LMS), como o Moodle, bem como dados oriundos de MOOCs, registros de interações, tempos de resposta, sequências de ações e até fatores afetivos (PAPAMITSIOU; ECONOMIDES, 2014). Por meio desses dados, é possível identificar padrões de comportamento estudantil e desenvolver modelos preditivos e prescritivos capazes de apoiar decisões educacionais em diferentes níveis. Estudos recentes apontam para a diversificação dessas aplicações, com avanços no uso de algoritmos adaptativos, análise de sentimentos e visualizações interativas Wang, Lai e Huang (2024), Montuori et al. (2023), Santos et al. (2024).

Dentre as técnicas mais empregadas no campo, destacam-se: (i) a modelagem

de comportamento estudantil, por meio da análise quantitativa e qualitativa das interações com o ambiente de aprendizagem (PAPAMITSIOU; ECONOMIDES, 2014); (ii) a aplicação de algoritmos de mineração de dados e aprendizado de máquina para melhorar a personalização e a avaliação da aprendizagem (FERGUSON, 2012); (iii) a análise de redes sociais (SNA), voltada à identificação de relações entre estudantes, professores e conteúdos (FERGUSON, 2012); (iv) a análise de logs e sequências de ações, com o objetivo de inferir estratégias de aprendizagem (PAPAMITSIOU; ECONOMIDES, 2014); (v) o uso de algoritmos de agrupamento e classificação, como redes neurais, para segmentar estudantes com base em seu desempenho (PAPAMITSIOU; ECONOMIDES, 2014); e (vi) a análise de texto em fóruns e atividades discursivas, com foco em aspectos como gênero discursivo e engajamento (PAPAMITSIOU; ECONOMIDES, 2014).

A aplicação dessas técnicas permite o suporte à tomada de decisão educacional em múltiplos níveis. Em nível instrucional, o LA pode ser utilizado para a identificação de alunos em risco, especialmente nos estágios iniciais do curso, viabilizando intervenções precoces (LONG; SIEMENS, 2011; PAPAMITSIOU; ECONOMIDES, 2014; GAŠEVIĆ; DAWSON; SIEMENS, 2015). Ferramentas como o *Course Signals*, por exemplo, foram desenvolvidas com esse intuito, sinalizando o risco de fracasso acadêmico para cada estudante (GAŠEVIĆ; DAWSON; SIEMENS, 2015). Além disso, o LA viabiliza a melhoria do feedback e da instrução, ao identificar dificuldades específicas dos estudantes, permitindo feedback personalizado e direcionado (GAŠEVIĆ; DAWSON; SIEMENS, 2015; PAPAMITSIOU; ECONOMIDES, 2014).

Outro impacto importante está na personalização da aprendizagem, por meio de sistemas adaptativos que ajustam recursos, conteúdos e sequências pedagógicas de acordo com o perfil e progresso do estudante (LONG; SIEMENS, 2011; PAPAMITSIOU; ECONOMIDES, 2014). Visualizações e dashboards, por sua vez, contribuem para o aumento da autorreflexão e autoconsciência dos alunos, ao comparar seu desempenho com o de seus pares (PAPAMITSIOU; ECONOMIDES, 2014; MONTUORI et al., 2023). Do ponto de vista institucional, o LA também pode embasar decisões administrativas, como a alocação de recursos, avaliação de programas e políticas educacionais (LONG; SIEMENS, 2011; SANTOS et al., 2024).

É importante destacar que a eficácia do uso de LA está diretamente relacionada à fundamentação pedagógica das ferramentas desenvolvidas (GAŠEVIĆ; DAWSON; SIEMENS, 2015). O julgamento humano permanece essencial na interpretação dos resultados e na tomada de decisões subsequentes, reforçando a necessidade de uma abordagem crítica e informada na utilização desses dados (PAPAMITSIOU; ECONOMIDES, 2014; SANTOS et al., 2024).

Assim, o Learning Analytics representa uma intersecção entre dados, tec-

nologia e pedagogia, oferecendo caminhos promissores para o aperfeiçoamento do processo de ensino-aprendizagem, especialmente em contextos educacionais mediados por tecnologia Wang, Lai e Huang (2024), Santos et al. (2024).

2.2 MOODLE

O Moodle (*Modular Object-Oriented Dynamic Learning Environment*) é uma plataforma de código aberto amplamente utilizada em instituições de ensino superior ao redor do mundo. Desenvolvido por Martin Dougiamas com base em princípios de construtivismo social, o Moodle busca promover ambientes colaborativos de aprendizagem (DOUGIAMAS; TAYLOR, 2003). Sua arquitetura modular e flexível permite a criação de cursos personalizados e a coleta detalhada de dados sobre as interações dos estudantes com o ambiente virtual, como acessos, participações em fóruns, envios de atividades e navegação entre recursos. Por essas características, o Moodle tornou-se uma das principais fontes de dados para pesquisas em Learning Analytics (PAPAMITSIOU; ECONOMIDES, 2014; GAŠEVIĆ; DAWSON; SIEMENS, 2015), sendo frequentemente utilizado em estudos sobre engajamento estudantil, personalização da aprendizagem e predição de evasão. A documentação oficial do projeto destaca sua ampla adoção global e a robustez de sua comunidade de desenvolvedores e educadores (Moodle Project, 2024).

Diversos estudos recentes têm explorado o potencial dos dados gerados pelo Moodle para prever a evasão estudantil no ensino superior. Poellhuber et al. (2023) propõem uma abordagem baseada em clusters de cursos, utilizando aprendizado de máquina para treinar modelos preditivos específicos para cada grupo de disciplinas, demonstrando a viabilidade de modelos em larga escala. Já Santos e Henriques (2023) adotam uma estratégia de análise agnóstica ao curso, com modelos treinados sobre múltiplas disciplinas, destacando o potencial dos logs do Moodle como preditores robustos de desempenho e risco. Em um estudo voltado ao contexto finlandês, Vaarma e Li (2024) compararam o poder preditivo dos dados do Moodle com históricos acadêmicos e dados demográficos, evidenciando que a “contagem de atividades no Moodle” foi uma das variáveis mais importantes para prever evasão ao longo do curso.

No contexto africano, Bouihi et al. (2024) exploram a regressão regularizada (L1 e L2) sobre dados acadêmicos extraídos de cursos no Moodle, atingindo acurácias elevadas mesmo com bases de dados pequenas, reforçando o papel do Moodle em contextos com menos infraestrutura. Em nível nacional, Cechinel et al. (2023) descrevem a arquitetura do sistema LANSE, que integra dados de interação com o Moodle em uma plataforma de análise preditiva em nuvem, com visualizações em dashboards para apoiar a tomada de decisão por professores e gestores. Esses trabalhos confirmam a relevância do Moodle não apenas como ambiente de ensino,

mas também como fonte confiável para a construção de modelos de predição de evasão baseados em Learning Analytics.

2.3 EVASÃO ESTUDANTIL E PREVISÃO DE BAIXO DESEMPENHO ACADÊMICO

A evasão estudantil e o baixo desempenho acadêmico representam desafios persistentes nas instituições de ensino superior. A identificação precoce de alunos em risco de evasão permite a formulação de políticas e intervenções eficazes para mitigar esses problemas.

Diversos estudos investigaram os principais fatores que impactam a retenção estudantil. O artigo de Doyaoen et al. (2024) analisou as taxas de retenção de estudantes em cursos de licenciatura e identificou que dificuldades acadêmicas, como notas baixas e reprovações, são os principais contribuintes para a evasão. Além disso, barreiras financeiras e fatores pessoais desempenham um papel significativo na decisão de interrupção dos estudos. O estudo destaca que cursos como Licenciatura em Matemática (BSED-Math) e Licenciatura em Educação Primária (BEED) apresentam as menores taxas de retenção.

De acordo com Kemda e Murray (2021), a reprovação em disciplinas do primeiro semestre é um preditor relevante de atraso na graduação. Além disso, dificuldades financeiras levam muitos alunos a trabalhar, reduzindo sua dedicação aos estudos e impactando negativamente seu desempenho acadêmico. A análise longitudinal realizada pelo mesmo estudo mostrou que há uma diferença significativa na média de reprovações entre alunos que concluíram seus cursos e aqueles que abandonaram. A presença de assistência financeira e políticas institucionais de apoio ao estudante também foram identificadas como fatores de influência na taxa de retenção.

Além dos desafios acadêmicos, fatores socioeconômicos desempenham um papel crucial na evasão estudantil. Segundo Nikolaidis et al. (2022), a intenção do aluno de abandonar a universidade está fortemente associada à sua percepção do progresso de aprendizagem. Fatores como interação com colegas e professores, eficácia do ensino, desempenho em exames e absenteísmo influenciam diretamente o risco de evasão. Da mesma forma, o estudo de Obaido et al. (2023) aponta que a evasão na área de Ciência da Computação está frequentemente relacionada à falta de habilidades de gerenciamento de projetos, dificuldades com o conteúdo do curso e ausência de feedback adequado dos professores.

Diante desses desafios, estratégias de intervenção têm sido amplamente estudadas. Nikolaidis et al. (2022) propõem um modelo baseado na Teoria da Integração de Tinto e no Modelo de Atrito de Bean, sugerindo que o GPA (média das notas) pode atuar como um mediador no risco de evasão, especialmente para alunos do primeiro

ano.

A evasão estudantil e o baixo desempenho acadêmico são fenômenos complexos influenciados por múltiplos fatores interconectados, incluindo dificuldades acadêmicas, financeiras, pessoais e programáticas. A literatura sugere que a implementação de intervenções precoces e sistemas de apoio abrangentes são essenciais para mitigar esses problemas. Além disso, compreender a percepção dos alunos sobre seu próprio progresso acadêmico pode fornecer informações valiosas para o desenvolvimento de estratégias mais eficazes de retenção estudantil.

Para enfrentar esses desafios de maneira mais eficiente, técnicas baseadas em ML têm sido cada vez mais empregadas na previsão de evasão estudantil. Essas abordagens analisam grandes volumes de dados educacionais e identificam padrões que podem indicar risco de abandono, permitindo que instituições desenvolvam estratégias preventivas. A próxima seção explora os principais modelos de aprendizado de máquina utilizados para prever a evasão e o desempenho acadêmico dos alunos.

2.4 FUNDAMENTOS DE MACHINE LEARNING

O aprendizado de máquina (*Machine Learning* – ML) é um campo multidisciplinar da inteligência artificial que se dedica ao desenvolvimento de algoritmos capazes de aprender padrões a partir de dados e melhorar seu desempenho com a experiência. Segundo Mitchell (2013), um programa de computador aprende com uma experiência E em relação a uma tarefa T e uma medida de desempenho P se seu desempenho na tarefa T , conforme medido por P , melhora com a experiência E . Essa definição ressalta a característica central do ML: a adaptação baseada em dados empíricos.

Para James et al. (2017), o aprendizado estatístico — que engloba o ML — consiste em um conjunto de ferramentas que combinam estatística e ciência da computação, voltadas à modelagem e compreensão de dados complexos. Os autores defendem que é essencial compreender os modelos, suas suposições e *trade-offs*, evitando tratá-los como caixas-pretas. Essa perspectiva também é enfatizada por Mitchell (2013), ao destacar a influência de campos como psicologia, neurobiologia e estatística no desenvolvimento de algoritmos de aprendizado, incluindo conceitos como viés e variância para entender erros de generalização.

Um estudo clássico conduzido por Caruana e Niculescu-Mizil (2006) comparou empiricamente diversos algoritmos supervisionados, reforçando a importância da avaliação empírica das hipóteses no desempenho preditivo. Os autores também apontam que esses algoritmos são úteis na predição de probabilidades calibradas, especialmente em domínios onde decisões dependem do grau de confiança, como na previsão de evasão escolar.

Os paradigmas centrais do aprendizado de máquina incluem o aprendizado supervisionado, não supervisionado, semissupervisionado e por reforço. No contexto educacional, o mais utilizado é o aprendizado supervisionado, no qual o modelo aprende uma função de mapeamento entre entradas (como características dos estudantes) e saídas (como evasão ou permanência) a partir de exemplos rotulados (KOTSIANTIS, 2007). Essa abordagem permite construir classificadores que associam perfis de comportamento a categorias de risco.

Além do paradigma supervisionado, há princípios teóricos que fundamentam o aprendizado de máquina:

- **Avaliação empírica de hipóteses:** A verificação experimental da precisão dos modelos é essencial para o desenvolvimento de sistemas confiáveis (JAMES et al., 2017; MITCHELL, 2013; CARUANA; NICULESCU-MIZIL, 2006).
- **Aprendizado Bayesiano:** Adota uma abordagem probabilística para representar incertezas e atualizar crenças com base em evidências, permitindo modelos interpretáveis e dinâmicos (MITCHELL, 2013; JAMES et al., 2017).
- **Aprendizado baseado em instâncias:** Algoritmos como o *k-Nearest Neighbors* generalizam localmente e são adequados para dados com distribuição não linear ou com baixa densidade (KOTSIANTIS, 2007; MITCHELL, 2013).
- **Algoritmos genéticos:** Inspiram-se na evolução biológica, utilizando mecanismos como seleção e recombinação para resolver problemas de busca e otimização em espaços complexos (MITCHELL, 2013).
- **Aprendizado analítico:** Integra conhecimento prévio e raciocínio dedutivo ao processo de indução, promovendo uma combinação entre dados empíricos e regras explícitas (MITCHELL, 2013).

Kotsiantis (2007) também propõem a distinção entre algoritmos indutivos e métodos *lazy*, que adiam a generalização até o momento da predição — o que pode ser útil em contextos como o educacional, onde decisões em tempo real são desejáveis.

Em síntese, o aprendizado de máquina se baseia em fundamentos sólidos de estatística, inferência e lógica, com aplicações amplas e crescentes na análise de dados. Esses princípios sustentam o desenvolvimento de modelos robustos e adaptativos, capazes de apoiar intervenções mais eficazes em contextos educacionais, como a previsão de evasão escolar e o apoio à tomada de decisão pedagógica.

2.5 MACHINE LEARNING NA EDUCAÇÃO: APLICAÇÕES E DESAFIOS

O uso de aprendizado de máquina (ML) no contexto educacional tem ganhado destaque nas últimas décadas, sobretudo pela sua capacidade de identificar padrões ocultos nos dados dos estudantes e apoiar intervenções pedagógicas mais eficazes. Em particular, o ML tem se mostrado promissor na prevenção da evasão escolar, ao permitir a identificação precoce de alunos em risco com base em dados históricos e comportamentais (IDRISSI, 2023).

Entre as abordagens mais comuns na educação, destaca-se o *aprendizado supervisionado*, que utiliza dados rotulados para treinar modelos capazes de prever eventos futuros, como o abandono escolar. Essa abordagem é especialmente adequada quando há registros históricos suficientes sobre o desempenho e o engajamento dos estudantes (NAGY; MOLONTAY, 2024).

Modelos supervisionados podem ser empregados para classificar estudantes em categorias de risco, como "propensos a desistir" ou "com alta probabilidade de conclusão", viabilizando sistemas de alerta precoce que auxiliam educadores e gestores na tomada de decisão (VILLAR; ANDRADE, 2024). Esses sistemas tornam possível a implementação de intervenções personalizadas, baseadas em evidências, com foco na retenção estudantil.

Dentre os algoritmos mais utilizados para essa finalidade, destacam-se os modelos baseados em árvores de decisão, especialmente os do tipo *Gradient Boosting*, como o XGBoost, LightGBM e CatBoost. Esses modelos apresentam desempenho robusto e são particularmente eficazes em conjuntos de dados educacionais com variáveis categóricas, como curso, modalidade e frequência de acesso ao ambiente virtual de aprendizagem (NGUYEN et al., 2024).

Apesar de seus benefícios, a aplicação do ML na educação também enfrenta desafios. Um dos principais é a interpretação dos modelos, especialmente quando são utilizados algoritmos complexos que operam como "caixas-pretas". A compreensão dos fatores que influenciam as previsões é fundamental para garantir a confiança dos educadores nas recomendações geradas pelos sistemas. Outro desafio importante diz respeito à qualidade e à disponibilidade dos dados, que nem sempre estão organizados ou atualizados de forma adequada nas instituições de ensino.

Por fim, a integração dos sistemas preditivos com os fluxos institucionais existentes exige um trabalho interdisciplinar que combine conhecimento técnico, pedagógico e administrativo. Superar esses desafios é essencial para que o aprendizado de máquina possa cumprir seu potencial como ferramenta estratégica na promoção do sucesso acadêmico e na redução das taxas de evasão.

2.6 MODELOS DE MACHINE LEARNING PARA PREVISÃO DE EVASÃO ESCOLAR

Diversos modelos de aprendizado de máquina têm sido amplamente utilizados para prever a evasão estudantil e o desempenho acadêmico, viabilizando a identificação precoce de alunos em risco e a implementação de intervenções proativas (JAIN; SOLANKI, 2019; MACARINI et al., 2019; QUEIROGA et al., 2020; SULAK; KOKLUB, 2024). Esses modelos analisam grandes volumes de dados provenientes de diferentes fontes — como registros acadêmicos, interações em ambientes virtuais de aprendizagem (AVA) e informações socioeconômicas — para detectar padrões associados ao abandono escolar.

Modelos estatísticos tradicionais, como a regressão logística, árvores de decisão e máquinas de vetores de suporte (SVM), têm sido frequentemente empregados nesse contexto. Entretanto, métodos mais avançados, como os modelos de conjunto (*ensemble models*), vêm ganhando destaque na literatura recente. Técnicas como Random Forest, XGBoost (*eXtreme Gradient Boosting*), LightGBM (*Light Gradient Boosting Machine*) e CatBoost (*Categorical Boosting*) demonstram excelente desempenho ao combinar múltiplos classificadores para reduzir viés e variância, alcançando maior generalização. Estudos como os de Nguyen et al. (2024) e Kumar et al. (2025) indicam que LightGBM e CatBoost superaram outros algoritmos em termos de precisão na predição da evasão.

Além dos modelos baseados em árvores, arquiteturas de redes neurais artificiais também têm sido investigadas. Por exemplo, Sulak e Koklub (2024) apresentou uma abordagem baseada em redes neurais que alcançou F1-score de 0,76. No entanto, estudos como os de Ndunagu et al. (2024), que utilizaram redes neurais recorrentes (LSTM) e convolucionais (CNN), apontam que a eficácia desses modelos depende da estrutura dos dados e da complexidade das relações entre as variáveis, nem sempre superando métodos mais tradicionais.

Quanto às variáveis mais relevantes para o treinamento dos modelos, destacam-se características demográficas e socioeconômicas (como idade, ocupação dos responsáveis, situação de bolsas), dados acadêmicos (notas, histórico escolar, créditos acumulados) e, principalmente, informações sobre o engajamento no AVA (acessos, participação em fóruns, submissão de tarefas). Trabalhos como os de Dewi et al. (2023) e Alghamdi, Soh e Li (2025) reforçam que logs de interação em plataformas como o Moodle são preditores especialmente eficazes.

Apesar dos avanços, um dos principais desafios ainda é a interpretabilidade dos modelos mais complexos. Ferramentas como o LIME (*Local Interpretable Model-Agnostic Explanations*) vêm sendo utilizadas para explicar decisões individuais de

modelos preditivos, tornando suas previsões mais compreensíveis e confiáveis para os educadores (NGUYEN et al., 2024).

Outro ponto crítico diz respeito à necessidade de previsões em tempo real, de modo que instituições educacionais possam reagir de forma rápida e personalizada. Nesse sentido, Ohkawauchi e Tanaka (2024) investigaram métodos capazes de processar logs em tempo real, mostrando que a análise contínua das interações dos estudantes pode elevar a eficácia dos sistemas de alerta.

Em suma, a escolha do modelo preditivo mais adequado dependerá de fatores como a disponibilidade e qualidade dos dados, o equilíbrio entre acurácia e interpretabilidade e a capacidade do modelo de generalizar entre diferentes perfis e contextos educacionais. Entre os modelos mais promissores, destaca-se o CatBoost, não apenas por seu desempenho superior com variáveis categóricas, mas também por sua robustez com dados numéricos, boa capacidade de generalização e mecanismos de regularização incorporados. Essas características justificam uma análise mais aprofundada do algoritmo na seção seguinte.

2.7 O CatBoost e Sua Aplicação na Predição de Reprovação

O CatBoost é um algoritmo de aprendizado de máquina desenvolvido pela Yandex, baseado em *Gradient Boosting*, que se destaca por sua eficiência na modelagem de dados tabulares. Embora seja amplamente reconhecido por sua capacidade nativa de lidar com variáveis categóricas, o CatBoost também apresenta desempenho competitivo em conjuntos de dados puramente numéricos — como é o caso deste estudo. Sua arquitetura baseada em árvores simétricas, aliada a técnicas avançadas de regularização e robustez contra overfitting, torna-o uma escolha apropriada para tarefas de classificação em contextos com dados desbalanceados e com variáveis derivadas de logs educacionais, como frequência de acesso, semanas no curso e métricas temporais acumuladas (PROKHORENKOVA et al., 2019).

Em estudos recentes, o CatBoost demonstrou desempenho superior na previsão de evasão e baixo desempenho acadêmico, mesmo quando comparado a algoritmos como Regressão Logística, Random Forest e LightGBM (NGUYEN et al., 2024). Por exemplo, em uma análise preditiva, o CatBoost obteve as maiores pontuações de AUC (0.774) e precisão média (0.847), superando métodos tradicionais e demonstrando sua eficácia na identificação de alunos em risco. Além disso, sua capacidade de lidar com dados desbalanceados contribui para sua precisão preditiva, especialmente em cenários em que há um número significativamente menor de alunos reprovados em relação aos aprovados (NAGY; MOLONTAY, 2024).

2.8 FONTES DE DADOS E ENGENHARIA DE CARACTERÍSTICAS

A eficácia dos modelos preditivos depende diretamente da qualidade e variedade dos dados disponíveis. Diferentes estudos exploraram fontes como registros acadêmicos, interações no LMS e fatores socioeconômicos e psicológicos para a previsão de evasão (OQAIDI; AOUHASSI; MANSOURI, 2022; AGUIAR et al., 2018).

Pesquisas indicam que os dados de LMS são altamente relevantes para prever o desempenho acadêmico. Por exemplo, Dewi et al. (2023) mostrou que características como frequência de login, participação em fóruns e realização de avaliações contribuem significativamente para melhorar a precisão dos modelos. Além disso, uma abordagem baseada em agrupamento foi utilizada para analisar padrões de evasão estudantil Poellhuber et al. (2023), identificando perfis de comportamento dentro do LMS que permitem intervenções mais direcionadas.

Uma revisão abrangente das variáveis de logs utilizadas em modelos preditivos foi conduzida por Wang e Mousavi (2022), demonstrando que a combinação de diferentes métricas de engajamento pode resultar em modelos mais robustos. No presente estudo, características como dias de acesso, frequência de logs e tipos de atividade foram incorporadas como variáveis-chave para a modelagem preditiva.

2.9 MÉTRICAS DE AVALIAÇÃO DE MODELOS PREDITIVOS

A avaliação do desempenho de modelos preditivos é uma etapa essencial em projetos de ciência de dados aplicados à educação, especialmente no contexto da previsão de evasão escolar. As métricas de avaliação orientam tanto a escolha quanto a otimização dos modelos, oferecendo critérios objetivos para comparar abordagens distintas e selecionar aquelas mais eficazes em atingir os objetivos do estudo.

No caso específico da previsão de abandono escolar, o desequilíbrio entre as classes, comumente uma minoria de estudantes que evadem, torna crucial a utilização de métricas que vão além da acurácia tradicional. Métricas como Recall, F1-Score e AUC-ROC fornecem uma visão mais abrangente sobre a capacidade do modelo em identificar corretamente os estudantes em risco, minimizando falsos negativos e maximizando o potencial de intervenção precoce. Nesta seção, são discutidas as principais métricas utilizadas na literatura e neste estudo, com destaque para suas características, aplicações e limitações no contexto educacional.

2.9.1 Acurácia na Previsão de Evasão Escolar

No contexto da análise do desempenho de modelos preditivos para a previsão de evasão escolar, a Acurácia desempenha um papel fundamental como métrica inicial para avaliar a capacidade geral do modelo em classificar corretamente os estudantes. A

acurácia representa a proporção de todas as previsões feitas pelo modelo que estavam corretas, ou seja, o número de estudantes corretamente classificados como desistentes ou não desistentes em relação ao total de estudantes analisados.

Estudos indicam que a acurácia é frequentemente utilizada na avaliação de diferentes modelos de *machine learning* para previsão da evasão escolar. Por exemplo, um modelo baseado em árvores de decisão apresentou acurácia significativa na previsão de abandono, matrícula e graduação (ZHOU; AGARWAL, 2024). Além disso, a comparação entre diversos algoritmos mostrou que *Random Forest* frequentemente obteve os melhores desempenhos em termos de acurácia (BOUIHI et al., 2024; DEB et al., 2024). Outra abordagem analisou técnicas de *ensemble*, como *stacking* e *blending*, na previsão de evasão em MOOCs, reportando acurácia superior a 82% (PUTRA; UTAMI, 2024). Modelos baseados em regressão logística, tanto com quanto sem regularização, também demonstraram desempenho competitivo, com acurácias de teste em torno de 81% (BOUIHI et al., 2024).

Apesar de sua ampla utilização, é essencial compreender as limitações da acurácia no contexto específico da previsão de evasão escolar. Embora forneça uma visão geral do desempenho do modelo, a acurácia pode ser enganosa em conjuntos de dados desbalanceados, um cenário comum na previsão de evasão, onde o número de estudantes que permanecem na escola pode ser significativamente maior do que o número de desistentes. Nesses casos, um modelo pode alcançar uma alta acurácia simplesmente prevendo a classe majoritária na maior parte do tempo, sem identificar corretamente os estudantes em risco de evasão.

Por esse motivo, a acurácia é frequentemente reportada em conjunto com outras métricas, como Precisão, Recall e F1-Score, que permitem uma avaliação mais detalhada do desempenho do modelo (DEB et al., 2024; PUTRA; UTAMI, 2024). Essas métricas fornecem informações cruciais sobre a capacidade do modelo de identificar corretamente os casos positivos (estudantes em risco de evasão) sem gerar um excesso de falsos positivos (estudantes previstos como evadidos que, na realidade, continuam matriculados).

Por exemplo, um modelo pode apresentar uma acurácia razoável, mas um baixo *recall* para a classe de evasão, indicando que falha em identificar um grande número de estudantes que realmente irão abandonar os estudos. Alternativamente, pode apresentar uma precisão reduzida, significando que muitas previsões de evasão são, na verdade, falsos alarmes. Um estudo recente demonstrou dificuldades em distinguir estudantes matriculados de graduados, um desafio intensificado pelo desbalanceamento das classes (ZHOU; AGARWAL, 2024). Isso ilustra como uma acurácia aparentemente alta pode mascarar um desempenho insuficiente em classes específicas de interesse, como a de estudantes que efetivamente abandonam os estudos.

Dessa forma, a Acurácia é um ponto de partida relevante na avaliação do desempenho de modelos preditivos para evasão escolar, mas não deve ser considerada isoladamente. Para uma compreensão mais completa da eficácia do modelo na identificação de estudantes em risco, é essencial analisá-la em conjunto com métricas que avaliem o impacto de falsos positivos e falsos negativos, especialmente em cenários nos quais os custos desses erros podem ter implicações distintas. Estudos recentes enfatizam a importância de considerar outras métricas, como Precisão, Recall e F1-Score, para uma avaliação mais robusta e informativa (DEB et al., 2024; PUTRA; UTAMI, 2024).

2.9.2 Recall na Previsão de Evasão Escolar

O Recall, também conhecido como sensibilidade, é uma métrica de avaliação fundamental na análise de desempenho de modelos preditivos, especialmente em cenários com desbalanceamento de classes. Matematicamente, o Recall é definido como a razão entre o número de verdadeiros positivos (TP) e a soma entre verdadeiros positivos e falsos negativos (FN) (KETUT et al., 2024):

$$\text{Recall} = \frac{TP}{TP + FN}$$

No contexto da previsão de evasão escolar, a classe positiva de interesse refere-se aos estudantes que efetivamente abandonam o curso. Assim, um valor elevado de Recall indica que o modelo é eficaz em identificar a maioria dos estudantes em risco, minimizando o número de falsos negativos — isto é, casos em que o estudante irá abandonar, mas não foi identificado pelo modelo como tal (KETUT et al., 2024).

Ketut et al. (2024) enfatizam a prioridade do Recall nesse contexto, argumentando que é preferível prever que um estudante abandonará o curso, mesmo que ele não o faça (falso positivo), do que deixar de identificar um estudante que efetivamente irá desistir (falso negativo). Isso se justifica pela possibilidade de intervenção precoce, ainda que desnecessária, em vez da perda de uma oportunidade crítica de suporte a um estudante que abandonará os estudos sem receber ajuda.

A importância do Recall para problemas de evasão escolar pode ser sintetizada em três aspectos principais:

- **Minimização de Falsos Negativos:** O principal objetivo de modelos preditivos nessa área é identificar estudantes em risco para que intervenções sejam realizadas. Um baixo Recall implica em elevado número de falsos negativos, o que significa que muitos estudantes que efetivamente abandonariam o curso não são detectados (KETUT et al., 2024; PSATHAS; CHATZIDAKI; DEMETRIADIS, 2023; ZHOU; AGARWAL, 2024).

- **Priorização da Cobertura:** Em problemas educacionais, é geralmente mais relevante capturar a maior quantidade possível de casos reais de evasão do que obter precisão absoluta em todas as previsões. Um Recall elevado assegura que a maioria dos estudantes em risco seja sinalizada para possível intervenção (KETUT et al., 2024; PSATHAS; CHATZIDAKI; DEMETRIADIS, 2023).
- **Impacto nas Intervenções:** Modelos com alto Recall são mais úteis para as instituições educacionais, pois aumentam a eficácia das estratégias de apoio ao permitir que recursos sejam direcionados aos estudantes que mais precisam (ZHOU; AGARWAL, 2024; KETUT et al., 2024).

Tanto Ketut et al. (2024) quanto Psathas, Chatzidaki e Demetriadis (2023) destacam o Recall como métrica central em seus estudos, especialmente devido à natureza desbalanceada dos dados — onde o número de estudantes que permanecem na instituição costuma ser muito superior ao número de evasores. Nessas situações, métricas como acurácia podem ser enganosas, já que um modelo pode alcançar alto desempenho apenas prevendo a classe majoritária (KETUT et al., 2024; PSATHAS; CHATZIDAKI; DEMETRIADIS, 2023).

No estudo de Ketut et al. (2024), a escolha do melhor modelo preditivo (k-Nearest Neighbors com SMOTE) foi orientada pelo maior valor de Recall alcançado (93,5%), indicando uma clara priorização da identificação de potenciais evasores. Da mesma forma, Psathas, Chatzidaki e Demetriadis (2023) também utilizaram o Recall como métrica principal ao avaliar modelos de previsão de evasão em cursos massivos abertos online (MOOCs), destacando a importância de reduzir falsos negativos.

Em síntese, o Recall é uma métrica essencial na avaliação de modelos de previsão de evasão escolar, pois foca na correta identificação dos estudantes em risco. Sua aplicação prática permite que instituições educacionais realizem intervenções direcionadas e oportunas, contribuindo para a redução das taxas de abandono e para a melhoria dos índices de sucesso acadêmico.

2.9.3 F1-Score na Previsão de Evasão Escolar

O F1-Score desempenha um papel fundamental na avaliação do desempenho de modelos preditivos em cenários com desequilíbrio de classes, como é o caso da previsão de evasão escolar. Essa métrica é especialmente útil quando a distribuição entre estudantes que evadem e os que permanecem no curso é desbalanceada, tornando métricas como a acurácia insuficientes para avaliar a real eficácia do modelo.

O F1-Score é definido como a média harmônica entre a precisão (precision) e o recall (revocação), proporcionando uma medida única que considera simultaneamente

a capacidade do modelo de evitar falsos positivos e falsos negativos. A fórmula para o cálculo é:

$$F1 = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

Neste contexto:

- A Precisão refere-se à proporção de estudantes previstos como evasores que realmente abandonam o curso.
- O Recall mede a proporção de estudantes que efetivamente evadem e que foram corretamente identificados pelo modelo.

Ao integrar ambas as métricas, o F1-Score permite uma avaliação mais robusta, especialmente na identificação da classe minoritária — os estudantes em risco de evasão. Em cenários onde a evasão é menos frequente, modelos podem alcançar alta acurácia apenas ao prever a classe majoritária, mascarando sua incapacidade de detectar estudantes em risco. O F1-Score mitiga esse problema ao penalizar desbalançamentos entre precisão e recall.

Diversos estudos utilizaram o F1-Score como métrica central para avaliar e comparar modelos de previsão de evasão escolar. Boteju, Tang e Brown (2024) destacam que a escolha do método de normalização impacta significativamente o F1-Score de modelos SVM, observando um valor médio de 0,779 com a técnica *One Hot Scaler*. Os autores evidenciaram uma diferença de até 0,27 no F1-Score entre modelos com e sem normalização de dados.

Cho, Yu e Kim (2023) analisaram o impacto de diferentes algoritmos de aprendizado e técnicas de oversampling (SMOTE, ADASYN e Borderline-SMOTE) sobre o F1-Score, encontrando que o LightGBM apresentou o melhor desempenho com um F1-Score de 0,840. De forma similar, Deb et al. (2024) identificaram o algoritmo *Random Forest* como o mais eficaz, com um F1-Score de 0,78.

No contexto de cursos online massivos (MOOCs), Putra e Utami (2024) utilizaram o F1-Score para comparar técnicas de *ensemble*, como *stacking* e *blending*, observando que o *blending* proporcionou os melhores resultados. Já Osemwegie, Amadin e Uduehi (2023) priorizaram o F1-Score e o recall para escolher o modelo de Regressão Logística para implantação prática, reforçando a importância dessas métricas em ambientes reais.

Outros trabalhos também relataram resultados significativos. Ridwan e Priyatno (2024) alcançaram um F1-Score máximo de 0,81 com o algoritmo XGBoost, utilizando uma divisão de dados de 80:20, refletindo um bom equilíbrio entre precisão e recall. Por

fim, Wijayaningrum, Kirana e Putri (2024) propuseram um *framework* com seleção de variáveis via *Random Forest* e tratamento de dados desbalanceados com SMOTE, obtendo um F1-Score de 0,9032 com um modelo MLP, superando abordagens anteriores e evidenciando a eficácia da engenharia de características aliada ao balanceamento de classes.

Em síntese, o F1-Score é uma métrica essencial para a avaliação de modelos de previsão de evasão escolar. Ao considerar simultaneamente precisão e recall, o F1-Score proporciona uma medida equilibrada da performance do modelo, especialmente em contextos com classes desbalanceadas. A análise do F1-Score permite comparar a eficácia de diferentes abordagens de modelagem e otimização, contribuindo para a identificação mais precisa dos estudantes em risco e para o aprimoramento de estratégias preventivas nas instituições de ensino.

2.9.4 AUC-ROC na Previsão de Evasão Escolar

O AUC-ROC (Área Sob a Curva da Característica de Operação do Receptor) é uma métrica amplamente utilizada para avaliar o desempenho de modelos de classificação binária, como aqueles aplicados à predição de evasão escolar. Essa métrica representa a capacidade do modelo de distinguir entre duas classes — por exemplo, estudantes que irão abandonar o curso versus estudantes que o concluirão — ao longo de diferentes limiares de decisão (CHEN et al., 2024).

O AUC-ROC varia de 0 a 1, onde valores mais próximos de 1 indicam maior capacidade discriminatória do modelo. Em outras palavras, um AUC-ROC elevado reflete um modelo capaz de diferenciar eficazmente entre estudantes em risco e aqueles não em risco.

Diversos estudos sobre predição de evasão escolar utilizaram o AUC-ROC como métrica de avaliação. Kim, Yoo e Kim (2023) reportaram um AUC-ROC médio de 0,935 ao aplicar quatro classificadores binários para prever se os estudantes abandonariam ou se graduariam, sendo o algoritmo *Random Forest* o mais eficaz, com AUC-ROC de 0,955. De forma semelhante, Sabbir et al. (2024) avaliaram diversos classificadores e obtiveram AUCs de 0,99 para *Random Forest*, 0,98 para *XGBoost*, 0,97 para *k-NN* e 0,94 para *Decision Tree*, indicando uma forte capacidade de discriminação entre as classes.

No entanto, apesar de sua popularidade, o AUC-ROC apresenta limitações importantes, especialmente em cenários com desequilíbrio de classes, algo comum em problemas de evasão escolar, nos quais o número de estudantes que efetivamente abandonam tende a ser menor do que o número de estudantes que permanecem. Chen et al. (2024) argumentam que o AUC-ROC pode ser uma métrica inadequada nesses casos, pois tende a ser insensível ao reconhecimento precoce — uma característica

essencial em aplicações onde intervenções antecipadas são desejáveis. Os autores também fazem referência ao estudo de Truchon e Bayly (2007), que reforça esse ponto ao destacar que o AUC-ROC pode ser enganoso em problemas onde a identificação precoce da classe positiva (evasores) é crítica.

Além disso, Chen et al. (2024) sugerem que, em situações com classes desbalanceadas, métricas alternativas como o AUC-PR (Área sob a Curva de Precisão-Revocação) são mais apropriadas. A pesquisa também identifica inconsistências entre as classificações de algoritmos baseadas em AUC-ROC e AUC-PR, apontando que confiar exclusivamente no AUC-ROC pode não refletir com precisão o desempenho de modelos voltados à identificação de estudantes em risco.

É relevante notar que nem todos os estudos se baseiam no AUC-ROC como principal métrica. Sulak e Koklu (2025), por exemplo, avaliaram modelos de previsão de evasão com base em métricas derivadas da matriz de confusão, como acurácia, precisão, recall e F1-score, sem mencionar o uso de AUC-ROC, o que reforça a necessidade de considerar múltiplas perspectivas na avaliação dos modelos.

Em síntese, embora o AUC-ROC seja uma métrica amplamente reportada e útil para avaliar a capacidade geral de discriminação de um modelo de predição de evasão escolar, é essencial reconhecer suas limitações. Em contextos com dados desbalanceados e quando a detecção precoce é prioritária, outras métricas — como o AUC-PR, o F1-score e o recall — podem oferecer avaliações mais relevantes e alinhadas com os objetivos educacionais. Uma análise abrangente do desempenho dos modelos deve, portanto, incluir múltiplas métricas para garantir uma interpretação robusta e contextualizada da eficácia preditiva.

2.10 TÉCNICAS DE BALANCEAMENTO E OTIMIZAÇÃO DE MODELOS

A previsão de evasão de estudantes universitários é um tema crucial na educação superior, pois permite a identificação precoce de alunos em risco, possibilitando intervenções mais eficazes. No entanto, um dos principais desafios nessa tarefa é o desequilíbrio de classes, que ocorre quando há uma disparidade significativa na quantidade de estudantes que continuam seus estudos em relação àqueles que evadem. Esse desequilíbrio pode comprometer a precisão dos modelos preditivos, pois os algoritmos de aprendizado de máquina tendem a favorecer a classe majoritária — ou seja, os estudantes que permanecem na universidade — em detrimento da classe minoritária, composta por aqueles que abandonam o curso. Como resultado, modelos treinados em dados desbalanceados frequentemente apresentam baixa capacidade de prever a evasão, o que pode dificultar a tomada de decisão por parte das instituições de ensino (CU et al., 2024).

Diante desse desafio, diversas técnicas de balanceamento de classes podem ser aplicadas para mitigar os impactos do desbalanceamento nos modelos de aprendizado. Essas técnicas podem ser classificadas em abordagens no nível dos dados (*data-level*), no nível do algoritmo (*algorithm-level*) ou híbridas. No nível dos dados, o reequilíbrio pode ser feito por meio do aumento do número de exemplos da classe minoritária (*oversampling*) ou pela redução dos exemplos da classe majoritária (*undersampling*). No contexto da previsão de evasão, o *oversampling* tem sido amplamente adotado, pois permite aumentar a representatividade dos estudantes evadidos sem perda de informações dos dados originais (CU et al., 2024; GANDHI; PATEL; JAGTAP, 2024).

Dentre as técnicas mais utilizadas para *oversampling*, destacam-se o Synthetic Minority Over-sampling Technique (SMOTE) e o ADASYN. O SMOTE gera exemplos sintéticos interpolando características de alunos que evadiram e seus vizinhos mais próximos, evitando a simples duplicação de amostras e reduzindo o risco de sobreajuste (*overfitting*). Dey e Pratap (2023) afirmam que essa abordagem tem sido amplamente utilizada na literatura e conta com variantes aprimoradas, como o I_SMOTE (*Improved Synthetic Minority Over-sampling Technique*), que introduz uma nova métrica de similaridade e um mecanismo de seleção por roleta para melhorar a qualidade das amostras sintéticas (CU et al., 2024).

O ADASYN, por sua vez, também gera amostras sintéticas, mas de maneira adaptativa, priorizando a criação de novos exemplos em regiões onde há menor densidade de estudantes evadidos. Enquanto o SMOTE distribui as novas amostras de forma mais uniforme, o ADASYN ajusta a geração conforme a distribuição dos dados, o que pode melhorar a capacidade do modelo de identificar padrões de evasão em cenários mais complexos (GANDHI; PATEL; JAGTAP, 2024). Segundo Husain et al. (2025), essa adaptação pode resultar em previsões mais precisas, especialmente quando combinada com estratégias de remoção de ruídos, como no método SMOTEENN (*SMOTE + Edited Nearest Neighbors*).

Além do balanceamento de classes, a otimização multiobjetivo desempenha um papel essencial na modelagem preditiva da evasão, pois diferentes critérios de desempenho precisam ser equilibrados. No contexto da previsão de evasão, os modelos podem ter que considerar tanto a precisão da predição quanto a precocidade da identificação dos alunos em risco. Um dos algoritmos mais utilizados para otimização multiobjetivo é o NSGA-II, que busca encontrar um conjunto de soluções Pareto-ótimas, permitindo que múltiplos objetivos sejam considerados simultaneamente (HAO et al., 2022).

Embora o NSGA-II não tenha sido amplamente explorado no contexto específico da previsão de evasão, Pan et al. (2024) abordam a necessidade de otimizar

múltiplos objetivos nesse problema, destacando a importância de equilibrar a precisão da predição com a precocidade da detecção de alunos em risco. Para isso, os autores propõem uma abordagem baseada em aprendizado por reforço multiobjetivo (*multi-objective reinforcement learning*, MORL), utilizando um processo de decisão de Markov para otimizar simultaneamente a acurácia e a precocidade das previsões. Essa estratégia visa minimizar tanto os falsos positivos quanto os falsos negativos, garantindo que as intervenções institucionais sejam realizadas no momento mais adequado. O princípio de equilibrar múltiplos critérios de desempenho é fundamental na otimização multiobjetivo e compartilha conceitos aplicados no NSGA-II (HAO et al., 2022).

O NSGA-II é um dos algoritmos mais amplamente utilizados para otimização multiobjetivo e foi diretamente abordado em estudos como os de Hao et al. (2022) e Zheng e Doerr (2024). Hao et al. (2022) apresentam o NSGA-II como um método clássico de ordenação não dominada, enfatizando sua capacidade de manter diversidade na população de soluções e encontrar compromissos ótimos entre diferentes objetivos. No entanto, os autores também apontam limitações do algoritmo, como sua tendência a cair em ótimos locais e sua velocidade de convergência lenta em problemas complexos. Já Zheng e Doerr (2024) realizam uma análise teórica detalhada do NSGA-II, demonstrando que sua eficiência se deteriora conforme o número de objetivos aumenta, especialmente a partir de três objetivos. No contexto da previsão de evasão, onde múltiplos fatores influenciam a tomada de decisão (exemplo: acurácia do modelo, precocidade da predição, taxa de falsos alarmes), é essencial considerar algoritmos capazes de lidar com essas múltiplas dimensões de forma eficiente.

Dessa forma, o equilíbrio entre precisão e precocidade na previsão da evasão estudantil é um problema que pode se beneficiar de técnicas de otimização multiobjetivo, como o NSGA-II. A escolha entre as diferentes estratégias de balanceamento de dados, seja SMOTE, ADASYN ou métodos híbridos, assim como a abordagem de otimização utilizada, dependerá das características do conjunto de dados e dos objetivos específicos da análise.

2.11 OTIMIZAÇÃO MULTIOBJETIVO COM NSGA-II

O NSGA-II emergiu como uma ferramenta poderosa de otimização multiobjetivo em diversas áreas, incluindo mineração de dados educacionais, onde equilibrar objetivos conflitantes como precisão e recall é crucial. Originalmente proposto como um algoritmo de otimização, a força do NSGA-II reside em sua capacidade de identificar um conjunto diversificado de soluções Pareto-ótimas enquanto mantém a eficiência computacional, essencial para otimizar modelos de aprendizado de máquina em grandes conjuntos de dados (DEB et al., 2002).

Neste estudo, o NSGA-II foi empregado para otimizar a seleção de modelos e

hiperparâmetros. Sua capacidade multiobjetivo permitiu equilibrar a precisão preditiva para estudantes em risco de evasão com outros indicadores-chave de desempenho. Diferente da otimização tradicional de objetivo único, o NSGA-II permite a otimização simultânea de múltiplos objetivos conflitantes ao gerar uma fronteira de Pareto de soluções não dominadas, possibilitando uma exploração mais abrangente dos trade-offs entre métricas como recall e F1-score.

O algoritmo NSGA-II integra diversos componentes críticos que aumentam sua robustez e adaptabilidade. Entre esses, destaca-se sua abordagem rápida de ordenação não dominada, que organiza soluções em camadas hierárquicas com base na dominância de Pareto, permitindo a seleção eficiente de soluções de alta qualidade ao longo das gerações. A distância de aglomeração (*crowding distance*) dentro do NSGA-II garante que as soluções sejam distribuídas uniformemente ao longo da fronteira de Pareto, evitando a convergência para regiões subótimas e melhorando a diversidade das soluções (VERMA; PANT; SNASEL, 2021).

O NSGA-II pode incorporar uma estratégia elitista, que mantém as melhores soluções ao longo das gerações, acelerando a convergência e garantindo a qualidade das soluções, sendo particularmente útil em espaços de busca de alta dimensão. Essa abordagem é especialmente vantajosa na otimização de modelos de aprendizado de máquina, onde equilibrar múltiplos objetivos frequentemente leva a um cenário de solução complexo (YUSOFF; NGADIMAN; ZAIN, 2011). A capacidade do algoritmo de lidar com problemas de otimização combinatória de forma eficiente, conforme descrito nas revisões de Yusoff, Ngadiman e Zain (2011) e Verma, Pant e Snasel (2021), reforça sua aplicabilidade neste estudo, onde espaços de busca multidimensionais e ajuste de hiperparâmetros são desafios-chave.

2.12 TESTES ESTATÍSTICOS PARA COMPARAÇÃO DE DISTRIBUIÇÕES

A avaliação da similaridade entre os conjuntos de treinamento e teste é uma etapa essencial no desenvolvimento de modelos de aprendizado de máquina, pois permite verificar se o modelo treinado será capaz de generalizar adequadamente para dados não vistos. Nesse contexto, os testes de Kolmogorov-Smirnov (KS) e de Mann-Whitney U (U-Test) são amplamente utilizados como ferramentas estatísticas não paramétricas para comparar distribuições e avaliar a existência de diferenças significativas entre dois conjuntos de dados.

O teste de Kolmogorov-Smirnov é um teste de bondade de ajuste que quantifica a distância entre duas funções de distribuição acumulada (FDCs), seja entre uma FDC empírica e uma distribuição teórica, ou entre duas FDCs empíricas. Na análise comparativa entre conjuntos de treinamento e teste, o teste KS de duas amostras verifica se ambas foram extraídas da mesma população. A estatística do teste é baseada na maior

diferença absoluta entre as FDCs empíricas dos dois conjuntos. Um valor pequeno da estatística KS, ou um valor- p alto, sugere que não há diferenças significativas entre as distribuições comparadas, indicando uma alta similaridade entre os dados. Segundo Massey (1951), o teste de Kolmogorov-Smirnov é uma ferramenta robusta para avaliação da bondade de ajuste em contextos diversos. Já Goldman e Kaplan (2018) demonstram que esse teste pode ser interpretado como um procedimento de testes múltiplos sobre os quantis das distribuições, permitindo identificar regiões específicas de divergência enquanto se controla a taxa de erro familiar (FWER). De forma similar, Özçomak et al. (2013) destacam o uso do teste KS para verificar se duas amostras independentes provêm da mesma distribuição populacional.

Por sua vez, o teste de Mann-Whitney U — também conhecido como teste de Wilcoxon rank-sum — é utilizado para comparar duas amostras independentes com o objetivo de avaliar se as populações subjacentes diferem em termos de tendência central. Esse teste opera com base nos postos (ranks) das observações combinadas e testa se uma das amostras tende a apresentar valores sistematicamente maiores do que a outra. Um valor- p elevado indica ausência de evidências estatísticas para afirmar que há diferença entre os grupos. O teste foi introduzido por Mann e Whitney (1947) como um método para avaliar a dominância estocástica entre duas variáveis aleatórias, sendo particularmente útil em contextos onde não se pode assumir normalidade das distribuições. De acordo com Özçomak et al. (2013), o teste de Mann-Whitney é apropriado para verificar se duas amostras independentes pertencem à mesma população, sendo uma alternativa ao teste KS com foco na mediana ou distribuição geral dos dados.

A utilização conjunta dos testes KS e Mann-Whitney permite uma avaliação abrangente da similaridade entre conjuntos de dados, fornecendo evidências estatísticas para aceitar ou rejeitar a hipótese de que os conjuntos de treinamento e teste são amostras da mesma distribuição. Resultados significativos nesses testes podem sugerir a ocorrência de *dataset shift*, ou mudança de distribuição, o que comprometeria a capacidade de generalização dos modelos preditivos. Como se tratam de testes não paramétricos, não impõem suposições rígidas sobre a forma das distribuições subjacentes, sendo apropriados para uma ampla variedade de tipos de dados. No estudo de Özçomak et al. (2013), os autores realizaram uma comparação do poder estatístico entre os testes KS e Mann-Whitney sob diferentes condições de curtose e assimetria, concluindo que a escolha entre os dois testes deve considerar as características específicas das distribuições em análise.

2.13 DESAFIOS E LIMITAÇÕES

Um dos principais desafios no desenvolvimento de modelos preditivos para evasão estudantil é o problema do desbalanceamento de classes, onde o número de alunos que abandonam o curso pode ser muito alto ou muito baixo, dependendo do tipo de curso. Esse desbalanceamento pode levar a modelos tendenciosos, que apresentam bom desempenho geral, mas falham na identificação precisa de alunos em risco. Diversas técnicas, como o SMOTE e o ADASYN, foram empregadas para lidar com esse problema (CHO; YU; KIM, 2023; MACARINI et al., 2019).

Neste estudo, utilizamos o ADASYN para reamostragem e o NSGA-II para busca de hiperparâmetros. A escolha da arquitetura e do ADASYN foi baseada no melhor desempenho obtido durante nossa busca por hiperparâmetros.

Embora ADASYN e NSGA sejam eficazes, eles podem não ser sempre as soluções mais adequadas, dependendo das restrições computacionais. No entanto, o NSGA, assim como outros algoritmos genéticos, oferece flexibilidade para se adaptar a essas restrições. Por exemplo, estratégias de mutação podem ser ajustadas para melhorar a exploração, técnicas de *crossover* podem ser otimizadas e o número de gerações e o tamanho da população podem ser reduzidos para se alinhar aos recursos computacionais disponíveis.

O acesso a dados abrangentes sobre os estudantes é frequentemente limitado, resultando em conjuntos de dados pequenos e restritos. Isso representa um desafio significativo para a generalização e o desempenho dos modelos preditivos. Técnicas eficazes de processamento de dados, incluindo engenharia de características e otimização de modelos, são essenciais para mitigar essas limitações. Uma abordagem promissora é o uso de algoritmos genéticos para otimização de busca, incluindo o ajuste de hiperparâmetros. Esses algoritmos têm demonstrado grande potencial na melhoria da precisão e generalização dos modelos, ao explorar sistematicamente o espaço de hiperparâmetros para encontrar as configurações ideais (QUEIROGA et al., 2020). A Tabela 1 resume os estudos apresentados nesta seção.

Embora o conjunto de dados utilizado neste estudo incluía 567 instâncias únicas de aluno-curso em 23 cursos distintos, seu volume limitado impõe restrições significativas à capacidade de generalização do modelo. Apesar da variedade de cursos contemplados, o número relativamente pequeno de instâncias de aluno-curso restringe a capacidade do modelo de capturar padrões mais amplos de engajamento e comportamento de evasão. Essa limitação pode afetar a robustez dos achados, pois um conjunto de dados maior permitiria uma distribuição mais representativa dos comportamentos de aprendizagem em diferentes contextos acadêmicos. Além disso, o conjunto de dados não inclui informações longitudinais ou dados de formatos de

ensino assíncronos, que poderiam oferecer insights mais profundos sobre a evolução do engajamento dos alunos ao longo do tempo. Estudos futuros devem buscar validar o modelo proposto utilizando conjuntos de dados maiores, com maior participação estudantil, permitindo uma avaliação mais abrangente de sua adaptabilidade e eficácia.

Além disso, fatores externos, como status socioeconômico, qualidade do ensino e motivação dos alunos, que são conhecidos por influenciar significativamente as taxas de evasão (KRÜGER; BRITTO; BARDDAL, 2023; KOK et al., 2024), não foram incluídos neste estudo. Nossa escolha de basear a análise exclusivamente em dados de plataformas de ensino online (por exemplo, logs do Moodle) foi intencional, visando maximizar a aplicabilidade da nossa abordagem para instituições com recursos limitados. Essa decisão permite que qualquer instituição com acesso a dados básicos do LMS adote nosso modelo sem a necessidade de coleta adicional de dados. No entanto, reconhecemos que a incorporação dessas variáveis externas, quando disponíveis, poderia refinar ainda mais as previsões do modelo, proporcionando uma compreensão mais abrangente dos fatores que contribuem para a evasão estudantil.

Tabela 1 – Resumo dos Trabalhos Relacionados Fonte: Elaborado pelo autor

Trabalho	Objetivo	Técnica	Algoritmo	Nível Educa- cional	Características Utilizadas	Métricas
(JAIN; SO-LANKI, 2019)	Predizer o desempenho estudantil em múltiplas turmas	Classificação	Árvores de Decisão, Floresta Aleatória, XGBoost	Superior	Notas, frequência, dados demográficos	Acurácia, Precisão, Recall, F-measure: 95
(MACARINI et al., 2019)	Predição de sucesso em ambientes de aprendizado híbridos	Classificação	Floresta Aleatória, Naïve Bayes, Árvores de Decisão	Superior	Engajamento no LMS, desempenho	AUC ROC: 90%
(SU; LIN; LIU, 2022)	Explorar comportamentos de aprendizagem e prever desempenho	Aprendizado de Máquina	Regressão Logística, SVM, Árvores de Decisão, Floresta Aleatória	Superior	Dados demográficos, métricas de desempenho, dados comportamentais	Acurácia: 0.83, F1: 0.88, AUC: 0.66 (RNA)

Continua na próxima página

Resumo dos Artigos Utilizados nos Trabalhos Relacionados (continuação)

Trabalho	Objetivo	Técnica	Algoritmo	Nível Educa- cional	Características Utilizadas	Métricas
(CHO; YU; KIM, 2023)	Predição de evasão baseada em dados de engajamento	Classificação	Floresta Aleatória, Regressão Logística, Árvores de Decisão	Superior	Dados de engajamento, uso do LMS, evasão	Acurácia: 0.955, F1-Score: 0.840, Precisão: 0.86, Recall: 0.81 (LightGBM)
(SARKER, 2021)	Aplicação de aprendizado de máquina em contextos educacionais reais	Classificação	Regressão Logística, Árvores de Decisão, Floresta Aleatória	-	Dados contextuais, PCA, ANOVA, Correlação de Pearson, RFE, Qui-quadrado	-
(DEWI et al., 2023)	Predição precoce de evasão em aprendizagem online	Classificação	Naïve Bayes, Regressão Logística e SVM	Online	Logs de interação no LMS e engajamento	Acurácia: 0.99, Precisão: 0.98, Recall: 0.96, F1: 0.97 (SVM)
(OQAI; AOUHASSI; MANSOURI, 2022)	Propor um modelo de previsão de evasão usando aprendizado de máquina	Classificação	Regressão Logística, Árvores de Decisão, Floresta Aleatória	Superior	Dados acadêmicos, engajamento, dados demográficos	-
(POELLHUBER et al., 2023)	Modelo de predição de evasão baseado em agrupamento no Moodle LMS	Agrupamento, Classificação	Regressão Logística, Floresta Aleatória, Gradient Boosting, XGBoost, AdaBoost	Superior	Engajamento, logs do LMS, design do curso	Acurácia: 0.94, F1-Score: 0.33, Precisão: 0.56, Recall: 0.26, AUC ROC: 0.62

Continua na próxima página

Resumo dos Artigos Utilizados nos Trabalhos Relacionados (continuação)

Trabalho	Objetivo	Técnica	Algoritmo	Nível Educa- cional	Características Utilizadas	Métricas
(WANG; MOUSAVI, 2022)	Análise de variáveis de logs para previsão de desempenho acadêmico	Meta-Análise	Vários	Superior	Variáveis de logs do LMS	Preditores significativos de desempenho acadêmico
(QUEIROGA et al., 2020)	Predição de alunos em risco na educação técnica a distância	Classificação, Otimização de Hiperparâmetros	Árvores de Decisão, Floresta Aleatória, AdaBoost	Secundário	Logs de interação estudantil	AUC ROC: 0.91
(KRÜGER; BRITTO; BARDAL, 2023)	Aprendizado de máquina explicável para predição de evasão	Classificação	Floresta Aleatória, Árvores de Decisão, Regressão Logística, AdaBoost, XGBoost	Infantil, Fundamental, Médio	Dados demográficos e de engajamento	Precisão: 0.92, Recall: 0.72, AUC-PR: 0.68, KS: 0.86 (XGBoost)
(NAGY; MOLONTAY, 2024)	Predição de evasão interpretável: Intervenção personalizada baseada em XAI	Classificação	CatBoost, NGBoost, EBC, Regressão Logística, XGBoost, Gradient Boosting Classifier, entre outros	Superior	Notas pré-matrícula, Dados Demográficos e Indicadores de Campo de Estudo	AUC ROC 0.77, Precisão: 0.85 (CatBoost)
(SULAK; KOKLUB, 2024)	Predição de evasão estudantil usando aprendizado de máquina	Classificação	Árvores de Decisão, Floresta Aleatória, RNA	Superior	Dados Demográficos, Acadêmicos, Financeiros	Acurácia 0.77, Precisão 0.76, Recall: 0.77 e F1-Score 0.76 (RNA)

Continua na próxima página

Resumo dos Artigos Utilizados nos Trabalhos Relacionados (continuação)

Trabalho	Objetivo	Técnica	Algoritmo	Nível Educa- cional	Características Utilizadas	Métricas
(NDUNAGU et al., 2024)	Aprendizado profundo para prever a taxa de evasão em educação a distância	Classificação	LSTM, Rede Neural Convolutacional Unidimensional (1DCNN)	Superior	Dados acadêmicos, administrativos, econômicos e pessoais	Acurácia 0.57 (LSTM)
(KOK et al., 2024)	Predição do risco de evasão e priorização de fatores contribuintes	Classificação	Regressão Logística	Superior	Dados demográficos, acadêmicos e de engajamento	Acurácia: 85%, Precisão: 83%, Recall: 80%, F1-Score: 81%, AUC: 0.87

2.14 CONSIDERAÇÕES SOBRE O ESTADO DA ARTE E CONTRIBUIÇÕES DESTE TRABALHO

A literatura recente demonstra avanços significativos no uso de dados extraídos do Moodle para fins de predição de evasão, com diversas técnicas de aprendizado de máquina sendo empregadas em contextos institucionais variados. Estudos como os de Poellhuber et al. (2023) e Vaarma e Li (2024) mostram que abordagens baseadas em agrupamentos de cursos ou em variáveis de frequência de acesso possuem bom desempenho preditivo, ainda que muitas vezes limitadas a contextos institucionais específicos ou amostras restritas. Por sua vez, trabalhos como Santos e Henriques (2023) e Cechinel et al. (2024) têm buscado desenvolver modelos mais genéricos, capazes de generalizar para diferentes disciplinas ou instituições.

Apesar desses avanços, ainda existem lacunas importantes. Muitas abordagens limitam-se a modelos treinados em dados semanais estáticos, desconsiderando a estrutura temporal completa dos cursos. Além disso, há pouca discussão sobre a comparação entre estratégias de modelagem semanais e abordagens que consideram o curso como um todo. Outra limitação recorrente refere-se à baixa capacidade de generalização dos modelos quando aplicados a cursos ou contextos ainda não vistos, o que compromete sua aplicabilidade prática.

Este trabalho busca contribuir para o estado da arte ao propor uma abordagem comparativa entre modelos semanais e um modelo único com codificação temporal,

treinado com dados de múltiplas disciplinas. A partir de logs extraídos do Moodle, utilizam-se técnicas de balanceamento de classes (ADASYN) e otimização multiobjetivo (NSGA-II) para maximizar simultaneamente diferentes métricas de desempenho, como acurácia, recall e F1-score. Com isso, pretende-se avançar no desenvolvimento de modelos mais robustos e transferíveis, que possam ser utilizados em escala por instituições interessadas em identificar precocemente estudantes em risco de evasão.

3 METODOLOGIA

Esta seção descreve a metodologia adotada neste estudo, que se baseia no projeto LANSE (CECHINEL et al., 2024), abrangendo pré-processamento de dados, treinamento de modelo e otimização de hiperparâmetros usando o NSGA-II (DEB et al., 2002). O objetivo foi prever a evasão estudantil e o fracasso em disciplinas, aproveitando um espaço de busca complexo que inclui seleção de modelo, seleção de características, ajuste de hiperparâmetros, normalização e configuração de oversampling. Foram conduzidos dois experimentos distintos, utilizando diferentes conjuntos de dados para treinamento e teste. A metodologia segue um processo estruturado, desde a coleta de dados até a avaliação do modelo, conforme ilustrado na Figura 1.

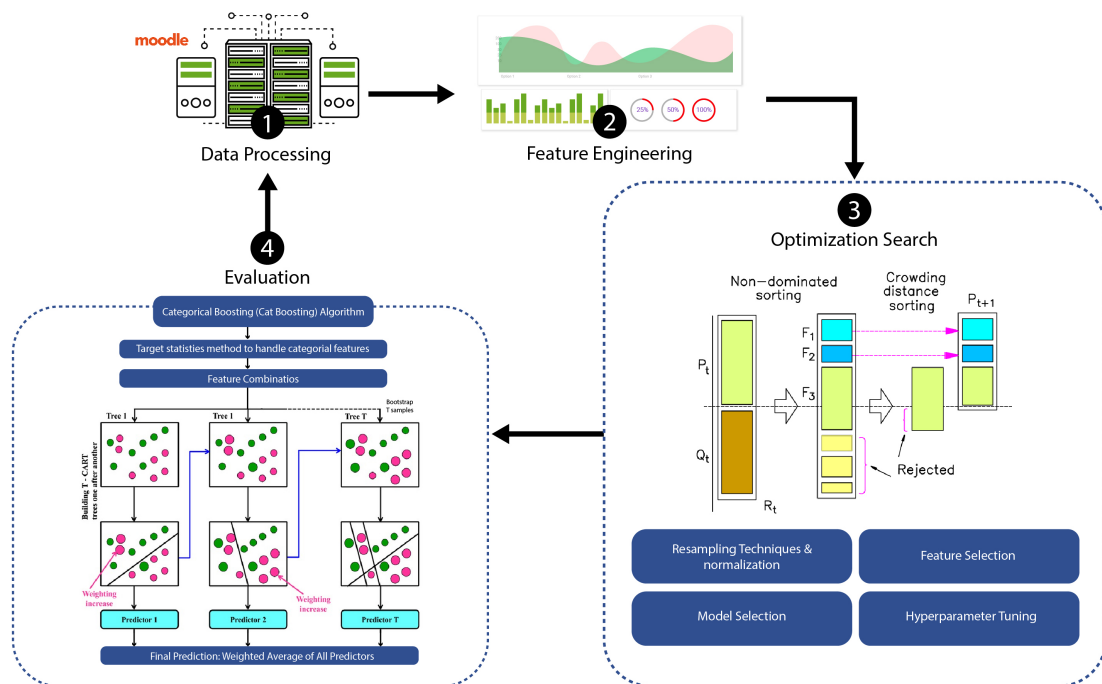


Figura 1 – Visão geral da metodologia Fonte: Elaborado pelo autor

A metodologia consiste em várias etapas principais. O primeiro passo, o processamento de dados, envolve a limpeza e formatação dos dados brutos para garantir consistência e preparação para a análise. Em seguida, a engenharia de características é aplicada para transformar ou criar novas variáveis que capturam padrões significativos nos dados, melhorando a precisão do modelo. O processo de otimização abrange a seleção de características para reduzir a dimensionalidade e focar nas variáveis mais relevantes, o ajuste de hiperparâmetros para refinar o desempenho do modelo e a escolha do algoritmo mais adequado para a tarefa. Técnicas de reamostragem,

como SMOTE e ADASYN, são empregadas para lidar com o desbalanceamento de classes. Por fim, o desempenho do modelo é rigorosamente avaliado usando métricas apropriadas, e novos ajustes são realizados para garantir robustez e precisão.

3.1 DESCRIÇÃO E PRÉ-PROCESSAMENTO DOS DADOS

O conjunto de dados utilizado neste estudo inclui 567 instâncias únicas de aluno-curso, com estudantes matriculados em 23 cursos distintos de diversas disciplinas. Inicialmente, todos os cursos foram configurados para 15 semanas, a duração mais comum dos cursos. No entanto, essa estrutura rígida limitava a flexibilidade dos modelos na generalização para cursos com diferentes durações. Para resolver esse problema, normalizamos o progresso de cada aluno com base na porcentagem do curso concluída, permitindo que o modelo fosse aplicado a cursos de diferentes durações.

O conjunto de dados é composto por 25 variáveis, derivadas principalmente dos registros de interação dos alunos no LMS. Essas variáveis capturam diversos aspectos do engajamento, atividade e progressão do aluno no curso. Um resumo dessas variáveis e suas descrições é apresentado na Tabela 2. Devido ao tamanho limitado do conjunto de dados, o sobreajuste foi uma preocupação importante, especialmente ao treinar modelos mais complexos. Para mitigar esse problema, aplicamos validação cruzada com cinco partições ($k=5$) durante o treinamento (JAMES et al., 2013) e implementamos critérios de parada antecipada em modelos como o CatBoost (DOROGUSH; ERSHOV; GULIN, 2018). Além disso, utilizamos técnicas de sobresampling, como o ADASYN (HE et al., 2008), para lidar com o desbalanceamento de classes e experimentamos diversas técnicas de normalização (PEDREGOSA et al., 2012) para padronizar os valores das variáveis.

Tabela 2 – Resumo das Variáveis Utilizadas no Modelo

Variável	Descrição
max_consecutive_days_with_access	Número máximo de dias consecutivos em que o aluno acessou o curso.
max_consecutive_days_without_access	Número máximo de dias consecutivos sem acesso ao curso.
first_log_days_diff	Dias entre o início do curso e o primeiro registro de acesso do aluno.

Continua na próxima página

Resumo das Variáveis Utilizadas no Modelo (continuação) Fonte: Elaborado pelo autor

Variável	Descrição
max_consecutive_days_with_access_week	Número máximo de dias consecutivos com acesso ao curso em uma semana.
max_consecutive_days_without_access_week	Número máximo de dias consecutivos sem acesso ao curso em uma semana.
logs	Número total de registros de acesso.
week_logs	Número de registros de acesso em uma semana.
daily_avg	Média diária de registros de acesso.
weekly_avg	Média semanal de registros de acesso.
days_with_logs	Número total de dias com registros de acesso.
days_with_logs_avg	Média de dias com registros de acesso.
days_with_logs_week	Dias com registros de acesso em uma semana.
days_with_logs_avg_week	Média de dias com registros de acesso em uma semana.
activity	Registros categorizados como atividades gerais.
content	Registros categorizados como interações com o conteúdo.
other	Registros que não se enquadram em categorias predefinidas.
report	Registros relacionados a atividades de relatório.
system	Registros que refletem interações com o sistema.

Continua na próxima página

Resumo das Variáveis Utilizadas no Modelo (continuação) Fonte: Elaborado pelo autor

Variável	Descrição
activity_week	Registros de atividade por semana.
content_week	Registros de interações com o conteúdo por semana.
other_week	Registros categorizados como "outros" por semana.
report_week	Registros de atividades de relatório por semana.
system_week	Registros de interação com o sistema por semana.
course progress*	Progresso do aluno no curso (semana atual / total de semanas).
total_weeks*	Número total de semanas do curso.

* Essas métricas foram utilizadas apenas no modelo que inclui semanas como variável.

Nosso estudo utilizou conjuntos de dados de 23 cursos distintos em diversas disciplinas para ambos os experimentos. Os conjuntos de dados do *Experimento1* e do *Experimento2* apresentaram estruturas semelhantes, incluindo o número de alunos únicos, cursos únicos e a distribuição dos resultados dos alunos. No entanto, o *Experimento2* introduziu duas novas variáveis relacionadas ao progresso do curso e à sua duração.

A Tabela 3 apresenta um resumo detalhado do conjunto de dados utilizado nos dois experimentos, incluindo o número de instâncias aluno-curso, número de alunos únicos, cursos únicos e a porcentagem de alunos aprovados ou reprovados nos conjuntos de treinamento e teste.

Tabela 3 – Resumo da Distribuição do Conjunto de Dados para os Experimentos 1 e 2 Fonte: Elaborado pelo autor

Métrica	Experimento 1		Experimento 2	
	Treinamento	Teste	Treinamento	Teste
Tamanho (instâncias aluno-curso)	2424	3107	3871	1660
Alunos Únicos	220	347	544	471
Cursos Únicos	11	12	23	23
Alunos Aprovados (%)	80.49%	63.82%	71.12%	71.14%
Alunos Reprovados (%)	19.51%	36.18%	28.88%	28.86%
Progresso do Curso e Duração como Variáveis	Não incluído		Incluído	



Figura 2 – Distribuição das Categorias de Cursos nos Conjuntos de Treinamento e Teste (*Experimento1*) Fonte: Elaborado pelo autor

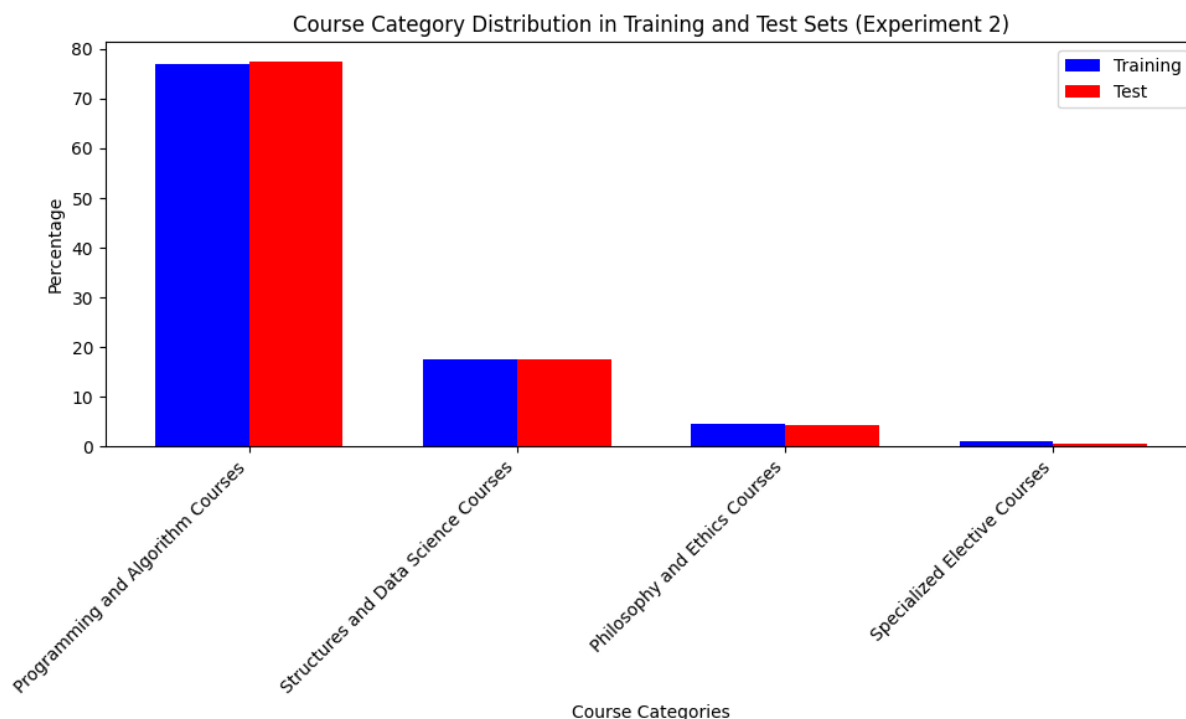


Figura 3 – Distribuição das Categorias de Cursos nos Conjuntos de Treinamento e Teste (*Experimento2*) Fonte: Elaborado pelo autor

Como observado na Tabela 3, ambos os experimentos envolveram conjuntos de treinamento com 3871 instâncias e conjuntos de teste com 1660 instâncias, garantindo tamanhos de dados consistentes nos dois experimentos. As categorias de cursos foram agrupadas em quatro categorias principais: Cursos de Programação e Algoritmos, Cursos de Estruturas e Ciência de Dados, Cursos de Filosofia e Ética e Cursos Optativos Especializados. A distribuição dessas categorias nos conjuntos de treinamento e teste de ambos os experimentos é visualizada nas Figuras 2 e 3.

3.2 ESTRATÉGIA DE OTIMIZAÇÃO

A escolha do modelo final foi baseada em uma comparação sistemática entre diferentes algoritmos, visando identificar aquele com maior desempenho na predição de alunos em risco de evasão. Inicialmente, testamos diversos classificadores amplamente utilizados em problemas educacionais, incluindo Floresta Aleatória, XGBoost, CatBoost, k-Nearest Neighbors (k-NN), Regressão Logística e Naïve Bayes. A seleção desses algoritmos foi motivada tanto pela sua recorrência na literatura quanto pela diversidade de abordagens (árvores, métodos lineares, baseados em instâncias).

Testes preliminares com uma plataforma AutoML (*Automated Machine Learning*) auxiliaram na triagem inicial das arquiteturas mais promissoras, que foram então incluídas na etapa posterior de otimização. O algoritmo NSGA-II foi empregado para

realizar uma busca multiobjetivo que englobou três dimensões: escolha do modelo, seleção de variáveis e ajuste de hiperparâmetros. Isso permitiu uma exploração robusta do espaço de busca, considerando simultaneamente métricas de desempenho como F1-score macro e recall para a classe minoritária.

Para os modelos baseados em *boosting*, como CatBoost e XGBoost, foram otimizados hiperparâmetros essenciais, incluindo taxa de aprendizado, número de estimadores e profundidade das árvores. Adicionalmente, avaliamos técnicas de normalização como *robust*, *standard* e *min-max scaling* (PEDREGOSA et al., 2012), assegurando compatibilidade entre variáveis com diferentes escalas. O desbalanceamento entre as classes foi tratado com ADASYN (HE et al., 2008), técnica que demonstrou sinergia com o CatBoost, melhorando sua sensibilidade na detecção de alunos em risco.

Como resultado desse processo comparativo e de otimização, o CatBoost foi selecionado como modelo final por apresentar os melhores resultados em termos de estabilidade, recall e F1-score para a classe de risco, além de manter boa interpretabilidade com auxílio dos valores SHAP.

3.3 MEDIÇÕES ESTATÍSTICAS E DE DISTÂNCIA PARA AVALIAÇÃO DO MODELO

Empregamos dois testes estatísticos para avaliar a similaridade entre os conjuntos de treinamento e teste: o teste de Kolmogorov-Smirnov (KS) e o teste de Mann-Whitney U (U-Test). Essas métricas servem a diferentes propósitos na análise da consistência entre distribuições:

O teste KS é um procedimento não paramétrico utilizado para determinar se duas amostras pertencem à mesma população. Ele faz isso medindo a maior diferença vertical entre as funções de distribuição empíricas das duas amostras, representando a maior divergência entre suas distribuições (NGUYEN; CARLIN; LEE, 2013). Um valor estatístico KS menor e um p -valor maior indicam maior similaridade entre os conjuntos de dados, enquanto diferenças significativas (valor KS maior e p -valor menor) sugerem possíveis discrepâncias que podem impactar o desempenho do modelo.

O teste de Mann-Whitney U é um teste não paramétrico que examina diferenças na tendência central (mediana) entre duas amostras independentes sem assumir normalidade (NACHAR, 2008). Um valor estatístico U menor e um p -valor reduzido indicam diferenças significativas nas tendências centrais, o que pode implicar em variações nas características dos conjuntos de treinamento e teste. Por outro lado, um p -valor maior sugere similaridade, indicando que o modelo provavelmente terá um desempenho consistente entre os conjuntos de dados.

Os resultados de ambos os testes são apresentados na Tabela 4.

Tabela 4 – Resultados dos Testes KS e Mann-Whitney U para os Experimentos 1 e 2 Fonte: Elaborado pelo autor

Métrica	<i>Experimento1</i>		<i>Experimento2</i>	
	Estatística	<i>p</i> -valor	Estatística	<i>p</i> -valor
Teste KS	0.069	3.48×10^{-131}	0.003	0.938
Teste Mann-Whitney U	2.09×10^9	6.33×10^{-49}	1.70×10^9	0.489

Ao interpretar esses resultados, o teste KS para o *Experimento1* apresentou um valor estatístico KS de 0.069 com um *p*-valor extremamente baixo (3.48×10^{-131}), indicando uma diferença estatisticamente significativa entre as distribuições dos conjuntos de treinamento e teste, sugerindo desafios na generalização do modelo. Em contraste, o *Experimento2* obteve um valor estatístico KS muito menor (0.003) com um *p*-valor de 0.938, indicando que não há diferença significativa entre as distribuições dos conjuntos de treinamento e teste, o que é favorável para a generalização. Da mesma forma, o teste de Mann-Whitney U no *Experimento1* revelou um *p*-valor baixo (6.33×10^{-49}), evidenciando uma diferença notável nas tendências centrais, o que pode impactar a capacidade do modelo de generalizar para dados não vistos. No entanto, no *Experimento2*, o alto *p*-valor (0.489) sugere que não há diferença significativa nas tendências centrais, reforçando o alinhamento entre os conjuntos de treinamento e teste. Esses achados indicam que o *Experimento2* se beneficia de uma maior consistência entre os dados de treinamento e teste, o que provavelmente melhora a confiabilidade e a generalização do desempenho do modelo.

3.4 TREINAMENTO E AVALIAÇÃO DO MODELO

O conjunto de dados foi dividido em dois experimentos para avaliar e aprimorar o desempenho do modelo sob diferentes configurações. No *Experimento1*, 220 alunos únicos foram alocados para o conjunto de treinamento e 357 para o conjunto de teste *holdout*, garantindo que os dados de teste fossem distintos dos utilizados no treinamento inicial dos modelos anteriores. Foi aplicada validação cruzada com 5 partições (JAMES et al., 2013) para avaliar a consistência do desempenho e minimizar o sobreajuste.

O *Experimento2* introduziu variáveis baseadas em semanas, permitindo que o modelo ajustasse dinamicamente seu aprendizado com base no progresso dos alunos. Além disso, utilizou um conjunto de teste aleatório correspondente a 30% do total de dados disponíveis. Essa divisão aleatória proporcionou uma avaliação mais ampla da capacidade de generalização do modelo, pois representava diferentes períodos de tempo e padrões de engajamento estudantil.

Inicialmente, foram treinados modelos distintos para cada semana (até a semana 15). O *Experimento2* simplificou esse processo ao introduzir variáveis de semanas relativas, como *semana atual / total de semanas* e *total de semanas do curso*, o que permitiu que um único modelo lidasse com dados de cursos com diferentes durações.

Após a realização desses experimentos e uma otimização extensiva, o CatBoost foi escolhido como a arquitetura final. Seu forte desempenho nos testes de validação cruzada e no conjunto *holdout* validou sua seleção, uma vez que o CatBoost demonstrou notável adaptabilidade a variáveis estáticas e dinâmicas. A capacidade nativa do CatBoost de lidar com dados categóricos e seus métodos eficazes de regularização minimizaram o sobreajuste, o que se mostrou particularmente útil no conjunto de dados desbalanceado.

3.5 ENGENHARIA DE CARACTERÍSTICAS

A engenharia de características concentrou-se na extração de variáveis significativas a partir dos registros de interação no LMS. Essas variáveis estavam relacionadas ao engajamento dos alunos e incluíam métricas como:

- **Contagem de Logs:** Número total de registros, registros semanais, médias diárias e tipos específicos de atividades (por exemplo, conteúdo, uso do sistema).
- **Padrões de Acesso:** Máximo de dias consecutivos com/sem acesso, dias com registros de acesso e média de dias com registros de acesso.
- **Características Temporais:** Número de dias desde o primeiro acesso, diferença de tempo entre acessos consecutivos e progressão relativa no curso, calculada como a razão entre a semana atual e a duração total do curso.

3.6 REPRODUTIBILIDADE DO ESTUDO

A metodologia proposta neste estudo pode ser replicada seguindo um processo estruturado, incluindo coleta de dados, pré-processamento, treinamento do modelo e avaliação. O conjunto de dados consiste em registros de interação dos alunos extraídos de um LMS, capturando padrões de acesso, participação em atividades e engajamento com o conteúdo do curso. Instituições que adotarem essa abordagem devem garantir a correta extração, anonimização e estruturação dos dados para que estejam alinhados com as características utilizadas neste estudo. Os dados devem ser organizados de forma que cada registro represente uma instância aluno-curso, contendo métricas de engajamento, como número total de registros, dias com acesso e progresso no curso, juntamente com o resultado final do aluno (aprovado ou reprovado).

O treinamento do modelo segue os passos descritos, incluindo pré-processamento para tratamento de valores ausentes, normalização de variáveis e aplicação de técnicas de balanceamento para lidar com o desbalanceamento de classes. Para uso prático, o modelo pode ser integrado a um sistema que visualize o engajamento dos alunos e suas previsões, como o Lanse, permitindo que educadores monitorem e intervenham quando necessário.

4 A PLATAFORMA LANSE

Com o objetivo de aplicar os modelos desenvolvidos em um ambiente real e contribuir diretamente com o processo de tomada de decisão nas instituições de ensino, este capítulo apresenta a plataforma LANSE. Esta solução computacional de *learning analytics* foi o ambiente no qual os modelos preditivos de evasão foram integrados, operando em conjunto com o AVA Moodle utilizado pela instituição parceira. A seguir, detalha-se o funcionamento da plataforma, seus componentes técnicos, a forma de coleta e processamento dos dados, e sua importância no apoio à redução da evasão estudantil.

A plataforma LANSE representa a materialização prática dos modelos preditivos de evasão estudantil desenvolvidos neste estudo, integrando-se diretamente ao ambiente Moodle utilizado pelas instituições de ensino. Sua arquitetura foi projetada para facilitar a coleta, processamento e visualização de dados educacionais com foco em Learning Analytics, apoiando a tomada de decisão pedagógica e administrativa (CECHINEL et al., 2024; CECHINEL et al., 2023).

O sistema inicia seu funcionamento a partir de um plugin instalado no AVA, especificamente no Moodle. Esse plugin tem como principal função extrair continuamente os registros de interação dos estudantes com o ambiente, como acessos, visualizações de conteúdos e participação em atividades. Os dados são coletados em lotes (*batches*) e enviados para a plataforma LANSE na nuvem, onde são armazenados e processados (CECHINEL et al., 2024; CECHINEL et al., 2023).

Na camada de processamento em nuvem, os dados são tratados e utilizados por modelos preditivos automatizados, treinados com algoritmos de aprendizado de máquina. Esses modelos, cuja lógica é fundamentada em trabalhos prévios como os de Macarini et al. (2019) e Detoni et al. (2016), geram previsões semanais sobre o risco de evasão e reprovação. As previsões são disponibilizadas por meio de um painel interativo, voltado a educadores e gestores, com informações visuais que facilitam a interpretação dos resultados e a tomada de decisão (CECHINEL et al., 2024; CECHINEL et al., 2023).

A plataforma prioriza o uso de variáveis extraídas a partir da contagem de interações dos estudantes, assim como atributos derivados dessas contagens. Essa escolha se alinha a evidências da literatura, que apontam a eficácia dessas métricas na previsão do desempenho e engajamento estudantil (MACARINI et al., 2019; DETONI et al., 2016).

O papel do LANSE vai além da simples geração de previsões. Ele permite

o monitoramento contínuo dos padrões de acesso e engajamento dos estudantes, oferecendo visualizações detalhadas que auxiliam na identificação precoce de sinais de risco. Os modelos são atualizados semanalmente e apresentam seus resultados com um nível de confiança estimado, permitindo que os professores avaliem não apenas o risco atribuído a cada aluno, mas também a confiabilidade dessa estimativa (CECHINEL et al., 2024).

Ao identificar estudantes com maior probabilidade de evasão, o sistema envia alertas aos professores e coordenadores, facilitando a adoção de estratégias de intervenção pedagógica personalizadas. Isso inclui, por exemplo, o envio de mensagens automatizadas de apoio, a oferta de encontros de orientação ou a adaptação do conteúdo das aulas para aumentar o engajamento de alunos com risco identificado (CECHINEL et al., 2023).

O dashboard do LANSE também fornece uma visão geral das turmas, destacando métricas como a taxa média de acesso, a proporção de alunos em diferentes faixas de risco e a evolução dessas métricas ao longo do tempo. Essa perspectiva global auxilia tanto os educadores quanto os gestores institucionais no acompanhamento do progresso acadêmico em diferentes disciplinas e cursos (CECHINEL et al., 2024; CECHINEL et al., 2023).

Relatos preliminares indicam que o uso do LANSE tem contribuído significativamente para a redução da carga de trabalho dos coordenadores e professores, ao automatizar processos anteriormente realizados de forma manual. Antes da adoção do sistema, a identificação de estudantes em risco era baseada na intuição dos docentes ou em análises pontuais de relatórios de acesso. Com o LANSE, esse processo se torna sistematizado, baseado em dados e sustentado por modelos preditivos calibrados empiricamente (CECHINEL et al., 2024).

Em síntese, o LANSE consolida-se como uma solução de Learning Analytics que viabiliza a aplicação prática de modelos preditivos treinados a partir de dados do Moodle. Sua integração com o AVA, o uso de previsões automatizadas e a entrega de informações acionáveis por meio de dashboards interativos tornam a plataforma uma ferramenta valiosa para a promoção do sucesso acadêmico e a redução das taxas de evasão escolar (CECHINEL et al., 2024; CECHINEL et al., 2023).

5 RESULTADOS

5.1 MODELOS EXISTENTES PARA COMPARAÇÃO

Os modelos utilizados para comparação neste estudo foram previamente treinados utilizando dois algoritmos de aprendizado de máquina bem estabelecidos: XGBoost e Floresta Aleatória. Esses modelos, desenvolvidos em um estudo anterior, foram treinados em um conjunto de dados derivado principalmente dos registros de atividades dos alunos na plataforma Moodle. As principais características incluíam a contagem de logs (frequência das interações dos alunos) e os padrões de acesso (frequência e horário de acesso à plataforma). Essas variáveis foram projetadas para capturar o engajamento básico dos alunos como preditores de risco de evasão. Os dados e os modelos foram organizados semanalmente, ou seja, cada semana tinha um modelo correspondente. Foi utilizada uma busca Bayesiana para selecionar tanto os parâmetros do modelo quanto as estratégias para lidar com o desbalanceamento de classes, incluindo a implementação do SMOTE.

Esses modelos serviram como referências valiosas para a avaliação da abordagem deste estudo e foram implementados na aplicação em nuvem LANSE (CECHINEL et al., 2024). Apesar de sua utilidade, apresentavam algumas limitações. O conjunto de características era limitado a dados básicos de interação, sem uma análise mais profunda do comportamento dos alunos ou da qualidade do conteúdo acessado, nem do tempo gasto em atividades específicas. Essa limitação reduzia a capacidade dos modelos de capturar a complexidade do engajamento estudantil. Além disso, embora o SMOTE tenha sido aplicado para lidar com o desbalanceamento de classes, os modelos enfrentaram desafios para generalizar bem para novos dados, pois foi observado sobreajuste em alguns casos. Isso sugere que os modelos ocasionalmente aprenderam padrões específicos dos dados de treinamento, limitando seu desempenho preditivo em novos conjuntos de dados.

Embora tenha sido realizada uma busca Bayesiana para otimizar hiperparâmetros e lidar com o desbalanceamento de classes, ambos os modelos ainda dependiam de ajustes relativamente básicos e poderiam se beneficiar de uma otimização mais avançada. Técnicas mais sofisticadas e uma otimização mais profunda poderiam melhorar a capacidade dos modelos de capturar as complexidades dos dados. Optamos por utilizar o NSGA-II para explorar de forma eficiente os trade-offs entre múltiplos objetivos conflitantes. Melhorias futuras também poderiam expandir o conjunto de características para incluir métricas mais sofisticadas, como padrões comportamentais detalhados e interações específicas com o conteúdo, aprimorando a precisão e a capacidade de

generalização dos modelos para a identificação de alunos em risco.

5.2 MODELO PROPOSTO

Uma nova abordagem foi desenvolvida para superar as limitações do modelo existente. Foram exploradas diferentes arquiteturas e, ao final da otimização com NSGA-II, o algoritmo CatBoost foi selecionado. A nova abordagem empregou a otimização de hiperparâmetros utilizando NSGA-II, permitindo uma exploração abrangente do espaço de parâmetros. Esse processo de otimização incluiu a seleção de características, o ajuste fino de hiperparâmetros e a escolha da estratégia de reamostragem (seleção e configuração do algoritmo).

Inicialmente, o modelo foi treinado com dados de cursos com duração fixa de 15 semanas, como na abordagem original. No entanto, em uma fase subsequente, a duração dos cursos foi normalizada para acomodar diferentes extensões. O desempenho do modelo foi comparado nos cenários de cursos fixos com 15 semanas e com duração normalizada.

As principais melhorias incluíram:

- **Seleção de características:** O novo modelo incorporou variáveis adicionais, como progresso relativo no curso, e forneceu uma análise mais granular das métricas semanais de engajamento, incluindo o tipo de log registrado.
- **Otimização de hiperparâmetros:** O NSGA-II foi utilizado para otimizar os modelos. O modelo CatBoost incluiu parâmetros ajustados como *seed*, taxa de aprendizado, profundidade das árvores e número de árvores, resultando em um modelo mais refinado.
- **Lidando com o desbalanceamento de classes:** Diversas estratégias de reamostragem foram testadas, sendo o ADASYN escolhido para superamostrar a classe minoritária (alunos evadidos). Essa abordagem mitigou o desbalanceamento do conjunto de dados ao tratar a sub-representação de alunos que abandonaram o curso.

A Figura 4 apresenta a frente de Pareto tridimensional obtida com o algoritmo NSGA-II, otimizando simultaneamente três métricas: acurácia, F1-score macro e F1-score da classe de alunos reprovados. Os pontos coloridos representam indivíduos (modelos) gerados ao longo das gerações, sendo que a escala de cor indica a geração correspondente — do roxo (inicial) ao amarelo (mais evoluído). É possível observar uma clara convergência das soluções em direção a uma borda superior da frente de Pareto, sugerindo que o algoritmo foi eficiente em encontrar compromissos ótimos entre as métricas conflitantes.

Notadamente, muitos dos melhores indivíduos alcançaram altos valores de F1-score para a classe minoritária (reprovados), sem comprometer substancialmente a acurácia ou o desempenho macro geral. Isso é relevante, pois indica que o modelo conseguiu equilibrar bem o foco em uma classe crítica (os alunos em risco), sem sacrificar o desempenho global. Esse comportamento valida o uso do NSGA-II como estratégia de otimização robusta, especialmente em cenários desbalanceados como o da reprovação estudantil.

Além disso, a análise da densidade de pontos nas gerações finais indica que a população foi se concentrando em regiões do espaço de busca com desempenho consistente em múltiplas métricas, o que reforça a estabilidade da solução. Para o modelo final, selecionado a partir dessa frente, utilizaram-se probabilidades de mutação, cruzamento e troca de 10%, 70% e 40%, respectivamente, equilibrando custo computacional e qualidade preditiva.

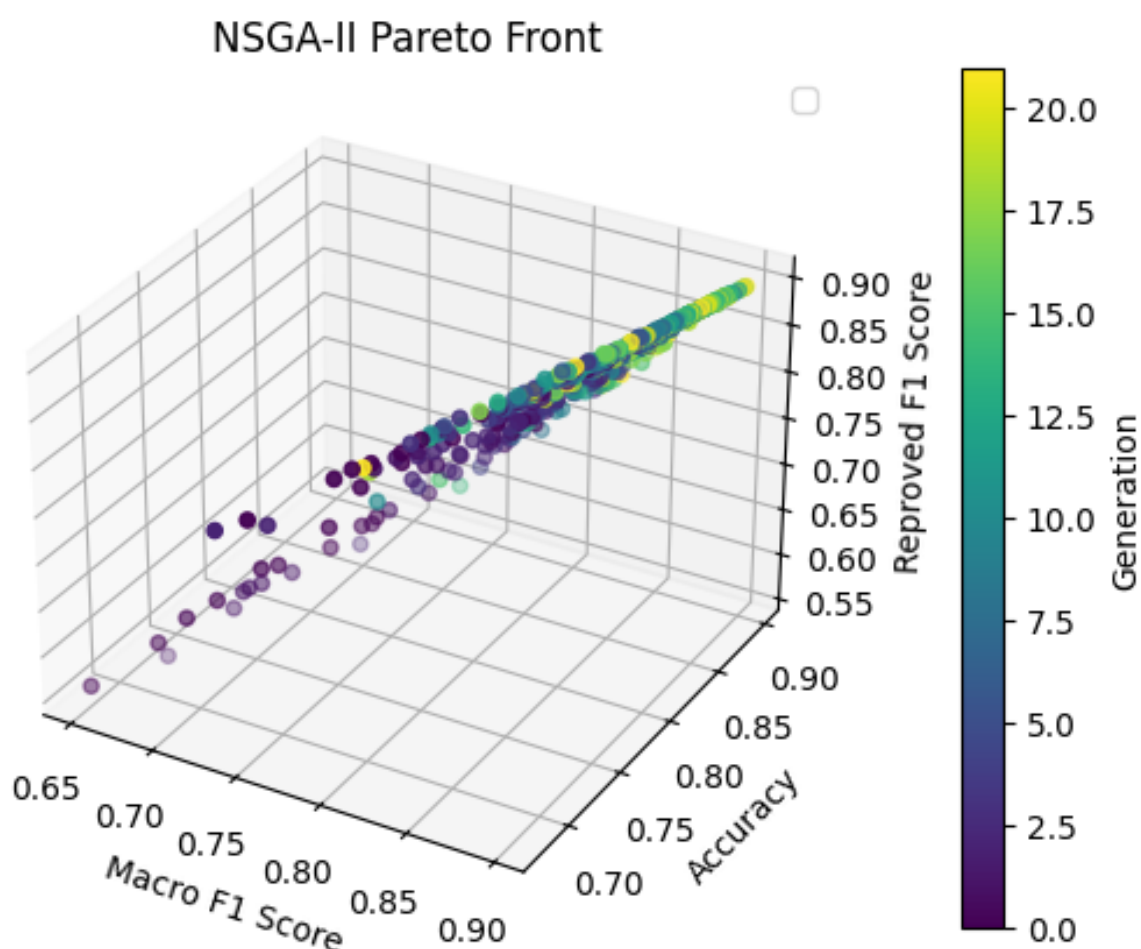


Figura 4 – Frente de Pareto 3D da busca NSGA-II mostrando a convergência das populações em termos de acurácia, F1-score macro e F1-score para alunos reprovados. Fonte: Elaborado pelo autor

5.3 MÉTRICAS DE DESEMPENHO

O desempenho do modelo existente e do modelo proposto, tanto na versão fixa de 15 semanas quanto no modelo mais flexível que utiliza a semana relativa (atual/total) e o total de semanas como variáveis. Foi comparado utilizando métricas-chave como acurácia, F1-score, recall e $AUC - ROC$. As Tabelas 5 e 6 resumem o desempenho dos modelos. Mesmo o modelo flexível foi avaliado com duração de 15 semanas nos testes *holdout*, conforme mostrado nas tabelas, permitindo uma comparação direta com os modelos anteriores.

Tabela 5 – Comparação de Desempenho no Conjunto de Teste Holdout. Novos modelos possuem um modelo distinto para cada semana (abordagem antiga). Fonte: Elaborado pelo autor

Sem.	Modelo	F1		Rec.		Acc.	AUC
		Aprov.	Reprov.	Aprov.	Reprov.		
1	Antigo	0.56	0.28	0.63	0.24	0.46	0.44
	Novo	0.55	0.25	0.53	0.27	0.44	0.40
3	Antigo	0.54	0.42	0.54	0.42	0.49	0.48
	Novo	0.55	0.50	0.46	0.64	0.53	0.56
5	Antigo	0.65	0.25	0.83	0.18	0.52	0.40
	Novo	0.65	0.40	0.59	0.47	0.56	0.55
7	Antigo	0.62	0.32	0.70	0.27	0.51	0.50
	Novo	0.74	0.27	0.69	0.18	0.61	0.46
9	Antigo	0.72	0.28	0.95	0.17	0.59	0.61
	Novo	0.66	0.43	0.76	0.35	0.57	0.56
11	Antigo	0.71	0.33	0.85	0.24	0.60	0.63
	Novo	0.65	0.35	0.73	0.29	0.55	0.50
13	Antigo	0.79	0.16	0.98	0.09	0.66	0.61
	Novo	0.74	0.36	0.81	0.29	0.63	0.54
15	Antigo	0.77	0.12	0.97	0.06	0.63	0.60
	Novo	0.76	0.34	0.89	0.24	0.65	0.52

Tabela 6 – Comparação de Desempenho no Conjunto de Teste Holdout. Modelo único vs múltiplos modelos por semana

Sem.	Modelo	F1		Rec.		Acc.	AUC
		Aprov.	Reprov.	Aprov.	Reprov.		
1	Múltiplos	0.63	0.49	0.56	0.60	0.57	0.63
	Único	0.73	0.51	0.73	0.50	0.65	0.66
3	Múltiplos	0.82	0.57	0.86	0.51	0.75	0.74
	Único	0.83	0.58	0.88	0.57	0.76	0.84
5	Múltiplos	0.82	0.55	0.78	0.63	0.74	0.76
	Único	0.92	0.74	0.94	0.68	0.86	0.92
7	Múltiplos	0.91	0.79	0.89	0.82	0.87	0.90
	Único	0.94	0.86	0.95	0.85	0.90	0.95
9	Múltiplos	0.90	0.74	0.95	0.65	0.85	0.90
	Único	0.97	0.94	0.96	0.93	0.95	1.00
11	Múltiplos	0.76	0.58	0.70	0.68	0.69	0.78
	Único	0.89	0.81	0.98	0.91	0.94	0.98
13	Múltiplos	0.92	0.67	0.92	0.64	0.87	0.88
	Único	0.98	0.96	1.00	0.93	0.99	0.97
15	Múltiplos	0.94	0.82	0.96	0.77	0.91	0.98
	Único	0.96	0.91	0.97	0.89	0.95	0.99

5.3.1 Comparação de Desempenho do Modelo

Esta seção compara o desempenho de duas metodologias de modelo: *Único* e *Múltiplos*. A comparação é baseada em métricas-chave de desempenho, como F1-score para alunos aprovados e reprovados e a Área Sob a Curva (AUC). As Figuras 5, 6 e 7 resumem os resultados ao longo das semanas.

5.3.1.1 F1-Score para Alunos Aprovados

A Figura 5 ilustra o desempenho do F1-score na previsão de alunos que concluíram o curso com aprovação. O modelo *Único* supera consistentemente o modelo *Múltiplos* em todas as semanas. A partir da Semana 2, o F1-score do modelo *Único* aumenta de forma constante, atingindo cerca de 0,95 na Semana 8, enquanto o modelo *Múltiplos* apresenta maior variabilidade, com uma queda notável por volta da Semana 12. Esses resultados sugerem que o modelo *Único* oferece maior estabilidade e precisão na previsão de alunos aprovados ao longo do tempo.

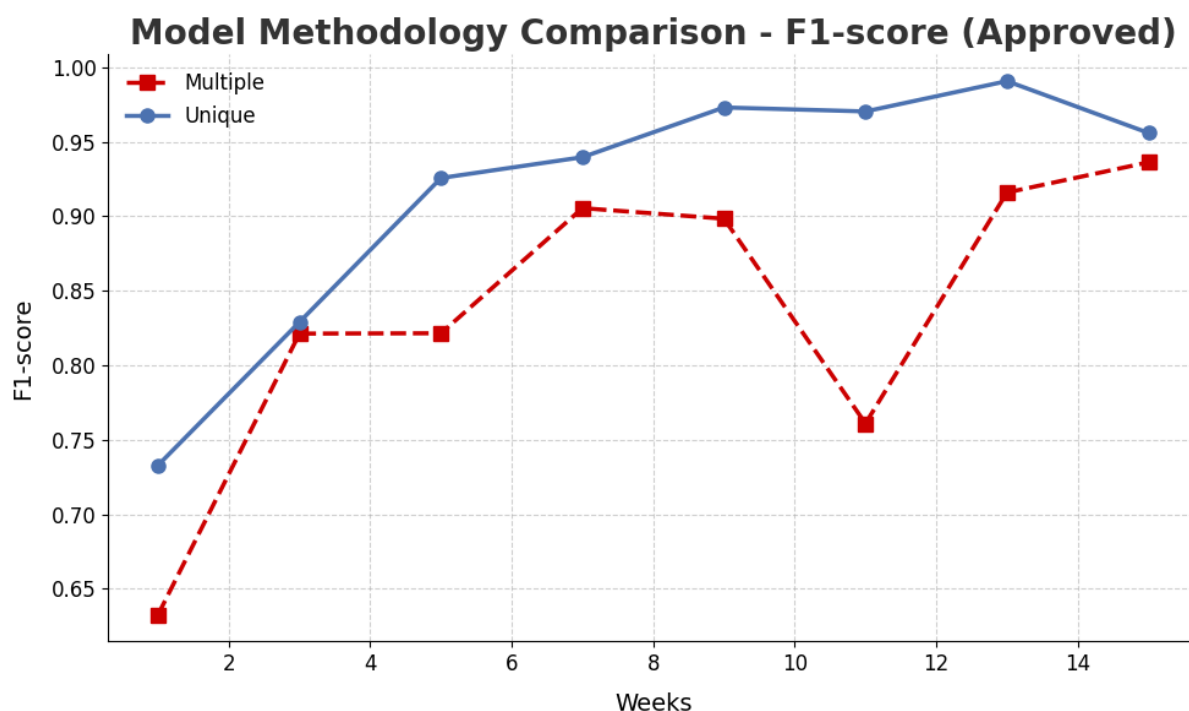


Figura 5 – Comparação de Metodologia do Modelo - F1-score (Aprovados) Fonte: Elaborado pelo autor

5.3.1.2 F1-Score para Alunos Reprovados

A Figura 6 apresenta o F1-score para alunos que foram reprovados no curso. Semelhante aos resultados dos alunos aprovados, o modelo *Único* melhora significativamente o desempenho em comparação com o modelo *Múltiplos*. O F1-score do modelo *Único* começa em 0,5 e sobe para aproximadamente 0,9 na Semana 8, mantendo esse desempenho até o final do curso. Em contraste, o modelo *Múltiplos* exibe pontuações consideravelmente mais baixas, especialmente nas Semanas 6 e 12, onde o desempenho cai significativamente. Essa análise sugere que o modelo *Único* é mais adequado para identificar alunos em risco de reprovação.

5.4 Comparação do AUC Score

A Figura 7 corrobora ainda mais a superioridade do modelo *Único*. O AUC para o modelo *Único* ultrapassa consistentemente 0.90 após a Semana 4, enquanto o modelo *Múltiplos* apresenta maior flutuação, com quedas significativas por volta das Semanas 6 e 12. Os valores de AUC mais estáveis do modelo *Único* refletem sua melhor capacidade de generalização ao longo das semanas, proporcionando previsões mais confiáveis tanto para alunos aprovados quanto para alunos em risco.

O AUC, mostrado na Figura 7, reforça o desempenho superior do modelo *Único*. Após a Semana 3, o AUC do modelo *Único* ultrapassa consistentemente 0.90, atingindo

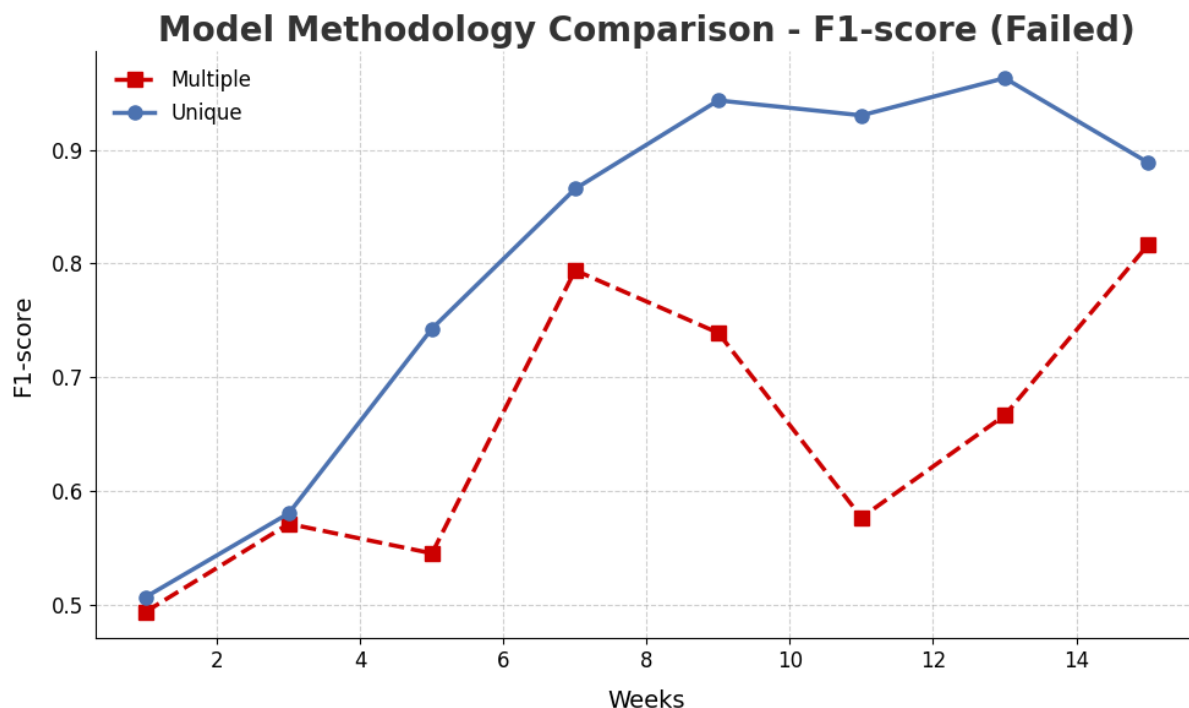


Figura 6 – Comparação de Metodologia do Modelo - F1-score (Reprovados) Fonte: Elaborado pelo autor

quase 1.00 na Semana 9 e mantendo esse nível nas semanas seguintes. Por outro lado, o modelo *Múltiplos* apresenta maior variabilidade, com quedas no AUC por volta das Semanas 7 e 11. Os valores de AUC mais estáveis do modelo *Único* indicam sua maior capacidade de generalização ao longo das semanas, oferecendo previsões mais confiáveis tanto para alunos aprovados quanto para alunos em risco.

5.5 Importância das Características

Para compreender melhor o processo de tomada de decisão do modelo, utilizamos SHAP para interpretar as contribuições das características nas previsões. A Figura 8 apresenta o gráfico de resumo do SHAP para o modelo final, destacando como cada característica influencia a previsão de evasão ou sucesso dos alunos.

O gráfico SHAP revela que características como *total_weeks* (duração total do curso em semanas), *days_with_logs_avg* (média de dias em que um aluno registrou atividade até a semana atual) e *max_consecutive_days_without_access* (máximo de dias consecutivos sem acessar a plataforma do curso) tiveram o maior impacto nas previsões do modelo para evasão e reprovação dos alunos. Especificamente, alunos com menos semanas de engajamento ou períodos prolongados sem acesso à plataforma foram mais propensos a serem classificados como em risco. Além disso, os resultados sugerem que alunos matriculados em cursos mais longos (com maior número total de semanas) apresentam maior probabilidade de evasão ou reprovação.

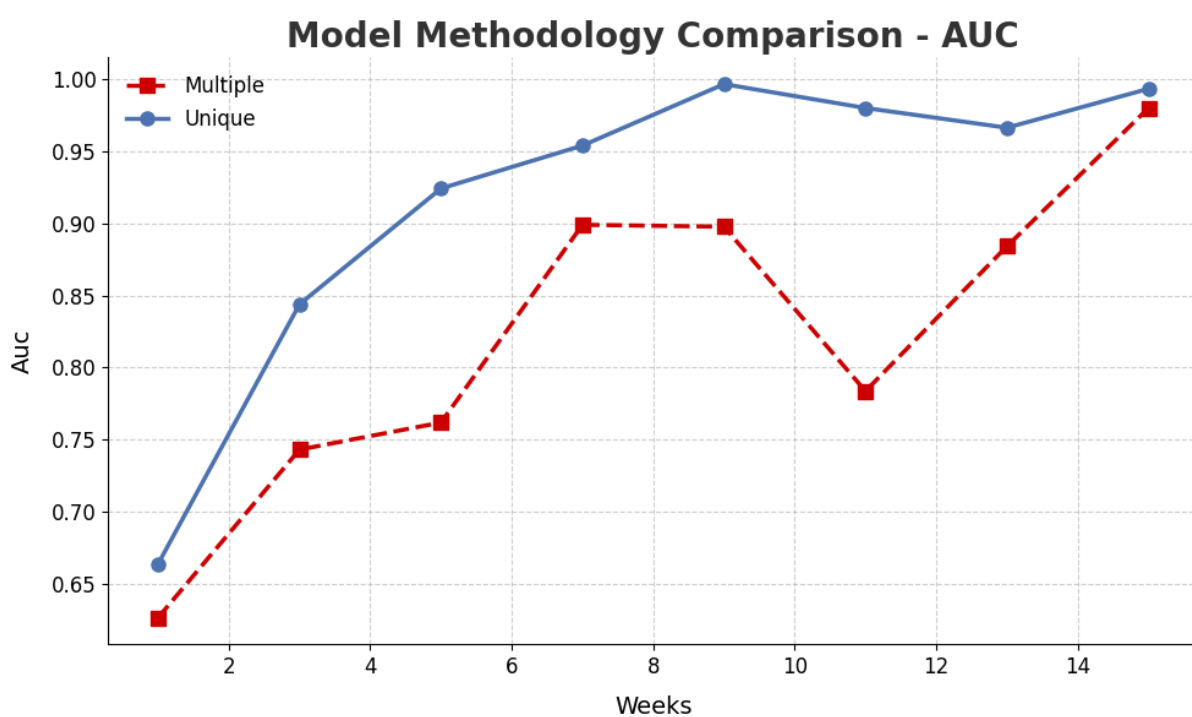


Figura 7 – Comparação de Metodologia do Modelo - AUC Fonte: Elaborado pelo autor

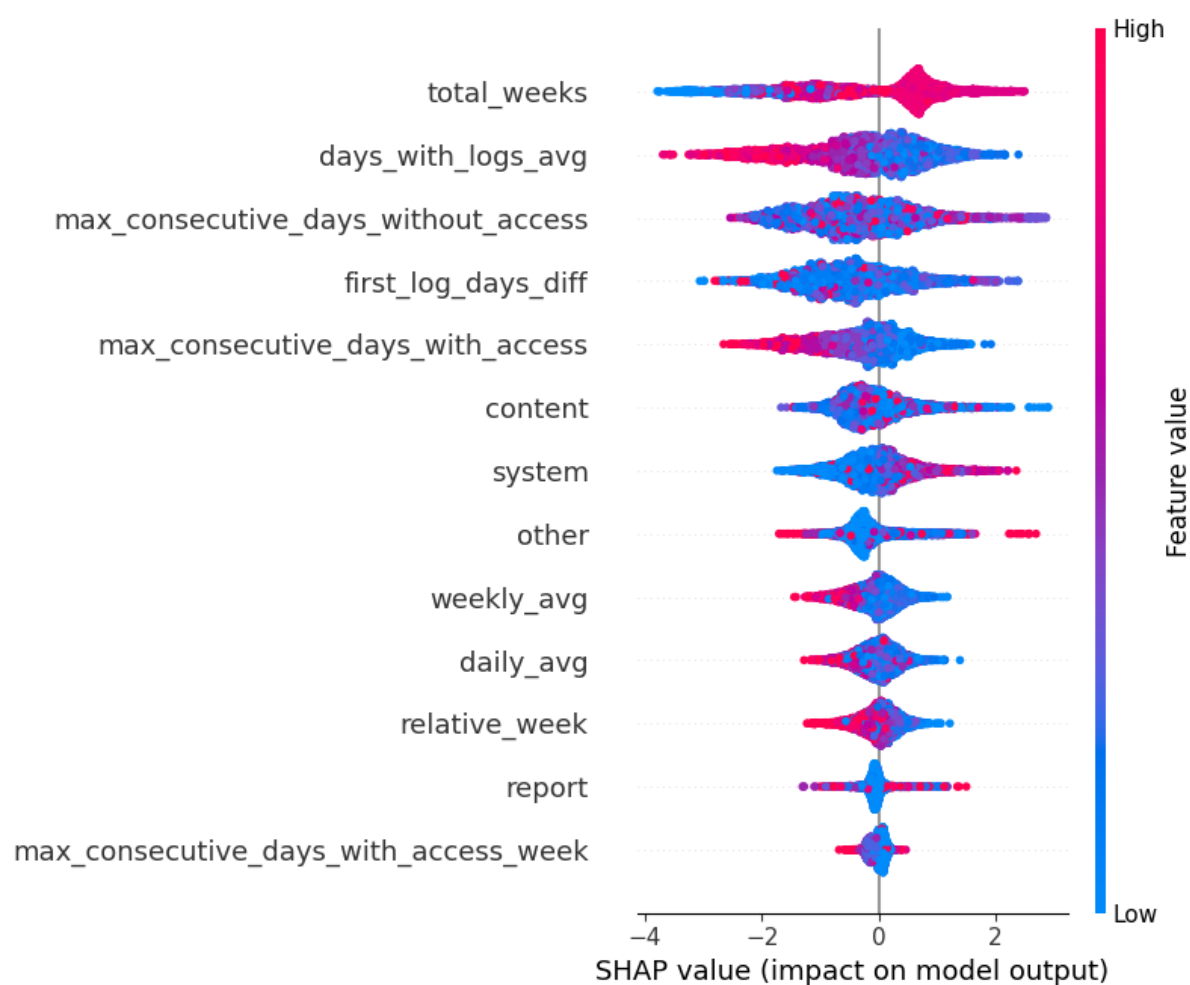


Figura 8 – Valores SHAP para importância das características. Vermelho representa valores mais altos da característica, e azul representa valores mais baixos. O eixo x mostra o impacto na saída do modelo. Fonte: Elaborado pelo autor

6 DISCUSSÃO

A comparação entre os modelos antigos (Random Forest e XGBoost) e o novo modelo (CatBoost com otimização via NSGA-II) demonstra melhorias claras no desempenho da abordagem proposta, particularmente em termos de recall, F1-score, acurácia e AUC.

F1-Score: Conforme observado nas Tabelas 5 e 6, o modelo proposto alcançou consistentemente F1-scores mais altos em comparação com os modelos antigos, especialmente na detecção de alunos em risco (classe de reprovados). Por exemplo, no *Experimento1*, na Semana 5, o F1-score para a classe de reprovados aumentou de 0,25 (modelo antigo) para 0,40 (modelo novo), enquanto no *Experimento2*, o F1-score melhorou de 0,55 (múltiplos modelos) para 0,74 (modelo único). Esses resultados destacam a capacidade significativamente aprimorada do novo modelo para identificar alunos em risco de reprovação.

Recall: O novo modelo apresentou escores de recall mais altos na maioria das semanas para a classe de reprovados, demonstrando sua capacidade de identificar uma maior proporção de alunos em risco. Por exemplo, no *Experimento1*, na Semana 7, o recall para a classe de reprovados diminuiu de 0,27 (modelo antigo) para 0,18 (modelo novo). No entanto, no *Experimento2*, o recall melhorou de 0,82 (múltiplos modelos) para 0,85 (modelo único). Isso demonstra uma melhora significativa na capacidade do novo modelo de identificar alunos em risco no *Experimento2*, tornando-o particularmente eficaz para intervenções precoces.

Acurácia: Assim como no caso do recall, a acurácia mostrou melhorias notáveis no *Experimento2*. No *Experimento1*, na Semana 9, a acurácia para a classe de reprovados diminuiu ligeiramente de 0,59 (modelo antigo) para 0,57 (modelo novo). No entanto, no *Experimento2*, a acurácia foi maior, subindo de 0,85 (múltiplos modelos) para 0,95 (modelo único). Isso indica que o modelo no *Experimento2* apresentou um desempenho muito melhor na identificação correta tanto de alunos reprovados quanto aprovados, destacando suas capacidades preditivas aprimoradas.

AUC: O AUC também demonstrou melhorias significativas, indicando a maior capacidade do modelo de distinguir entre alunos em risco e não em risco. No *Experimento1*, na Semana 9, o AUC diminuiu ligeiramente de 0,61 (modelo antigo) para 0,56 (modelo novo). No entanto, no *Experimento2*, o AUC melhorou significativamente, subindo de 0,90 (múltiplos modelos) para 1,00 (modelo único). Isso mostra que o modelo no *Experimento2* possui uma capacidade discriminatória muito mais forte na identificação de alunos em risco, aprimorando ainda mais sua aplicação prática para intervenções

precoces.

O modelo proposto, que incorpora a duração do curso como uma característica e utiliza técnicas robustas de otimização, demonstrou melhorias substanciais em relação aos modelos anteriores. Os modelos antigos eram limitados por dependerem apenas de dados acumulados até uma determinada semana, enquanto o novo modelo aproveita a otimização de hiperparâmetros via NSGA-II e integra características mais granulares, incluindo a duração e o progresso do curso. Isso permite que o modelo utilize de forma mais eficaz os dados de todas as semanas disponíveis. Como resultado, o novo modelo oferece melhor generalização ao longo do tempo, maior flexibilidade, escalabilidade e previsões significativamente mais precisas dos resultados dos alunos.

Esses resultados são particularmente importantes na mineração de dados educacionais, pois a detecção precoce de alunos em risco de reprovação pode permitir intervenções oportunas, ajudando, em última instância, a reduzir as taxas de evasão e melhorar o sucesso acadêmico geral dos alunos.

6.1 ANÁLISE DE ERROS E CONSIDERAÇÕES PEDAGÓGICAS

Embora os resultados quantitativos mostrem ganhos claros, é importante considerar os casos em que o modelo falhou. Por exemplo, falsos positivos podem representar alunos com engajamento irregular, mas que, por outros fatores (como motivação pessoal ou ajuda externa), conseguem ser aprovados. Já falsos negativos podem incluir alunos que acessam regularmente o ambiente virtual, mas não conseguem acompanhar o conteúdo, o que reforça a necessidade de indicadores mais ricos de aprendizagem além do mero acesso. Esses casos evidenciam a importância de integrar dados comportamentais e acadêmicos com o julgamento docente no uso dos modelos.

6.2 IMPORTÂNCIA DAS VARIÁVEIS: INTERPRETAÇÃO COM SHAP

A interpretação do modelo por meio da análise SHAP, apresentada na Figura 8, revela que as variáveis com maior impacto nas previsões foram *total_weeks* (duração total do curso), *days_with_logs_avg* (média de dias com acesso até a semana de corte) e *max_consecutive_days_without_access* (máximo de dias consecutivos sem acessar o AVA). A variável *total_weeks*, em particular, representa o número total de semanas de duração do curso em que o aluno está matriculado. Essa informação foi fundamental para contextualizar o progresso e o comportamento do estudante ao longo do tempo, permitindo que o modelo diferenciasse padrões de engajamento conforme a carga temporal total do curso.

Os estudantes com baixa frequência de acesso ou longos períodos de inatividade foram mais propensos a serem classificados como em risco, o que está de

acordo com estudos anteriores sobre engajamento acadêmico em ambientes virtuais (WANG; MOUSAVI, 2022). Por outro lado, cursos mais longos apresentaram uma tendência maior de reprovação, possivelmente devido à maior exigência de manutenção do engajamento ao longo do tempo.

Apesar da forte associação entre essas variáveis e o desfecho de reprovação, seu uso não configura circularidade, uma vez que os dados são capturados antes da conclusão do curso e utilizados de forma preditiva. O modelo é aplicado em tempo real, antes da decisão final sobre o rendimento do aluno, e os logs analisados refletem apenas o comportamento anterior à previsão. Além disso, o gráfico SHAP mostra que o modelo integra múltiplas dimensões comportamentais e temporais, como a semana relativa no curso, o tempo até o primeiro acesso e os padrões de permanência, o que reduz a dependência de uma única variável e aumenta a robustez da predição.

6.3 DISCUSSÃO SOBRE A ESTRATÉGIA DE MODELAGEM: MODELO ÚNICO VS MÚLTIPLOS MODELOS

Além dos ganhos em desempenho observados nas métricas de F1-score, recall e AUC, a adoção do modelo único representa uma mudança importante na estratégia de modelagem preditiva em comparação com a abordagem baseada em múltiplos modelos por semana. A modelagem semanal, utilizada na versão original do projeto LANSE, exigia treinar e manter um modelo distinto para cada semana do curso, o que aumentava a complexidade operacional e dificultava a aplicação em contextos com cursos de diferentes durações. Por outro lado, o modelo único, ao incorporar variáveis como *relative_week* e *total_weeks*, demonstrou maior capacidade de generalização e escalabilidade, pois pode ser aplicado em qualquer ponto do curso com base em informações temporais relativas. Essa mudança de paradigma não apenas melhora a eficiência computacional, como também viabiliza intervenções mais flexíveis, adaptadas a diferentes contextos institucionais e formatos de curso.

Embora a literatura sugira que modelos especializados por semana possam capturar nuances temporais com maior precisão, no presente estudo o modelo único demonstrou desempenho superior. Isso se deve ao fato de que ele foi treinado com um volume de dados maior e mais variado, o que favoreceu sua capacidade de generalização. Além disso, a inclusão de variáveis temporais como *relative_week* permitiu simular o efeito da segmentação semanal, sem a perda de amostras ou o risco de sobreajuste característico dos modelos semanais com menor número de instâncias. Essa abordagem unificada também capturou interações entre diferentes momentos do curso, o que pode explicar sua vantagem em termos de estabilidade e acurácia.

Assim, a adoção do modelo único representa uma solução mais robusta e prática para cenários educacionais com variabilidade de estrutura e duração, explicando

seu desempenho superior mesmo diante de expectativas teóricas contrárias.

6.4 COMPARAÇÃO DE MÉTRICAS COM ESTUDOS EXISTENTES

Nesta seção, comparamos o desempenho do nosso modelo em ambos os experimentos com as métricas relatadas em estudos anteriores. A Tabela 7 resume as métricas de desempenho dos trabalhos relacionados e dos nossos experimentos, incluindo acurácia, recall, F1-score e AUC. Essa comparação enfatiza que nosso estudo focou no F1-score e no recall para alunos reprovados, enquanto os estudos comparados podem não ter fornecido métricas especificamente para essa categoria.

Tabela 7 – Comparação de Métricas de Desempenho entre Estudos

Estudo	F1-Score	Recall	Acurácia	AUC
Exp. 1 (Semana 7) Novo	0.79	0.82	0.87	0.90
Exp. 2 (Semana 7) Único	0.86	0.85	0.90	0.95
(JAIN; SOLANKI, 2019)	0.95	0.95	0.95	-
(CHO; YU; KIM, 2023)	0.84	0.81	0.955	-
(DEWI et al., 2023)	0.97	0.96	0.99	-
(POELLHUBER et al., 2023)	0.33	0.26	0.94	0.62
(SU; LIN; LIU, 2022)	0.88	-	0.83	0.66
(MACARINI et al., 2019)	-	-	-	0.90
(KRÜGER; BRITTO; BARDDAL, 2023)	-	0.72	-	0.86
(QUEIROGA et al., 2020)	-	-	-	0.91
(KOK et al., 2024)	0.81	0.80	0.85	0.87
(NAGY; MOLONTAY, 2024)	-	-	-	0.77
(SULAK; KOKLUB, 2024)	0.76	0.77	0.77	-

Continua na próxima página

Comparação de Métricas de Desempenho entre Estudos (continuação) Fonte: Elaborado pelo autor

Estudo	F1-Score	Recall	Acurácia	AUC
(NDUNAGU et al., 2024)	-	-	0.57	-

Algumas observações-chave podem ser extraídas desta comparação:

- **Acurácia:** A acurácia alcançada no modelo do Experimento 2 foi de 0.90, sendo competitiva com (SU; LIN; LIU, 2022) (0.83), e alinhando-se de perto com (KOK et al., 2024) (0.85), além de se aproximar de (CHO; YU; KIM, 2023) (0.955). Nossa acurácia está ligeiramente abaixo de (DEWI et al., 2023) (0.99), mas ainda apresenta um desempenho robusto em comparação com estudos anteriores.
- **Recall:** O recall no Experimento 2 foi de 0.85, representando uma melhora em relação ao Experimento 1 (0.82) para alunos reprovados. Enquanto (DEWI et al., 2023) (0.96), (CHO; YU; KIM, 2023) (0.81) e (KRÜGER; BRITTO; BARDDAL, 2023) (0.72) relatam valores de recall mais altos, não está claro se esses valores focam especificamente nos alunos reprovados. A capacidade do nosso modelo de detectar alunos em risco, especificamente aqueles propensos à reprovação, destaca sua relevância na previsão de resultados acadêmicos adversos. Nossa taxa de recall supera (CHO; YU; KIM, 2023) (0.81) e (KRÜGER; BRITTO; BARDDAL, 2023) (0.72), aproximando-se de (KOK et al., 2024) (0.80) e (DEWI et al., 2023) (0.96).
- **F1-Score:** O F1-score de 0.86 no Experimento 2 para alunos reprovados representa uma melhoria substancial em relação ao Experimento 1 (0.79). Enquanto (CHO; YU; KIM, 2023) relatou um F1-score de 0.84, (DEWI et al., 2023) obteve 0.97 e (SU; LIN; LIU, 2022) 0.88, não está claro se essas pontuações estão focadas na mesma categoria de alunos que nosso estudo, que enfatiza os reprovados. Nosso F1-score supera (CHO; YU; KIM, 2023) (0.84) e se aproxima de (SU; LIN; LIU, 2022) (0.88), ficando atrás apenas de (DEWI et al., 2023) (0.97). Como diferentes estudos focam em categorias distintas de alunos, essas discrepâncias ressaltam possíveis diferenças na forma como as métricas foram calculadas entre os estudos.
- **AUC:** O AUC no Experimento 2, de 0.95, representa um avanço significativo em relação ao Experimento 1 (0.90). Esse valor está alinhado com (MACARINI et al., 2019) (0.90), (KRÜGER; BRITTO; BARDDAL, 2023) (0.86) e (QUEIROGA et al., 2020) (0.91), confirmando o forte poder discriminatório do modelo na

diferenciação entre alunos em risco e não em risco. Além disso, nosso modelo apresenta desempenho competitivo com (KOK et al., 2024) (0.87), reforçando sua confiabilidade.

Além das métricas de desempenho, é essencial comparar as abordagens algorítmicas e estratégias empregadas pelo nosso modelo e pelos estudos existentes. Estudos anteriores utilizaram modelos como Random Forest e XGBoost, conhecidos por sua robustez, mas que não contavam com a otimização avançada de hiperparâmetros utilizada em nosso estudo (JAIN; SOLANKI, 2019). Ao integrar o NSGA-II, nosso modelo demonstrou maior acurácia, recall e F1-score, especialmente para a classe minoritária de alunos em risco. Isso enfatiza a utilidade do NSGA-II no ajuste fino dos parâmetros do modelo para alcançar um melhor desempenho preditivo em diversas métricas.

Por exemplo, enquanto modelos anteriores apresentaram melhorias ao utilizar técnicas de oversampling, como o SMOTE, aplicamos o ADASYN para focar de maneira mais eficaz na classe minoritária. Isso levou a ganhos significativos na detecção de alunos em risco de evasão. O ADASYN permitiu uma amostragem mais direcionada para a classe minoritária, proporcionando um melhor equilíbrio e aprimorando a capacidade do modelo de generalizar em conjuntos de dados desbalanceados.

Nosso modelo introduz características dinâmicas baseadas no tempo—como a duração relativa do curso e métricas de engajamento semanais—que melhoram significativamente sua capacidade de generalização para cursos com diferentes durações. Estudos anteriores demonstraram a importância dos logs de interação no LMS como preditores de desempenho acadêmico (WANG; MOUSAVI, 2022). No entanto, nosso modelo expande essa abordagem ao incorporar a duração do curso como uma característica dinâmica, ajustando previsões com base no progresso do aluno ao longo do curso. Essa capacidade de lidar com cursos de diferentes durações e comportamentos dos alunos aumenta a robustez do modelo e o diferencia da literatura existente.

Além disso, nosso modelo oferece maior adaptabilidade a diferentes tipos de cursos. Enquanto pesquisas anteriores frequentemente focaram em disciplinas ou tipos de cursos específicos, como STEM, demonstramos que nosso modelo generaliza eficazmente em uma gama mais ampla de disciplinas. Por exemplo, dados socioeconômicos têm sido incorporados para prever taxas de evasão em diversos contextos educacionais (KRÜGER; BRITTO; BARDDAL, 2023). Em contraste, nosso estudo mostrou que o modelo pode ser generalizado para diferentes tipos de cursos, tornando-o aplicável em vários cenários educacionais. Essa flexibilidade destaca a aplicação prática do nosso modelo em diversos contextos institucionais, ampliando seu potencial para enfrentar desafios educacionais em diferentes áreas.

6.5 QUESTÕES DE PESQUISA

Com base nos resultados de nossos experimentos, apoiados por avaliações estatísticas e de distância, fornecemos as seguintes respostas às questões de pesquisa:

Questão 1: *Quão eficazmente os modelos de aprendizado de máquina, treinados com dados limitados de logs estudantis do Moodle, podem generalizar para prever resultados em conjuntos de dados maiores e diversificados, incluindo cursos não vistos anteriormente?* **Resposta.** No *Experimento1*, os modelos apresentaram uma generalização moderada, atingindo uma acurácia de 0.87 na semana 7, mas com um recall menor (0.82) e um F1-score de 0.79 para alunos em risco. Isso indica que os modelos tiveram dificuldades para identificar alunos em risco de maneira eficaz em conjuntos de dados não vistos, sem ajustes adicionais. Testes estatísticos (estatística KS de 0.069, valor- p próximo de zero; valor- p do teste Mann-Whitney U 6.33×10^{-49}) confirmaram diferenças significativas entre as distribuições de treinamento e teste, sugerindo que a homogeneidade dos dados limita a generalização do modelo. Este experimento destaca a importância da qualidade dos dados, especialmente ao trabalhar com um conjunto de dados limitado. Um pequeno conjunto de dados pode não ser suficiente para uma generalização robusta em contextos educacionais, onde os comportamentos dos alunos e os formatos dos cursos podem variar significativamente.

Questão 2: *Quais características nos logs estudantis são mais preditivas do sucesso ou engajamento, e essas características permanecem igualmente importantes em diferentes cursos?* **Resposta.** A análise SHAP e a seleção de características pelo NSGA-II destacaram as principais variáveis preditivas:

- **total_weeks:** Cursos mais longos correlacionam-se com maior probabilidade de evasão.
- **days_with_logs_avg e max_consecutive_days_without_access:** Engajamento regular e ausência de períodos prolongados sem acesso estão associados ao sucesso do aluno.
- **first_log_days_diff e relative_week:** Engajamento precoce e consistente favorecem melhores resultados acadêmicos.

A seleção consistente dessas características nos modelos indica a robustez dessas métricas, embora sua importância relativa possa variar dependendo da estrutura do curso.

Questão 3: *O algoritmo de otimização multiobjetivo NSGA-II pode efetivamente melhorar a acurácia preditiva e a confiabilidade dos modelos de aprendizado de máquina na identificação de alunos em risco em cenários de predição de evasão?* **Resposta.** A implementação do algoritmo NSGA-II produziu bons resultados, especialmente para a classe de alunos em risco, alcançando um F1-score de 0.86 na semana 7. A flexibilidade desse algoritmo genético permitiu uma exploração e exploração eficientes do espaço de busca, possibilitando a otimização simultânea de múltiplos aspectos, como seleção de modelo, seleção de características e ajuste de hiperparâmetros. Sua capacidade multiobjetivo nos permitiu otimizar simultaneamente o macro F1-score e o recall para alunos reprovados, equilibrando o desempenho para alunos em risco com a acurácia geral do modelo. A frente de Pareto 3D na Figura 4 ilustra como o NSGA-II convergiu para configurações ótimas em diferentes métricas, melhorando tanto o recall quanto a estabilidade do modelo.

Questão 4: *Como a acurácia preditiva e a estabilidade de um modelo único treinado com dados de todas as semanas se comparam com modelos separados treinados para cada semana?* **Resposta.** As métricas de desempenho do *Experimento2* revelam que um modelo único treinado com dados de todas as semanas superou os modelos individuais semanais. Notavelmente, o modelo único alcançou pontuações mais altas nas principais métricas, incluindo um F1-score de 0.86, um recall de 0.85 e um AUC de 0.95 na semana 7. Esses resultados sugerem que o modelo único, que tem acesso a todos os dados disponíveis, fornece previsões mais precisas do que um modelo treinado apenas com os dados da semana correspondente.

Implicações para Modelagem Preditiva na Educação: Os resultados indicam que a incorporação de características acumulativas de engajamento, a otimização de métricas-chave via NSGA-II e o uso de características baseadas em semanas melhoram a generalização, a estabilidade e a acurácia preditiva do modelo em diferentes contextos educacionais.

Esses achados podem ser incorporados em sistemas de alerta precoce nas instituições de ensino, integrando os modelos em dashboards analíticos e ferramentas de apoio à decisão docente. A capacidade de identificar estudantes em risco com antecedência suficiente permite que professores, tutores e gestores educacionais intervenham de forma personalizada, promovendo estratégias de retenção, orientação acadêmica e suporte psicopedagógico direcionado.

6.6 LIMITAÇÕES E CONSIDERAÇÕES SOBRE GENERALIZAÇÃO

Embora os resultados obtidos neste estudo sejam promissores, sua generalização para outras instituições e plataformas educacionais deve ser feita com cautela. Os dados utilizados foram extraídos exclusivamente de cursos ofertados por instituições conectadas ao projeto LANSE, que utilizam o ambiente Moodle como AVA. Portanto, as variáveis comportamentais utilizadas, como frequência de acesso e padrões de navegação, são específicas da estrutura de logs dessa plataforma.

Outras plataformas, como Canvas ou Blackboard, podem registrar dados de maneira diferente ou apresentar funcionalidades distintas que influenciam o comportamento dos estudantes. Além disso, fatores institucionais, culturais e pedagógicos podem impactar diretamente os padrões de engajamento e os fatores associados à evasão.

Assim, futuras pesquisas devem validar o modelo proposto em contextos institucionais variados, com dados provenientes de diferentes plataformas e perfis de curso, a fim de verificar sua robustez e ampliar sua aplicabilidade.

6.7 CONTRIBUIÇÕES CIENTÍFICAS

Este trabalho apresenta contribuições relevantes para a área de Learning Analytics, especialmente no contexto da predição de evasão estudantil com dados extraídos do Moodle. Primeiramente, propôs-se e avaliou-se um modelo preditivo único, com codificação temporal e variáveis acumuladas, capaz de generalizar para cursos com diferentes durações e estruturas, superando a limitação de modelos semanais estáticos. Em segundo lugar, utilizou-se a otimização multiobjetivo com o algoritmo NSGA-II para ajustar simultaneamente múltiplas métricas de desempenho, como o F1-score macro e o recall da classe de risco, resultando em maior estabilidade e precisão na identificação de estudantes em risco. Por fim, realizou-se uma comparação empírica entre abordagens baseadas em múltiplos modelos semanais e uma abordagem de modelagem integrada, demonstrando que o modelo único com variáveis temporais não apenas apresenta desempenho superior, mas também maior aplicabilidade prática para implementação em escala institucional.

7 CONSIDERAÇÕES FINAIS

Este capítulo apresenta uma síntese dos principais achados da pesquisa, discutindo os resultados obtidos em relação às questões de pesquisa e objetivos estabelecidos. Também são apontadas direções para investigações futuras, com foco na ampliação do impacto e da aplicabilidade dos modelos propostos. Ao final, é registrada a publicação científica derivada deste trabalho, evidenciando sua relevância acadêmica.

7.1 CONCLUSÃO

Neste estudo, investigamos o impacto de arquiteturas avançadas de modelos e técnicas de balanceamento na acurácia, recall e confiabilidade das previsões para identificação de alunos em risco de evasão. Por meio de dois experimentos, avaliamos como diferentes abordagens influenciam a capacidade dos modelos de aprendizado de máquina em generalizar para novos cursos e conjuntos de dados.

O *Experimento 1* forneceu uma linha de base utilizando um modelo tradicional aplicado a um conjunto de teste *holdout*, explorando os desafios de generalização em cenários com dados limitados de LMS (Questão 1). No entanto, os testes estatísticos revelaram discrepâncias significativas entre as distribuições de treinamento e teste, limitando a aplicabilidade do modelo para novos cursos. Em contraste, o *Experimento 2* introduziu um modelo único, treinado com dados de todas as semanas e características baseadas na progressão do curso. Essa abordagem melhorou substancialmente a capacidade do modelo de se adaptar a diferentes estruturas de cursos, aumentando seu poder preditivo quando aplicado a novos dados.

Além disso, a análise SHAP e a otimização via NSGA-II permitiram identificar as principais características preditivas nos logs estudantis, respondendo à Questão 2. Fatores como *total_weeks*, *days_with_logs_avg* e *max_consecutive_days_without_access* foram consistentemente os mais influentes na previsão de evasão, independentemente do tipo de curso analisado. Isso confirma a importância dessas métricas de engajamento como preditores universais de sucesso ou abandono acadêmico.

A aplicação do NSGA-II mostrou-se uma estratégia de otimização eficaz, respondendo diretamente à Questão 3. Esse algoritmo permitiu a otimização simultânea de múltiplos objetivos, como recall e F1-score para alunos em risco, equilibrando melhor o desempenho do modelo. Os resultados indicam que a utilização de um processo de otimização multiobjetivo contribuiu significativamente para melhorar a precisão e a confiabilidade preditiva.

Os resultados do *Experimento 2* também responderam à Questão 4, eviden-

ciando que um modelo único, treinado com dados de todas as semanas, superou os modelos semanais individuais. Essa abordagem demonstrou maior estabilidade e generalização, permitindo previsões mais confiáveis ao longo do tempo. Esse achado reforça que o uso de características acumulativas de engajamento é uma solução mais escalável para a previsão de evasão estudantil em ambientes virtuais.

Do ponto de vista científico, este trabalho avança o estado da arte em Learning Analytics ao combinar modelagem temporal, explicabilidade (via SHAP), balanceamento de classes e otimização multiobjetivo (NSGA-II) em um contexto real de aplicação. A proposta de um modelo preditivo único, treinado com dados de múltiplas semanas e validado em cenários educacionais diversos, representa uma inovação metodológica com alto potencial de generalização. Esses achados fornecem evidências empíricas para a viabilidade de modelos preditivos robustos e adaptáveis, contribuindo diretamente para a literatura sobre evasão escolar, sistemas de alerta precoce e apoio à decisão em ambientes virtuais de aprendizagem.

Com esses avanços, este estudo contribui para o aprimoramento de técnicas de previsão de evasão estudantil, fornecendo uma base sólida para futuras pesquisas e aplicações em larga escala em ambientes de aprendizagem online.

7.2 TRABALHOS FUTUROS

Trabalhos futuros podem expandir este estudo de diversas maneiras. Primeiramente, testar o modelo em conjuntos de dados maiores e mais diversos ajudaria a validar ainda mais sua robustez e capacidade de generalização. Além disso, aprofundar a explicabilidade do modelo com técnicas avançadas de XAI poderia fornecer insights mais detalhados sobre os fatores que impulsionam a evasão, tornando as previsões mais acionáveis para educadores.

Outra direção relevante seria investigar o desempenho do modelo em diferentes áreas do conhecimento, como STEM, ciências sociais e humanidades, a fim de verificar se abordagens especializadas poderiam melhorar ainda mais a precisão preditiva. Também se destaca o potencial de incorporar dados de cadastro dos estudantes — como idade, turno, modalidade e trajetória educacional — e variáveis socioeconômicas, visando enriquecer os modelos com fatores contextuais que influenciam o comportamento acadêmico.

Por fim, recomenda-se a realização de estudos voltados à aplicação prática dos resultados em ambientes educacionais reais. Investigações futuras poderiam integrar o modelo a sistemas institucionais e avaliar seu uso como ferramenta de apoio à decisão pedagógica, verificando como as previsões influenciam intervenções, estratégias de orientação e políticas de retenção estudantil. Essa transição do modelo

preditivo para a ação educacional representa um passo fundamental para garantir que os avanços tecnológicos se traduzam em melhorias concretas no processo de ensino-aprendizagem.

7.3 ARTIGO PUBLICADO

Rebelo Marcolino, M., Reis Porto, T., Thompsen Primo, T. et al. Student dropout prediction through machine learning optimization: insights from moodle log data. Sci Rep 15, 9840 (2025). <https://doi.org/10.1038/s41598-025-93918-1>

Referências

AGUIAR, E. et al. Engagement vs performance: Using electronic portfolios to predict student outcomes. *Journal of Learning Analytics*, v. 5, n. 3, p. 93–108, 2018. Citado 2 vezes nas páginas 18 e 30.

ALGHAMDI, S.; SOH, B.; LI, A. A comprehensive review of dropout prediction methods based on multivariate analysed features of mooc platforms. *Multimodal Technologies and Interaction*, v. 9, n. 1, 2025. ISSN 2414-4088. Disponível em: <<https://www.mdpi.com/2414-4088/9/1/3>>. Citado na página 28.

BENNEDSEN, J.; CASPERSEN, M. E. Failure rates in introductory programming. *SIGCSE Bull.*, Association for Computing Machinery, New York, NY, USA, v. 39, n. 2, p. 32–36, jun. 2007. ISSN 0097-8418. Disponível em: <<https://doi.org/10.1145/1272848.1272879>>. Citado 2 vezes nas páginas 16 e 17.

BOTEJU, G.; TANG, L.; BROWN, M. S. Support vector machine for predicting student dropout under different normalization methods. *Excerpts from "Boteju et al. - 2024 - Support Vector Machine for Predicting Student Drop.pdf"*, 2024. Citado na página 34.

BOUIHI, B. et al. Prediction of higher education student dropout based on regularized regression models. *Engineering, Technology & Applied Science Research*, v. 14, n. 6, p. 17811–17815, dez. 2024. ISSN 1792-8036, 2241-4487. Citado 2 vezes nas páginas 23 e 31.

CARUANA, R.; NICULESCU-MIZIL, A. An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, p. 161–168, 2006. Citado 2 vezes nas páginas 25 e 26.

CECHINEL, C. et al. Learning analytics para moodle em uma arquitetura na nuvem: uma solução escalável para predição de risco acadêmico. In: *Anais do II Workshop de Aplicações Práticas de Learning Analytics em Instituições de Ensino no Brasil*. Porto Alegre, RS, Brasil: SBC, 2023. p. 128–137. ISSN 0000-0000. Disponível em: <<https://sol.sbc.org.br/index.php/wapla/article/view/26136>>. Citado 3 vezes nas páginas 23, 57 e 58.

CECHINEL, C. et al. Lanse: A cloud-powered learning analytics platform for the automated identification of students at risk in learning management systems. In: *AIED 2024 Workshops, CCIS 2150*. [S.l.]: Springer, 2024. Citado 6 vezes nas páginas 16, 45, 47, 57, 58 e 59.

CHEAH, C. S. Factors contributing to the difficulties in teaching and learning of computer programming: A literature review. *Contemporary Educational Technology*, v. 12, n. 2, p. ep272, 2020. Citado 3 vezes nas páginas 16, 17 e 18.

CHEN, J. et al. Generalizing the auc-roc for unbalanced data, early retrieval and link prediction evaluation. jul. 2024. Disponível em: <<https://www.preprints.org/manuscript/202209.0277/v2>>. Citado 2 vezes nas páginas 35 e 36.

CHO, C.-H.; YU, Y. W.; KIM, H. G. A study on dropout prediction for university students using machine learning. *Applied Sciences*, MDPI, v. 13, n. 21, p. 12004, 2023. Disponível em: <<https://www.mdpi.com/2076-3417/13/21/12004/pdf?version=1698999402>>. Citado 5 vezes nas páginas 34, 41, 43, 71 e 72.

CU, N. G. et al. Increment of academic performance prediction of at-risk student by dealing with data imbalance problem. *Applied Computational Intelligence and Soft Computing*, v. 2024, n. 1, p. 4795606, jan. 2024. ISSN 1687-9724, 1687-9732. Citado 2 vezes nas páginas 36 e 37.

DEB, K. et al. A fast and elitist multiobjective genetic algorithm: Nsga-ii. *IEEE Transactions on Evolutionary Computation*, v. 6, n. 2, p. 182–197, 2002. Citado 2 vezes nas páginas 38 e 47.

DEB, S. et al. Predicting student dropout: A machine learning approach. *Excerpts from "Deb et al. - 2024 - Predicting Student Dropout A Machine Learning App.pdf"*, 2024. Citado 3 vezes nas páginas 31, 32 e 34.

DETONI, D. et al. Learning to identify at-risk students in distance education using interaction counts. *Revista de Informática Teórica e Aplicada*, v. 23, n. 2, p. 124, dez. 2016. ISSN 21752745, 01034308. Citado na página 57.

DEWI, M. A. et al. Machine learning algorithms for early predicting dropout student online learning. In: *2023 9th International Conference on Computing, Engineering, and Design (ICCED)*. [S.l.]: IEEE, 2023. p. 1–6. Citado 5 vezes nas páginas 28, 30, 43, 71 e 72.

DEY, I.; PRATAP, V. A comparative study of smote, borderline-smote, and adasyn oversampling techniques using different classifiers. In: *2023 3rd International Conference on Smart Data Intelligence (ICSMDI)*. Trichy, India: IEEE, 2023. p. 294–302. ISBN 978-1-66546-487-1. Disponível em: <<https://ieeexplore.ieee.org/document/10127891/>>. Citado na página 37.

DONCHE, V. Learning analytics voor hoger onderwijs: inleiding op het themanummer. *TIJDSCHRIFT VOOR HOGER ONDERWIJS*, 2023. Citado na página 21.

DOROGUSH, A. V.; ERSHOV, V.; GULIN, A. Catboost: gradient boosting with categorical features support. *ArXiv*, 2018. Disponível em: <<https://arxiv.org/abs/1810.11363>>. Citado na página 48.

DOUGIAMAS, M.; TAYLOR, P. C. Moodle: Using learning communities to create an open source course management system. 2003. Citado na página 23.

DOYAOEN, S. I. et al. Analysis of education student retention rates: Basis for policy formation. *International Journal of Social Science and Education Research Studies*, v. 4, n. 12, p. 1322–1328, 2024. Citado na página 24.

DUNICAN, E. Making the analogy: Alternative delivery techniques for first year programming courses. In: *Proceedings of the 14th annual workshop of the psychology of programming interest group*. [S.l.: s.n.], 2002. v. 2002, p. 89–99. Citado 2 vezes nas páginas 16 e 17.

FERGUSON, R. Learning analytics: drivers, developments and challenges. *International Journal of Technology Enhanced Learning*, v. 4, n. 5/6, p. 304, 2012. ISSN 1753-5255, 1753-5263. Citado 2 vezes nas páginas 21 e 22.

FREDRICKS, J. A.; BLUMENFELD, P. C.; PARIS, A. H. School engagement: Potential of the concept, state of the evidence. *Review of Educational Research*, v. 74, n. 1, p. 59–109, mar. 2004. ISSN 0034-6543, 1935-1046. Citado na página 16.

GANDHI, Y.; PATEL, A.; JAGTAP, V. Exploring strategies and algorithms for tackling class imbalance across varied data types. In: *2024 International Conference on Intelligent Computing and Emerging Communication Technologies (ICEC)*. Guntur, India: IEEE, 2024. p. 1–5. ISBN 9798331508432. Disponível em: <https://ieeexplore.ieee.org/document/10837043/>. Citado na página 37.

GAŠEVIĆ, D.; DAWSON, S.; SIEMENS, G. Let's not forget: Learning analytics are about learning. *TechTrends*, v. 59, n. 1, p. 64–71, jan. 2015. ISSN 8756-3894, 1559-7075. Citado 2 vezes nas páginas 22 e 23.

GOLDMAN, M.; KAPLAN, D. M. Comparing distributions by multiple testing across quantiles or cdf values. *Journal of Econometrics*, v. 206, n. 1, p. 143–166, set. 2018. ISSN 03044076. ArXiv:1708.04658 [math]. Citado na página 40.

GUPTA, U.; ZHENG, R. Z. Cognitive load in solving mathematics problems: Validating the role of motivation and the interaction among prior knowledge, worked examples, and task difficulty. *European Journal of STEM Education*, v. 5, n. 1, p. 05, 2020. Citado 2 vezes nas páginas 16 e 18.

HAO, J. et al. An improved nsga-ii algorithm based on adaptive weighting and searching strategy. *Applied Sciences*, v. 12, n. 22, p. 11573, nov. 2022. ISSN 2076-3417. Citado 2 vezes nas páginas 37 e 38.

HE, H. et al. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: *IEEE International Joint Conference on Neural Networks*. [S.l.: s.n.], 2008. p. 1322–1328. Citado 2 vezes nas páginas 48 e 53.

HUSAIN, G. et al. Smote vs. smoteenn: A study on the performance of resampling algorithms for addressing class imbalance in regression models. *Algorithms*, v. 18, n. 1, p. 37, jan. 2025. ISSN 1999-4893. Citado na página 37.

IDRISSI, A. *Modern Artificial Intelligence and Data Science: Tools, Techniques and Systems*. Cham: Springer Nature Switzerland, 2023. v. 1102. (Studies in Computational Intelligence, v. 1102). ISBN 978-3-031-33308-8. Citado na página 27.

JAIN, A.; SOLANKI, S. An efficient approach for multiclass student performance prediction based upon machine learning. In: *2019 International Conference on Communication and Electronics Systems (ICCES)*. [S.l.: s.n.], 2019. p. 1457–1462. Citado 5 vezes nas páginas 18, 28, 42, 71 e 73.

JAMES, G. et al. *An Introduction to Statistical Learning: with Applications in R*. 2. ed. [S.l.]: Springer, 2017. (Springer Texts in Statistics). ISBN 978-1461471370. Citado 2 vezes nas páginas 25 e 26.

JAMES, G. M. et al. *An introduction to statistical learning: with applications in R*. Corrected at 6th printing 2015. New York: Springer Science+Business Media, 2013. (Springer texts in statistics). ISBN 978-1-4614-7138-7. Citado 2 vezes nas páginas 48 e 54.

JENKINS, T. On the difficulty of learning to program. In: CITESEER. *Proceedings of the 3rd Annual Conference of the LTSN Centre for Information and Computer Sciences*. [S.l.], 2002. v. 4, p. 53–58. Citado 2 vezes nas páginas 16 e 17.

KEMDA, L. E.; MURRAY, M. Longitudinal analysis of academic dropout in higher education: Evidence from a south african university. *International Journal of Higher Education*, v. 10, n. 6, p. 153–167, 2021. Citado na página 24.

KETUT, I. et al. Optimizing dropout prediction in university using oversampling techniques for imbalanced datasets. *International Journal of Information and Education Technology*, 2024. Disponível em: <<https://api.semanticscholar.org/CorpusID:272241050>>. Citado 2 vezes nas páginas 32 e 33.

KIM, S.; YOO, E.; KIM, S. Why do students drop out? university dropout prediction and associated factor analysis using machine learning techniques. arXiv, n. arXiv:2310.10987, out. 2023. ArXiv:2310.10987 [cs]. Disponível em: <<http://arxiv.org/abs/2310.10987>>. Citado na página 35.

KOK, C. L. et al. A novel predictive modeling for student attrition utilizing machine learning and sustainable big data analytics. *Applied Sciences*, v. 14, n. 9633, 2024. Disponível em: <<https://www.mdpi.com/10.3390/app14219633>>. Citado 5 vezes nas páginas 42, 45, 71, 72 e 73.

KOTSIANTIS, S. B. Supervised machine learning: A review of classification techniques. *Informatica*, v. 31, n. 3, p. 249–268, 2007. Disponível em: <<http://www.informatica.si/index.php/informatica/article/view/148>>. Citado na página 26.

KRÜGER, J.; BRITTO, A.; BARDDAL, J. An explainable machine learning approach for student dropout prediction. *Expert Systems with Applications*, 2023. Citado 6 vezes nas páginas 18, 42, 44, 71, 72 e 73.

KUH, G. et al. *Using Evidence of Student Learning to Improve Higher Education*. United States: Jossey-Bass, 2014. ISBN 978-1-118-90366-7. Citado na página 18.

KUMAR, N. M. et al. Ai-powered early detection and prevention system for student dropout risk. *International Journal of Computational and Experimental Science and Engineering*, v. 11, n. 1, Jan. 2025. Disponível em: <<https://www.ijcesen.com/index.php/ijcesen/article/view/839>>. Citado na página 28.

LONG, P.; SIEMENS, G. Penetrating the fog: Analytics in learning and education. *EDUCAUSE review*, v. 46, n. 5, p. 31–40, 2011. Citado na página 22.

MACARINI, L. A. B. et al. Predicting students success in blended learning—evaluating different interactions inside learning management systems. *Applied Sciences*, MDPI, v. 9, n. 24, p. 5523, 2019. Citado 7 vezes nas páginas 18, 28, 41, 42, 57, 71 e 72.

MANN, H. B.; WHITNEY, D. R. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, v. 18, n. 1, p. 50–60, mar. 1947. ISSN 0003-4851. Citado na página 40.

MASSEY, F. J. The kolmogorov-smirnov test for goodness of fit. *Journal of the American Statistical Association*, v. 46, n. 253, p. 68–78, 1951. Citado na página 40.

MITCHELL, T. M. *Machine Learning*. 1. ed. [S.l.]: McGraw-Hill, 2013. ISBN 9780070428072. Citado 2 vezes nas páginas 25 e 26.

MONTUORI, L. et al. Learning analytics as data driven decision making in high education: a case study. In: *Proceedings INNODOCT/22. International Conference on Innovation, Documentation and Education*. Editorial Universitat Politècnica de València, 2023. p. 83–89. ISBN 978-84-13-96036-4. Disponível em: <http://ocs.editorial.upv.es/index.php/INNODOCT/INN2022/paper/view/15750>. Citado 2 vezes nas páginas 21 e 22.

Moodle Project. *Moodle – Open-source learning platform*. 2024. Disponível em: <https://moodle.org>. Citado na página 23.

MOTZ, B. et al. A lak of direction: Misalignment between the goals of learning analytics and its research scholarship. abr. 2022. Disponível em: <https://osf.io/bzahw>. Citado na página 21.

NACHAR, N. The mann-whitney u: A test for assessing whether two independent samples come from the same distribution. *Tutorials in Quantitative Methods for Psychology*, v. 4, p. 13–20, 03 2008. Citado na página 53.

NAGY, M.; MOLONTAY, R. Interpretable dropout prediction: Towards xai-based personalized intervention. *International Journal of Artificial Intelligence in Education*, v. 34, n. 2, p. 274–300, June 2024. ISSN 1560-4306. Disponível em: <https://doi.org/10.1007/s40593-023-00331-8>. Citado 4 vezes nas páginas 27, 29, 44 e 71.

NDUNAGU, J. N. et al. Deep learning for predicting attrition rate in open and distance learning (odl) institutions. *Computers*, v. 13, n. 9, 2024. ISSN 2073-431X. Disponível em: <https://www.mdpi.com/2073-431X/13/9/229>. Citado 3 vezes nas páginas 28, 45 e 72.

NGUYEN, C. D.; CARLIN, J. B.; LEE, K. J. Diagnosing problems with imputation models using the kolmogorov-smirnov test: a simulation study. *BMC Medical Research Methodology*, v. 13, p. 144, 2013. Disponível em: <https://doi.org/10.1186/1471-2288-13-144>. Citado na página 53.

NGUYEN, T. H. et al. A multivariate analysis of the early dropout using classical machine learning and local interpretable model-agnostic explanations. *CTU Journal of Innovation and Sustainable Development*, v. 16, n. Special issue: ISDS, p. 98–106, October 2024. ISSN 2815-6412, 2588-1418. Citado 3 vezes nas páginas 27, 28 e 29.

NIKOLAIDIS, P. et al. Predicting student attrition in higher education through the determinants of learning progress: A structural equation modelling approach. *Sustainability*, v. 14, n. 20, p. 13584, 2022. Citado na página 24.

OBAIDO, G. et al. Analysis of attrition studies within the computer sciences. *IEEE Access*, v. 11, p. 53736–53751, 2023. Citado na página 24.

OHKAWAUCHI, T.; TANAKA, E. A study on real-time prediction of course dropout students using lms logs. In: *2024 16th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*. [S.l.: s.n.], 2024. p. 224–230. Citado na página 29.

OQAIDI, K.; AOUHASSI, S.; MANSOURI, K. Towards a students' dropout prediction model in higher education institutions using machine learning algorithms. *International Journal of Emerging Technologies in Learning (iJET)*, International Association of Online Engineering (IAOE), v. 17, n. 18, p. 109–125, 2022. Citado 2 vezes nas páginas 30 e 43.

OSEMWEGIE, E. E.; AMADIN, F. I.; UDUCHI, O. M. Student dropout prediction using machine learning. *FUDMA JOURNAL OF SCIENCES*, v. 7, n. 6, p. 347–353, dez. 2023. ISSN 2616-1370, 2645-2944. Citado na página 34.

PAN, F. et al. Achieving optimal trade-off for student dropout prediction with multi-objective reinforcement learning. *PeerJ Computer Science*, v. 10, p. e2034, abr. 2024. ISSN 2376-5992. Citado na página 37.

PAPAMITSIOU, Z.; ECONOMIDES, A. Learning analytics and educational data mining in practice: A systematic literature review of empirical evidence. *Educational Technology & Society*, v. 17, p. 49–64, 10 2014. Citado 3 vezes nas páginas 21, 22 e 23.

PEDREGOSA, F. et al. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, v. 12, 01 2012. Citado 2 vezes nas páginas 48 e 53.

POELLHUBER, L.-V. et al. Cluster-based performance of student dropout prediction as a solution for large scale models in a moodle lms. In: *Proceedings of the 13th International Learning Analytics and Knowledge Conference (LAK 2023)*. [S.l.]: ACM, 2023. p. 592–598. Citado 5 vezes nas páginas 23, 30, 43, 45 e 71.

PROKHORENKOVA, L. et al. Catboost: unbiased boosting with categorical features. arXiv, n. arXiv:1706.09516, jan. 2019. ArXiv:1706.09516 [cs]. Disponível em: <http://arxiv.org/abs/1706.09516>. Citado na página 29.

PSATHAS, G.; CHATZIDAKI, T. K.; DEMETRIADIS, S. N. Predictive modeling of student dropout in moocs and self-regulated learning. *Computers*, v. 12, n. 10, p. 194, set. 2023. ISSN 2073-431X. Citado 2 vezes nas páginas 32 e 33.

PUTRA, M. R. P.; UTAMI, E. Comparative analysis of hybrid model performance using stacking and blending techniques for student drop-out prediction in mooc. *JURNAL RESTI (Rekayasa Sistem dan Teknologi Informasi)*, v. 8, n. 3, p. 346 – 354, 2024. Citado 3 vezes nas páginas 31, 32 e 34.

QUEIROGA, E. M. et al. A learning analytics approach to identify students at risk of dropout: A case study with a technical distance education course. *Applied Sciences*, MDPI, v. 10, n. 11, p. 3998, 2020. Citado 7 vezes nas páginas 16, 18, 28, 41, 44, 71 e 72.

RIDWAN, A.; PRIYATNO, A. M. Predict students' dropout and academic success with xgboost. *Journal of Education and Computer Applications*, v. 1, n. 2, p. 1–8, dez. 2024. ISSN 3047-8308. Citado na página 34.

SABBIR, W. et al. Improving predictive analytics for student dropout: A comprehensive analysis and model evaluation. In: *2024 11th International Conference on Computing for Sustainable Global Development (INDIACom)*. New Delhi, India: IEEE, 2024. p. 951–956. ISBN 978-93-80544-51-9. Disponível em: <https://ieeexplore.ieee.org/document/10498520/>. Citado na página 35.

SANTOS, F. J. D. et al. Learning analytics: transformando dados em estratégias pedagógicas. *OBSERVATÓRIO DE LA ECONOMÍA LATINOAMERICANA*, v. 22, n. 4, p. e4093, abr. 2024. ISSN 1696-8352, 1696-8352. Citado 3 vezes nas páginas 21, 22 e 23.

SANTOS, R.; HENRIQUES, R. Predicting student performance from moodle logs in higher education: A course-agnostic approach. In: *Education and New Developments 2023 - Volume 2*. [S.l.: s.n.], 2023. Citado 2 vezes nas páginas 23 e 45.

SARKER, I. H. Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science*, Springer, v. 2, n. 3, p. 160, 2021. Citado na página 43.

SHOHAG, S.; BAKAUL, M. A machine learning approach to detect student dropout at university. *International Journal of Advanced Trends in Computer Science and Engineering*, v. 10, p. 3101–3107, 12 2021. Citado 2 vezes nas páginas 16 e 18.

SHYNARBEK, N. et al. Forecasting dropout in university based on students' background profile data through automated machine learning approach. In: *2022 IEEE International Conference on Smart Information Systems and Technologies (SIST)*. IEEE, 2022. p. 1–4. Disponível em: <https://dx.doi.org/10.1109/SIST54437.2022.9945715>. Citado 2 vezes nas páginas 16 e 17.

SU, Y.-S.; LIN, Y.-D.; LIU, T.-Q. Applying machine learning technologies to explore students' learning features and performance prediction. *Frontiers in Neuroscience*, 2022. Citado 3 vezes nas páginas 42, 71 e 72.

SULAK, S. A.; KOKLU, N. Predicting student dropout using machine learning algorithms. *Intelligent Methods in Engineering Sciences*, p. 3, jan. 2025. ISSN 2979-9236. Citado na página 36.

SULAK, S. A.; KOKLUB, N. Predicting student dropout using machine learning algorithms. *Intelligent Methods In Engineering Sciences*, v. 3, n. 3, p. 91–98, September 2024. Disponível em: <https://imiens.org/index.php/imiens/article/view/62>. Citado 3 vezes nas páginas 28, 44 e 71.

TRUCHON, J.-F.; BAYLY, C. I. Evaluating virtual screening methods: Good and bad metrics for the “early recognition” problem. *Journal of Chemical Information and Modeling*, v. 47, n. 2, p. 488–508, mar. 2007. ISSN 1549-9596, 1549-960X. Citado na página 36.

VAARMA, M.; LI, H. Predicting student dropouts with machine learning: An empirical study in finnish higher education. *Technology in Society*, v. 76, p. 102474, mar. 2024. ISSN 0160791X. Citado 2 vezes nas páginas 23 e 45.

VERMA, S.; PANT, M.; SNASEL, V. A comprehensive review on nsga-ii for multi-objective combinatorial optimization problems. *IEEE Access*, v. 9, p. 57757–57791, 2021. Citado na página 39.

VIHAVAINEN, A.; PAKSULA, M.; LUUKKAINEN, M. Extreme apprenticeship method in teaching programming for beginners. In: *Proceedings of the 42nd ACM Technical Symposium on Computer Science Education*. New York, NY, USA: Association for Computing Machinery, 2011. (SIGCSE '11), p. 93–98. ISBN 9781450305006. Disponível em: <<https://doi.org/10.1145/1953163.1953196>>. Citado 2 vezes nas páginas 16 e 17.

VILLAR, A.; ANDRADE, C. R. V. D. Supervised machine learning algorithms for predicting student dropout and academic success: A comparative study. *Discover Artificial Intelligence*, v. 4, n. 1, p. 2, January 2024. ISSN 2731-0809. Citado na página 27.

WANG, Q.; MOUSAVI, A. Which log variables significantly predict academic achievement? a systematic review and meta-analysis. *British Journal of Educational Technology*, v. 54, 10 2022. Citado 4 vezes nas páginas 30, 44, 70 e 73.

WANG, Y.; LAI, Y.; HUANG, X. Innovations in online learning analytics: A review of recent research and emerging trends. *IEEE Access*, v. 12, p. 166761–166775, 2024. ISSN 2169-3536. Citado 2 vezes nas páginas 21 e 23.

WATSON, C.; LI, F. W. Failure rates in introductory programming revisited. In: *Proceedings of the 2014 conference on Innovation & technology in computer science education - ITiCSE '14*. Uppsala, Sweden: ACM Press, 2014. p. 39–44. ISBN 978-1-4503-2833-3. Disponível em: <<http://dl.acm.org/citation.cfm?doid=2591708.2591749>>. Citado 2 vezes nas páginas 16 e 17.

WIJAYANINGRUM, V. N.; KIRANA, A. P.; PUTRI, I. K. Student academic performance prediction framework with feature selection and imbalanced data handling. *Jurnal Ilmiah Kursor*, v. 12, n. 3, p. 123–134, maio 2024. ISSN 2301-6914, 0216-0544. Citado na página 35.

YUSOFF, Y.; NGADIMAN, M.; ZAIN, A. Overview of nsga-ii for optimizing machining process parameters. *Procedia Engineering*, v. 15, 08 2011. Citado na página 39.

ZHENG, W.; DOERR, B. Runtime analysis for the nsga-ii: Proving, quantifying, and explaining the inefficiency for many objectives. *IEEE Transactions on Evolutionary Computation*, v. 28, n. 5, p. 1442–1454, out. 2024. ISSN 1089-778X, 1089-778X, 1941-0026. Citado na página 38.

ZHOU, F.; AGARWAL, N. Student performance prediction based on decision trees. *Quest Journals Journal of Research in Applied Mathematics*, v. 10, p. 114–120, 2024. Citado 3 vezes nas páginas 31, 32 e 33.

ÖZÇOMAK, M. et al. Comparison of the powers of the kolmogorov-smirnov two-sample test and the mann-whitney test for different kurtosis and skewness coefficients using the monte carlo simulation method. *Journal of Statistical and Econometric Methods*, v. 2, p. 81–98, 01 2013. Citado na página 40.