Análise de Desempenho Acadêmico Utilizando Machine Learning

Márcio Éric Lamêgo Valente[1] **, José Santo de Moura Neto**[2]

[1] [2]Curso de Engenharia de Software
Instituto de Ciências Exatas e Tecnologia (ICET)
Universidade Federal do Amazonas (UFAM)
Itacoatiara, AM, Brasil

marcio.valente@ufam.edu.br[1], jose-santo.moura@ufam.edu.br[2]

Abstract. *Technological advances in recent decades have revolutionized several areas of knowledge, and Artificial Intelligence (AI) stands out as one of the main driving forces of this transformation. Within this field, Machine Learning (ML) emerges as one of the most promising subareas, allowing systems to learn from data, identify patterns, and make predictions. In the educational context, ML has been applied to improve the quality of teaching and the student experience, such as predicting academic performance and identifying students at risk of dropping out. This article proposes the development of predictive models using ML to analyze academic performance in universities. By identifying patterns and predicting students' future success or difficulties, the project seeks to provide a basis for personalized interventions, improving the effectiveness of educational strategies and increasing academic success rates. The application of ML in education not only improves student retention but also contributes to the optimization of institutional resources.*

Resumo. *O avanço tecnológico nas últimas décadas tem revolucionado diversas áreas do conhecimento, e a Inteligência Artificial (IA) se destaca como uma das principais forças motrizes dessa transformação. Dentro desse campo, o Machine Learning (ML) emerge como uma das subáreas mais promissoras, permitindo que sistemas aprendam com dados, identifiquem padrões e façam previsões. No contexto educacional, o ML tem sido aplicado para melhorar a qualidade do ensino e a experiência dos estudantes, como na predição do desempenho acadêmico e na identificação de estudantes em risco de evasão. Este artigo propõe o desenvolvimento de modelos preditivos utilizando ML para analisar o desempenho acadêmico em universidades. Ao identificar padrões e prever o sucesso ou dificuldades futuras dos estudantes, o projeto busca fornecer uma base para intervenções personalizadas, aprimorando a eficácia das estratégias educacionais e aumentando as taxas de sucesso acadêmico. A aplicação de ML na educação não apenas melhora a retenção de alunos, mas também contribui para a otimização de recursos institucionais.*

1. Introdução

1.1. Tema

O avanço tecnológico nas últimas décadas tem revolucionado diversas áreas do conhecimento, e a Inteligência Artificial (IA) se destaca como uma das principais forças mo-

trizes dessa transformação. A IA, definida como a capacidade de máquinas simularem inteligência humana para realizar tarefas complexas, tem sido aplicada em setores como saúde, finanças, transporte e, mais recentemente, na educação [Russell and Norvig 2016].

Dentro desse campo, o Machine Learning (ML) emerge como uma das subáreas mais promissoras, permitindo que sistemas aprendam com dados, identifiquem padrões e façam previsões sem serem explicitamente programados para cada tarefa [Mitchell 1997]. O ML é baseado em algoritmos que podem ser supervisionados, não supervisionados ou de reforço, dependendo da natureza do problema a ser resolvido. Em problemas supervisionados, o modelo é treinado com dados rotulados, enquanto em problemas não supervisionados, o sistema identifica padrões sem rótulos prévios. Já o aprendizado por reforço envolve a interação do sistema com um ambiente, onde ele aprende a tomar decisões com base em recompensas e penalidades [Goodfellow et al. 2016]. Essas técnicas têm sido amplamente utilizadas para prever comportamentos, otimizar processos e tomar decisões baseadas em dados, tornando-se uma ferramenta essencial para organizações que buscam inovação e eficiência.

1.2. Problema

No contexto educacional, o ML tem sido aplicado para melhorar a qualidade do ensino e a experiência dos estudantes. Por exemplo, algoritmos de ML podem prever o desempenho acadêmico, identificar estudantes em risco de evasão e personalizar o aprendizado de acordo com as necessidades individuais dos alunos [Filho et al. 2020]. [Kantorski et al. 2016] demonstraram a eficácia do ML ao prever a evasão em uma instituição pública de ensino superior, alcançando uma acurácia impressionante de 98%. Além disso, [Bastos et al. 2024] utilizaram técnicas de Mineração de Dados Educacionais (EDM) para analisar fatores críticos de evasão na Universidade Federal do Sul da Bahia (UFSB), destacando a importância do tempo de permanência e do coeficiente de rendimento (CR) como preditores significativos.

Este artigo propõe o desenvolvimento de modelos preditivos utilizando ML para analisar o desempenho acadêmico em universidades. Ao identificar padrões e prever o sucesso ou dificuldades futuras dos estudantes, o projeto busca fornecer uma base para intervenções personalizadas, aprimorando a eficácia das estratégias educacionais e aumentando as taxas de sucesso acadêmico. A aplicação de ML na educação não apenas melhora a retenção de alunos, mas também contribui para a otimização de recursos institucionais, permitindo que universidades direcionem esforços de forma mais eficiente e estratégica.

1.3. Objetivos

1.3.1. Objetivo Geral

Desenvolver um modelo preditivo utilizando técnicas de Machine Learning para analisar o desempenho acadêmico dos alunos, com o intuito de identificar padrões de desempenho, prever resultados futuros e propor intervenções personalizadas para melhorar o rendimento dos estudantes.

1.3.2. Objetivos Específicos

- Pesquisar algoritmos e ferramentas de Machine Learning adequados para a análise de dados acadêmicos.
- Avaliar a eficácia de diferentes modelos preditivos de Machine Learning aplicados aos dados coletados.
- Propor intervenções pedagógicas baseadas nos resultados do modelo preditivo para melhorar o desempenho acadêmico dos alunos.

1.4. Justificativa

A evasão e o baixo desempenho acadêmico são desafios significativos enfrentados pelas instituições de ensino superior no Brasil. Com o crescimento do volume de dados educacionais, surgem oportunidades para utilizar técnicas de Machine Learning na análise e previsão do rendimento dos estudantes. Essas ferramentas permitem identificar padrões de comportamento e desempenho, possibilitando ações preventivas e intervenções pedagógicas mais eficazes. Assim, este projeto se justifica pela necessidade de soluções inovadoras que contribuam para a melhoria da qualidade do ensino e para o aumento da permanência e do sucesso acadêmico dos alunos.

2. Referencial Teórico

A aplicação de técnicas de Inteligência Artificial (IA), especialmente o Machine Learning (ML), na área da educação tem ganhado destaque nos últimos anos. ML pode ser definido como um conjunto de algoritmos que permitem que sistemas aprendam a partir de dados históricos para realizar previsões ou tomar decisões com o mínimo de intervenção humana [Mitchell 1997, Goodfellow et al. 2016]. Dentre as abordagens mais comuns no contexto educacional estão os algoritmos de classificação (como árvores de decisão, SVM e redes neurais) e de regressão.

A área conhecida como Mineração de Dados Educacionais (Educational Data Mining – EDM) é responsável por investigar padrões ocultos em grandes volumes de dados educacionais com o objetivo de melhorar os processos de ensino e aprendizagem [Filho et al. 2020]. Diversos estudos demonstram a eficácia de modelos preditivos na identificação de estudantes em risco de evasão [Kantorski et al. 2016, Bastos et al. 2024], na previsão de desempenho acadêmico [Fernandez-García et al. 2021] e na proposição de intervenções pedagógicas mais eficazes [Najera et al. 2018].

A literatura também destaca a importância da integração de variáveis socioeconômicas, comportamentais e acadêmicas na construção de modelos robustos de predição [Matz et al. 2023]. Tais variáveis ampliam a compreensão do contexto individual do aluno, promovendo ações educacionais mais precisas e inclusivas. Em um estudo realizado na Universidade Federal do Sul da Bahia, por exemplo, o tempo de permanência no curso e o coeficiente de rendimento foram destacados como fatores preditivos significativos de evasão [Bastos et al. 2024].

Adicionalmente, iniciativas como as de [Bakker et al. 2023] e [Cabrera et al. 2023] demonstram que modelos preditivos baseados em IA podem alcançar altos níveis de precisão na previsão de desempenho acadêmico, mesmo em populações com características diversas, como estudantes autistas ou de educação a distância.

Por fim, a combinação de revisões sistemáticas da literatura [Kitchenham and Charters 2007, Petticrew and Roberts 2008] e métodos empíricos baseados em dados reais consolida o uso do ML como ferramenta científica na educação, promovendo um ciclo contínuo de melhoria baseado em evidências.

3. Metodologia

A metodologia será dividida em quatro etapas principais:

3.1. Etapa 1 – Pesquisa de Algoritmos de ML para Análise Acadêmica

A primeira etapa envolverá a revisão das técnicas e ferramentas de Machine Learning (ML) disponíveis, com foco em algoritmos de classificação e regressão que possam ser aplicados à análise de desempenho acadêmico. A pesquisa será conduzida utilizando fontes acadêmicas como IEEE, ACM e Google Scholar, com o objetivo de identificar as melhores práticas e as abordagens mais promissoras para a criação do modelo preditivo.

3.2. Etapa 2 – Coleta e Pré-processamento dos Dados Acadêmicos

Nesta etapa, serão utilizados dados acadêmicos obtidos a partir de uma base pública disponível no Kaggle, contendo informações como notas, frequência, participação em atividades extracurriculares, entre outros. O pré-processamento dos dados incluirá a limpeza, normalização e transformação dos dados para garantir a qualidade e a consistência necessárias para a análise. Serão definidos os atributos que melhor representam o desempenho acadêmico e que possam ser usados para a construção do modelo preditivo.

3.3. Etapa 3 – Desenvolvimento e Avaliação dos Modelos Preditivos

Utilizando os algoritmos de ML selecionados, serão gerados modelos preditivos para analisar o desempenho acadêmico dos alunos. Os modelos serão treinados e testados com os dados pré-processados, e sua acurácia será avaliada por meio de métricas como precisão, recall e F1-score. A avaliação também considerará a capacidade dos modelos de prever corretamente os alunos em risco de baixo desempenho.

3.4. Etapa 4 – Propostas de Intervenções Pedagógicas Baseadas nos Resultados dos Modelos

Após a comparação e seleção do modelo mais eficaz, serão propostas intervenções pedagógicas baseadas nos resultados obtidos. Essas intervenções poderão incluir tutorias personalizadas, ajustes no currículo e outras estratégias voltadas para os alunos identificados como de maior risco. O objetivo é melhorar o desempenho acadêmico geral e aumentar as taxas de sucesso dos alunos na UFAM.

4. Cronograma

O cronograma estimado para a execução do projeto é apresentado na Tabela 1.

Tabela 1. Cronograma das Etapas do Projeto.

Etapa	Início	Término
Pré-processamento dos dados	20/05/2025	20/06/2025
Geração e avaliação dos modelos preditivos	20/05/2025	20/06/2025
Comparação dos classificadores de Machine Learning	20/06/2025	20/07/2025
Apuração de resultados e inferências finais	20/06/2025	20/07/2025

5. Resultados

Durante a fase inicial deste projeto, foram realizadas análises preliminares das técnicas de Machine Learning (ML) voltadas para a previsão de desempenho acadêmico e evasão escolar. Os resultados obtidos forneceram uma visão inicial sobre as abordagens mais promissoras.

5.1. Pesquisa de Algoritmos de Machine Learning (Etapa 1)

A primeira etapa concentrou-se na investigação de algoritmos de classificação e regressão para a análise do desempenho acadêmico dos alunos. Fontes acadêmicas renomadas, como IEEE, ACM e Google Scholar, foram analisadas para identificar as melhores práticas na previsão de evasão escolar. Estudos iniciais indicam que modelos baseados em árvores de decisão, como o Random Forest, apresentam potencial significativo para a identificação de padrões em dados acadêmicos. O Random Forest se destaca pela capacidade de lidar com um grande volume de variáveis e fornecer previsões robustas.

A pesquisa indicou que o modelo gerou representações gráficas de árvores de decisão que ilustram a influência de fatores acadêmicos e socioeconômicos na predição da evasão, bem como o código utilizado em diferentes etapas do treinamento do modelo, incluindo a definição de variáveis de entrada e saída, o treinamento de um modelo de árvore de decisão, e a construção do Random Forest.

5.2. Mapeamento Sistemático da Literatura (MSL)

Um Mapeamento Sistemático da Literatura (MSL) foi planejado com o objetivo de analisar publicações científicas e identificar ameaças de segurança e soluções em Redes Sociais Online (RSO), do ponto de vista do pesquisador.

As questões de pesquisa (QP) do MSL foram: ”Quais algoritmos de Machine Learning apresentam maior eficácia na previsão de desempenho acadêmico?”(QP1) e ”Como técnicas de Machine Learning podem ser aplicadas para prever a evasão escolar no ensino superior com base em dados acadêmicos?”(QP2).

A estratégia de busca utilizou a biblioteca digital Elsevier Scopus, focando em publicações científicas (artigos de conferências e periódicos com revisão por pares) em inglês, da área de Ciência da Computação. A string de busca foi construída com base no critério PICOC (População: ”higher education”; Intervenção: ”machine learning”; Resultados: ”school dropout”).

Os critérios de seleção (Tabela 4 no documento original) foram aplicados em dois filtros para garantir a relevância dos artigos. Inicialmente, cerca de 70 artigos foram retornados; após a seleção, 11 artigos foram incluídos, correspondendo a aproximadamente 15,7% do total inicial.

A extração de dados classificou informações da publicação e dados temáticos. Em relação às questões de pesquisa, 66,7% dos artigos abordaram a aplicação de técnicas de ML para prever evasão escolar e o uso de dados acadêmicos para predição, enquanto 33,3% abordaram o desenvolvimento de ferramentas ou sistemas baseados em IA para mitigar a evasão.

5.3. Execução do Experimento de Classificação (Etapas 2 e 3)

Para dar seguimento aos objetivos propostos, foi conduzido um experimento prático utilizando uma base de dados pública do Kaggle, contendo 4424 registros e 35 variáveis sobre o desempenho de estudantes. O problema foi modelado como uma tarefa de classificação multiclasse, com o objetivo de prever se um aluno se enquadraria em uma das três categorias: **Evasão** (Dropout), **Graduado** (Graduate) ou **Matriculado** (Enrolled).

5.3.1. Pré-processamento e Engenharia de Atributos

A etapa inicial consistiu no pré-processamento e na engenharia de atributos para enriquecer o conjunto de dados. Duas novas variáveis foram criadas para capturar de forma mais direta o desempenho acadêmico:

- `total_approved_units`: A soma de todas as unidades curriculares em que o aluno foi aprovado.
- `approval_rate`: A taxa de aprovação, calculada como a razão entre as unidades aprovadas e as unidades matriculadas.

Após esta etapa, foi aplicado um método de seleção de atributos, o `SelectKBest` com o teste ANOVA F-test, para identificar as 25 variáveis mais relevantes para o modelo, otimizando o processo de treinamento e reduzindo ruídos.

5.3.2. Desenvolvimento e Avaliação dos Modelos

Foram treinados e avaliados seis algoritmos de Machine Learning distintos, incluindo abordagens baseadas em árvores (Random Forest, Gradient Boosting, XGBoost, Extra Trees) e outros métodos clássicos (Regressão Logística, SVM). A Tabela 2 apresenta os resultados comparativos de acurácia, área sob a curva ROC (ROC-AUC) e F1-Score para cada modelo em sua versão inicial. O desempenho dos modelos também pode ser visualizado na Figura 1.

Tabela 2. Métricas de desempenho dos modelos de classificação.

Modelo	Acurácia	ROC-AUC	F1-Score	CV Score (± std)
Random Forest	0.7593	0.9047	0.7654	0.7544 (±0.0063)
Gradient Boosting	0.7661	0.8995	0.7579	0.7624 (±0.0068)
XGBoost	**0.7740**	**0.9043**	**0.7661**	**0.7660 (±0.0086)**
Extra Trees	0.7277	0.9005	0.7414	0.7485 (±0.0087)
Regressão Logística	0.7503	0.8866	0.7474	0.7598 (±0.0133)
SVM	0.7198	0.8913	0.7342	0.7398 (±0.0125)

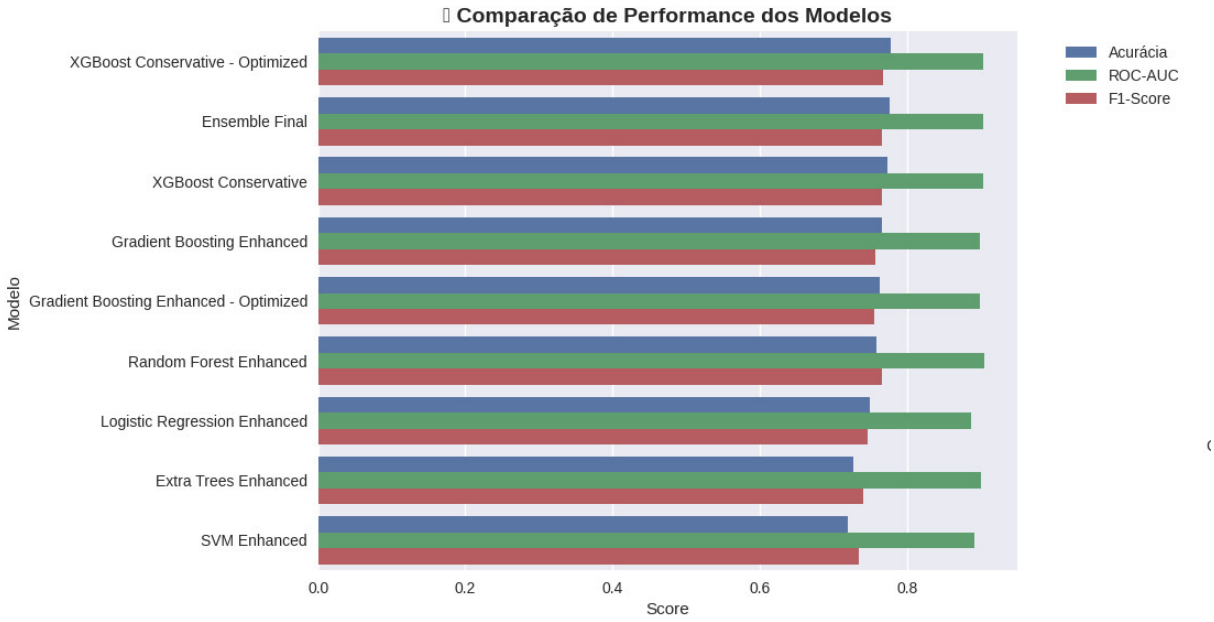

Figura 1. Comparativo de Acurácia e F1-Score entre os modelos.

Os modelos XGBoost e Gradient Boosting, que apresentaram os melhores resultados iniciais, foram submetidos a um processo de otimização de hiperparâmetros utilizando `GridSearchCV`. O modelo XGBoost otimizado alcançou uma acurácia de **0.7774**. Adicionalmente, um modelo de ensemble, combinando os três melhores classificadores via `VotingClassifier`, foi construído, atingindo uma acurácia de 0.7763.

5.3.3. Análise do Modelo Final

O modelo **XGBoost Otimizado** foi selecionado como o melhor preditor. O relatório de classificação detalhado, apresentado na Tabela 3, revela uma alta capacidade de predição para as classes ”Graduate”(recall de 92%) e ”Dropout”(recall de 76%). No entanto, o modelo apresentou dificuldade em identificar corretamente os alunos da classe ”Enrolled”, com um recall de apenas 41%. A Figura 2 ilustra a matriz de confusão do modelo, onde é possível observar visualmente esses acertos e erros.

Tabela 3. Relatório de Classificação do Modelo XGBoost Otimizado.

Classe	Precision	Recall	F1-Score	Support
Dropout	0.82	0.76	0.79	284
Enrolled	0.55	0.41	0.47	159
Graduate	0.81	0.92	0.86	442
Accuracy			**0.78**	**885**
Macro Avg	0.73	0.70	0.71	885
Weighted Avg	0.77	0.78	0.77	885

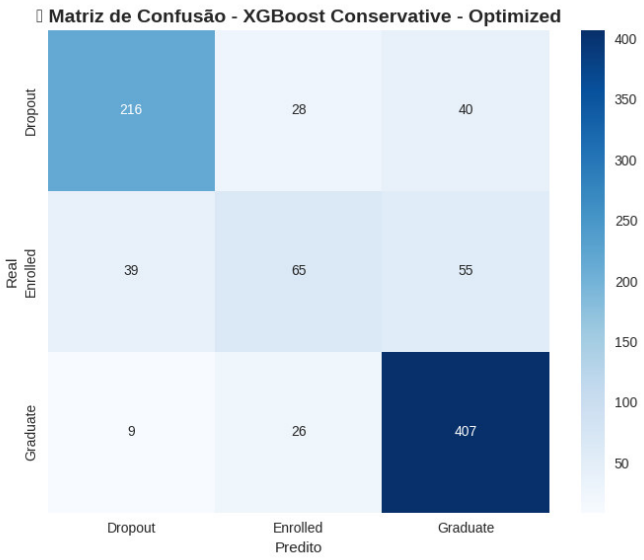

Figura 2. Matriz de Confusão do modelo XGBoost Otimizado.

A análise de importância dos atributos revelou que a variável `approval_rate` (taxa de aprovação) é, de longe, o fator mais preditivo, com um peso de 0.37, seguida por `Tuition fees up to date` (mensalidades em dia) e `Curricular units 2nd sem (approved)` (unidades aprovadas no 2º semestre), conforme destacado na Figura 3. Isso indica que o desempenho acadêmico contínuo e a situação financeira do aluno são determinantes para seu percurso na universidade.

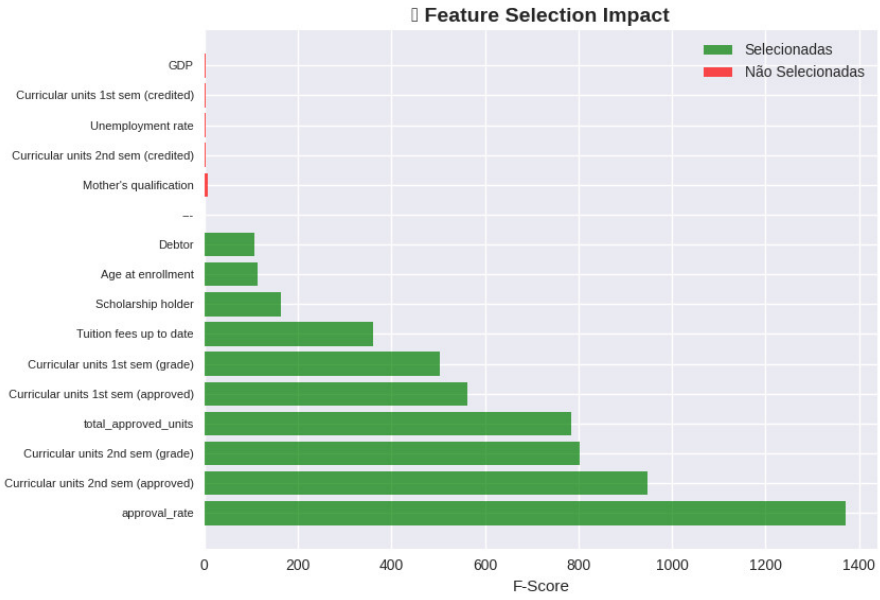

Figura 3. As 10 features mais importantes segundo o modelo XGBoost.

6. Conclusão

Este trabalho demonstrou a viabilidade e a eficácia da aplicação de técnicas de Machine Learning para a análise e predição do desempenho acadêmico em instituições de ensino superior. O objetivo de desenvolver um modelo preditivo para identificar padrões e prever resultados futuros foi alcançado com sucesso.

O modelo final, baseado no algoritmo XGBoost, atingiu uma acurácia geral de aproximadamente 78%, mostrando-se particularmente robusto na identificação de alunos com alta probabilidade de se graduarem (recall de 92%) e daqueles em risco iminente de evasão (recall de 76%). Estes resultados são promissores e fornecem uma base sólida para a implementação de ações preventivas.

A principal descoberta deste estudo reside na identificação dos fatores de maior impacto no percurso acadêmico. A "taxa de aprovação"(`approval_rate`) emergiu como o preditor mais significativo, reforçando que o acompanhamento contínuo do rendimento do aluno é fundamental. Além disso, fatores como a regularidade no pagamento das mensalidades (`Tuition fees up to date`) e o número de unidades curriculares aprovadas no semestre anterior também se mostraram cruciais.

A principal limitação do modelo foi a baixa capacidade de prever com precisão a classe "Matriculado"(Enrolled). Isso sugere que este é um estado mais transitório e complexo, cujos fatores determinantes podem não ter sido totalmente capturados pelas variáveis disponíveis.

Com base nos resultados, as intervenções pedagógicas propostas na Etapa 4 podem ser direcionadas com maior precisão. Alunos que apresentem uma queda na taxa de aprovação podem ser rapidamente identificados e direcionados para programas de tutoria e reforço acadêmico. Da mesma forma, a importância do fator financeiro aponta para a necessidade de políticas de auxílio e suporte ao estudante.

Como trabalhos futuros, sugere-se a incorporação de novas fontes de dados, como informações comportamentais do ambiente virtual de aprendizagem (AVA) ou dados socioeconômicos mais detalhados, para aprimorar a predição da classe "Matriculado". Adicionalmente, a implementação deste modelo como uma ferramenta piloto no ambiente da universidade permitiria validar sua eficácia em um cenário real e refinar continuamente suas previsões, contribuindo para a redução das taxas de evasão e o aumento do sucesso acadêmico.

Referências

Bakker, T. et al. (2023). **Predicting academic success of autistic students in higher education**. *Autism*, 27(6):1803–1816.

Bastos, A. C. F. L. C. et al. (2024). Processo decisório de gestão da evasão nos cursos de graduação da universidade federal do sul da bahia: contribuições da abordagem da mineração de dados educacionais. Dissertação (mestrado), Universidade Tecnológica Federal do Paraná.

Cabrera, E. et al. (2023). **AI-based application to predict student dropout in the National Learning Service SENA**. In *IEEE. 2023 XIII International Conference on Virtual Campus (JICV)*, pages 1–4.

Fernandez-García, A. J. et al. (2021). **A real-life machine learning experience for predicting university dropout at different stages using academic data**. *IEEE Access*, 9:133076–133090.

Filho, J. H., Vinuto, J., and Leal, F. (2020). **Aplicação de machine learning para previsão de evasão no ensino superior**. *Revista Brasileira de Informática na Educação*, 28(1):45–60.

Goodfellow, I. et al. (2016). *Deep learning*. MIT Press Cambridge.

Kantorski, V. et al. (2016). **Predição de evasão em instituições públicas de ensino superior utilizando machine learning**. In *Anais do Simpósio Brasileiro de Informática na Educação*.

Kitchenham, B. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE-2007-01, Keele University. Also available via Citeseer.

Matz, S. C. et al. (2023). **Using machine learning to predict student retention from socio-demographic characteristics and app-based engagement metrics**. *Scientific Reports*, 13(1):5705.

Mitchell, T. M. (1997). *Machine learning*. McGraw-Hill New York.

Najera, A. B. U., Calleja, J. d. l., et al. (2018). **Selección de tutores académicos en la educación superior usando árboles de decisión**. *REOP - Revista Española de Orientación y Psicopedagogía*, 29(1):108–124.

Petticrew, M. and Roberts, H. (2008). *Systematic reviews in the social sciences: A practical guide*. John Wiley & Sons.

Russell, S. J. and Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.